doi:10.1038/nature10532

Contents

List of Tables

List of Figures

1 Supporting text

1.1 Biological sample collection and RNA extraction

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the sampling strategy and about the RNA extraction techniques.

The 131 organ samples that provided the foundation for this study were obtained from various sources (Supplementary Table 1; sample overview in Supplementary Table 2). Informed consent for use of the human tissues for research was obtained in writing from all donors or the next of kin. All non-human primates used in this study suffered sudden deaths for reasons unrelated to the participation in this study and without any relation to the tissue used. All necessary permits to use the listed samples for biomedical research were obtained and can be provided upon request.

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 and maximize sample comparability, major parts of each organ (covering the different structures/cells) were dissected and homogenized before RNA extraction where possible. Given that the brain is a particularly heterogeneous tissue, we sampled from two major regions of the brain for each species (Supplementary Table 1): (i) prefrontal cortex/frontal lobe (humans and other primates) or entire brain except olfactory bulb and cerebellum (all other); (ii) cerebellum (cerebellar cortex). Notably, the cerebellum was chosen not only because it is a brain region with interesting functional roles (motor control and involvement in cognitive functions such as attention and language) but also because it is a well-defined and conserved brain region. It is thus structurally similar between species and easily dissectible, in spite of the major differences in brain size among the amniote species studied. With respect to the cortex/brain sampling, it is further noteworthy that previous studies suggested that while the cortical regions substantially differ from the cerebellum in terms of gene expression (which we account for by our sampling procedure), different regions within the cerebral cortex only show small expression differences[1]. These are likely negligible given the evolutionary time scales (and hence major expression divergences) considered in most of the presented analyses, as also indicated by the well-resolved gene expression phylogenies and the slow rate of expression change detected in the between-species comparisons of the cortex/brain regions. Generally, expression differences detected in our analyses may reflect functional changes of genes in a given tissue (or cell type) but also differences related to the structure, cellular composition, and size of tissues and cells that arose during evolution.

1.2 RNA sequencing

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the procedures that were used for RNA sequencing and for basecalling.

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 Bioanalyzer. To ensure data comparability, only samples with high RNA integrity (RIN) values (Supplementary Table 1) were used in this study. Indeed, there is no indication that results are affected by RIN value variability (*i.e.*, branch lengths and RIN values are not correlated; data not shown).

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 cDNA 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 Bioanalyzers. Potential influences on RNA sequencing results due to different experimenters preparing the libraries were ruled out on the basis of RNA-Seq data comparisons of replicate libraries prepared by the different experimenters (*i.e.*, expression levels derived from replicate libraries were highly correlated: Spearman's $\rho > 0.99$). The RNA-Seq libraries were each sequenced (76 cycles) in at least one lane of the Illumina Genome Analyzer IIx platform according to the manufacturer's specifications. Technical replicates (*i.e.*, sequencing the same library on different machines) were performed to rule out potential biases during the sequencing step (*i.e.*, expression levels between technical replicates were highly correlated: Spearman's $\rho > 0.98$).

After sequencing, we processed the fluorophore intensity files with the Ibis base caller (version 1.11)[2], in addition to applying the standard Illumina base calling algorithms. As illustrated in Supplementary Note Table 1 and Supplementary Note Figure 1 for a small subset of samples, we found that Ibis significantly increased the number of usable reads and drastically reduced the error rate. This improvement was more pronounced for the sequencing runs processed with early versions of the Illumina pipeline, but remained noticeable even for the latest Illumina release (GA pipeline 1.60). All subsequent analyses were performed on the Ibis-called reads.

1.3 Initial read mapping

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the initial read mapping procedure, which was mainly used for the annotation refinement procedure.

Read mapping with TopHat and Bowtie We extracted the reference genome sequences from Ensembl[3], release 57. We removed the haplotypic regions present in the human and orangutan genomes before the RNASeq analysis. For the bonobo, we used the chimpanzee genome sequence as a reference, since the bonobo genome sequence was not publicly available at the time when these analyses were performed.

To get an initial mapping of the RNASeq reads, we used TopHat[4] to align the reads on the reference genome sequences. TopHat internally uses the fast-aligner Bowtie[5], and performs an *ab initio* identification of splice junctions, without relying on genomic annotations. For the splice junction detection, we chose the parameters in order to allow intron sizes between 40 bp and 1 Mb. The anchor size (*i.e.* the minimum aligned length spanning each of the two exons that define a splice junction) was set at 8 bp, and we allowed 1 mismatch on the anchor region. We removed the threshold on the minimum isoform frequency, so that splice junctions belonging to rare isoforms would also be reported.

Definition of transcribed islands and splice junction coordinates We filtered the read alignments accepted by TopHat in order to remove the mapping ambiguity. To do this, we extracted the best mapping(s) for each read, based on the number of mismatches in the alignment, and we selected those reads for which the best mapping was unique. We then computed the per-base read coverage given by the filtered read alignments, and we extracted from it the coordinates of the "transcribed islands" (*i.e.*, the maximum contiguous segments with a non-null read coverage).

We extracted the splice junction coordinates from the gapped read alignments provided by TopHat, after removal of ambiguous mappings. We checked that the anchor size and mismatch requirements (defined above) were fulfilled. We kept only those junctions that were supported by at least one read aligned with at most 3 mismatches. We inferred the sense strand for the junction based on the splice sites and we considered for our analyses only those junctions that had GT-AG and GC-AG splice sites, for which the inference of the strand was reliable.

Splice junction detection for genes with retrocopies For the splice junction detection, TopHat uses only those reads that could not be aligned without gaps on the genome sequence[4]. This restriction results in a considerable increase of computation speed, but we suspected that it may also lead to reduced efficiency of splice site detection for one particular class of genes: those that have retrotransposed duplicates. Indeed, since the retrotransposed duplicate is a copy of the intronless mature mRNA, the junctions between consecutive exons of the parental gene can be found as such (without gaps) in the genome sequence.

To verify this intuition, we analyzed a set of genes with retrocopies for the orangutan genome (Kaessmann and Potrzebowski, unpublished). We measured the efficiency of the splice junction detection through the number of annotated junctions that were confirmed by TopHat, and we compared this measurement between genes with and without retrocopies. As expected, we found that, at equal expression level, the genes that have retrocopies are less well represented in the TopHat results: for example, for the female cerebellum sample, we found that 83% of the annotated junctions were confirmed by TopHat for genes without retrocopies, but only 72% were confirmed for genes with retrocopies (these results are for the genes found in the highest quartile of the expression level distribution).

To estimate the magnitude of this bias and to eventually correct for it, we developed a method for detecting retrocopies in a given genome. We focused on recent retrocopies (i.e. those that are highly similar in sequence to the parental gene), since more diverged retrocopies will not affect the sensitivity of the junction detection. Our procedure consists of several steps:

1. we define "exon blocks" based on the annotations (exons from the same gene that are separated by less than 20bp are collapsed; some genes that are annotated as multi-exonic may become in this way mono-exonic);

2. we mask on the genome sequence the exon blocks from multiexonic genes (as defined in the previous step) and the repeated sequences[6];

3. we then do a BLAST[7] search for nucleotide alignments (blastn) of the exon sequences of multi-exonic genes, against the exon-masked, repeat-masked genome;

4. we select BLAST hits with e-value $< 10^{-5}$, percent identity ≥ 90, percent insertions-deletions ≤ 10;

5. for each multi-exonic gene, we search for clusters of hits that are separated by less than 20bp, that align with at least 2 exons of that gene, and that are at least 75bp length (given that we generate 76bp-long reads, the presence of smaller retrocopies cannot affect the splice junction detection procedure).

6. we output the coordinates of the clusters detected at the previous step, as potential retrocopies.

We detected between ≈1,800 and ≈2,200 genes that generate retrocopies in Eutherian species (Supplementary Note Table 2). In comparison, for platypus and chicken, only ≈ 200 such genes were found. This observation is in agreement with previous notions that most retrocopies were generated by the retrotransposition machinery of the LINE L1 element[8]. While this LINE family is present and known to be (or have been) active in therian species, it is absent or inactive in monotremes and birds, and these lineages are thus less prone to retroposition[8].

To improve the sensitivity of the splice junction detection procedure, we masked the detected retrocopies in the reference genome sequence, and re-run TopHat on the masked sequence. As illustrated in Supplementary Note Table 2, the number of detected junctions for the parental genes was significantly increased after retrocopy-masking. For maximum sensitivity, in our final analyses we combined the two sets of splice junctions (detected on the unmasked and retrocopy-masked genome).

1.4 Refinement of genomic annotations

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the annotation refinement procedure.

Extension of gene models We extracted genome annotations from Ensembl[3] (release 57), for 9 of the 10 species in our dataset (excluding the bonobo, for which the genome sequence and annotations were not publicly available at the time when these analyses were performed), and used them as a starting point for our annotation refinement procedure.

For each sample in our dataset, we defined the set of transcribed islands and extracted the splice junctions based on the unambiguous read alignments provided by TopHat (see above). We filtered the transcribed islands to remove the regions with extremely low-coverage (supported by only one RNASeq read) and we selected the islands that were connected through splice junctions to other transcribed regions. To avoid ambiguity due to segmental duplications, we removed from our dataset the transcribed islands that were connected with exons of more than one gene.

For each gene, we defined "exon blocks" as the union of exon coordinates from all alternative transcripts annotated for that gene. We then searched for connected transcribed islands that were found within 100kb of annotated gene boundaries, and added those islands to the gene models, whenever the extremities of the splice junctions connected to the islands were found within the boundaries of previously-known exon blocks. We repeated this step until no more transcribed islands could be added to the gene models; the exon blocks coordinates and the gene boundaries were recomputed at each iteration.

Finally, the extended exon blocks obtained for all the RNASeq samples of a given species were collapsed into a single set. This annotation set will be referred to as "extended exon blocks" or "extended annotations" throughout the manuscript. Note that this set of exon blocks may include retained introns.

Validation and quantification of splice junctions We observed that the splice junctions annotated in Ensembl are not all found in the set of junctions detected with TopHat; the proportion of confirmed junctions varies between 57% and 82% in the different species (Supplementary Note Table 3). This discrepancy may be explained, at least in part, by the tissue-specificity of alternative transcripts; a fraction of these tissue-specific isoforms will not be represented in our dataset, which covers only 6 tissues/organs, but might be included in the Ensembl annotations. However, we were concerned that some annotated splice junctions might be missed by TopHat, due to the length (76bp) of our RNASeq reads. In Ensembl annotations, between 16% and 26% (according to the species) of all exons are shorter than 75bp, meaning that a single RNASeq read can often be aligned over two or more splice junctions. The read alignment performed by the initial releases of TopHat did not take into account this possibility, and the version of TopHat used here (1.0.13) provides a solution to this problem by splitting the reads into (non-overlapping) 25bp segments before performing the alignment. Since no estimations of the effectiveness of this solution were publicy available, to ensure the completeness of our annotations, we developed an additional method (termed here "multi-splice validation") for the validation of splice junctions, that takes into account the possibility for a read to span multiple exon junctions.

Our multi-splice validation method uses as an input the coordinates of the extended exon blocks (obtained as described above), and a set of splice junctions compiled from two sources: Ensembl annotations and TopHat-determined junctions for which the boundaries were found within the coordinates of the extended exon blocks. From this information, we construct for each splice junction the set of all 150bp ($= 2 \times (read\_length - 1)$) flanking sequences that can be associated with this splice junction in the isoform repertoire. The construction of these flanking sequences is done with a recursive algorithm, that can be summarized as follows:

To construct the right-hand flanking regions, of length $l = 75bp$, for a junction between exon blocks e_1 and e_2, starting at position s in e_1 and ending at position e in e_2:

1. extract all of the splice junctions start coordinates in exon block b_2 that are greater than e (denoted here $Start_{b2}$);
2. if not already included, add the end coordinate of exon block b_2 to $Start_{b2}$;
3. append the segments $e - t$ to the set of flanking regions, for each t in $Start_{b2}$;
4. extract the end coordinates (denoted here End_t) of all the splice junctions starting at t, for each t in $Start_{b2}$;
5. if End_t is empty, stop;
6. else, for each end coordinate e' in End_t, belonging to exon block b'_2, re-evaluate the remaining flanking length at $l = l - (t - e + 1)$, set $e = e'$, $b_2 = b'_2$;
7. if $l > 0$, go to step 1;
8. repeat while the length of the flanking regions is < 75.

The same procedure is used to construct the left-hand flanking sequences. Note that we did not attempt to apply this algorithm on genes with more than 200 splice junctions, since its computation time increases exponentially with the number of junctions; between 2 and 7 genes (depending on the species) were excluded for this reason. After constructing the set of flanking sequences, we used Bowtie[5] to align the RNASeq reads on these sequences. As above, we filtered the alignments to extract the unambiguously mapping reads. As for TopHat-detected junctions, we consider a splice junction to be validated if it is supported by at least one unambiguously mapping read, with an anchor size of 8bp and with at most 1 mismatch allowed in the anchor region. This procedure of junction validation succeeded in increasing the number of Ensembl-annotated splice junctions that are confirmed by our RNASeq data (Supplementary Note Table 3), but note that the improvement with respect to the TopHat-only results is relatively weak (2 to 5% more junctions confirmed).

We used the same set of flanking sequences to quantify the frequency of the splice junctions. To do this, for each validated junction, we count the number of unambiguously mapped RNASeq reads that align on one of its associated flanking sequences, with an anchor size of 4 bp, and at most 1 mismatch in the anchor region. These read counts were used to quantify isoform frequency for the constitutive exon definition (see section 1.5 below).

Refinement of exon coordinates The annotation extension procedure described above can result in including retained introns in the gene models. To preserve coherence with Ensembl annotations, in which transcripts known to contain retained introns are distinguished from the other isoforms, we developed a method that defines exon boundaries based on splice junction coordinates. The main principle is simple: the start position of a splice junction corresponds to the end coordinate of an exon, and *vice-versa*. For exons that have multiple 5' or 3' splice sites, we retain the outermost coordinates. The full algorithm can be summarized as follows:

1. we combine the sets of splice junctions obtained with our "multi-splice validation" procedure for each sample of a species into a single set of junctions;

2. for each gene, we extract the coordinates of the exon blocks obtained after the extension of gene models (see above);

3. for each exon block, we extract those splice junctions that have at least one extremity within the exon boundaries, and the other extremity within the boundaries of an exon block (potentially the same exon block) of the same gene;

4. we order the splice junction extremities found within an exon block, and label them as either "start junction" or "end junction" (points that are both a junction start and a junction end are excluded);

5. in the simplest case, if an exon block contains an "end junction" coordinate at position i, and a "start junction" coordinate at position j, with $i < j$, we define an exon with coordinates i - j;

6. if a "start junction" coordinate at position j_2 follows after another "start junction" coordinate at position j_1 (with $j_1 < j_2$), the end coordinate of the exon is extended at $j2$;

7. conversely, if an "end junction" coordinate at position j_2 follows after another "end junction" coordinate at position j_1 (with $j_1 < j_2$), the start coordinate of the exon remains set at $j1$;

8. if the first junction coordinate (denoted f) found within an exon block is strictly greater than the start of the exon block (denoted s):

   (a) if f is an "end junction" coordinate, the segment s - $f - 1$ is considered non-exonic;

   (b) if f is a "start junction" coordinate, the segment s - f is considered exonic, and s is assimilated to an exon start.

9. if the last junction coordinate (denoted l) found within an exon block is strictly smaller than the end of the exon block (denoted e):

   (a) if l is an "start junction" coordinate, the segment $l + 1$ - e is considered non-exonic;

   (b) if l is a "end junction" coordinate, the segment l - e is considered exonic, and e is assimilated to an exon end.

Note that for this procedure we considered only those splice junctions that were supported by at least 2 unambiguously mapping reads.

This dataset will be referred to as "refined exon blocks" or "refined annotations" throughout the manuscript.

Comparison with Ensembl annotations To validate our annotation extension and refinement procedure, we performed a series of comparisons with Ensembl annotations. We first analyzed the total length covered by known protein-coding genes, as computed based on Ensembl annotations and on our refined annotations. As shown in Supplementary Note Table 4, our refined annotations bring a significant increase in gene length for species with relatively poor genome annotations (such as platypus, opossum and chicken), whereas for well-studied species (such as human and mouse), the total gene length remains remarkably stable. Importantly, the increase in gene length is not determined solely by the sequencing depth, as one may have feared (Supplementary Note Table 4).

The exonic length of known protein-coding genes is also higher in the refined annotations than in Ensembl (Supplementary Note Table 4). However, not all Ensembl-annotated exons are found in our annotations, and *vice-versa* (Supplementary Note Table 5). To verify whether this discrepancy between the two annotation sets is not an artifact of our approach, we analyzed the evolutionary conservation of the different classes of exons. To do this, we extracted PhastCons[9] conservation scores from the UCSC genome browser[10], for 6 species in our dataset (human, orangutan, mouse, opossum, platypus, chicken - PhastCons data was not available for the other species). For human and mouse several sets of PhastCons scores were available; we used the one corresponding to the "vertebrate" set of species (as opposed to the "placental" set). Note that the PhastCons scores are not directly comparable among species, since different alignments were used for the computations. We analyzed three classes of exons: Ensembl exons that were also found in our refined annotations, exons found only in Ensembl (no overlap with our annotations) and exons found only in our annotation dataset. For each class of exons, we computed the mean PhastCons scores over 100bp, divided into 4 segments: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. As expected, for confirmed Ensembl exons, the conservation score is much higher within the exon than on the intron, with a peak in the ≈ 10 positions surrounding the exon boundaries - likely due to splicing signals (Supplementary Note Figures 2 to 4). The same pattern is observed for the exons found only in our annotations, however, for exons present only in Ensembl annotations the conservation score is not different from that of the neighbouring introns (Supplementary Note Figures 2 to 4). This result confirms that our annotation extension procedure was successful in adding genuine exons to the gene models. We note however that the mean conservation score is lower for new exons than for confirmed Ensembl exons; the difference between the two is particularly strong for the well-studied species (human and mouse), while for the other 4 species the effect is much weaker.

1.5 Definition of constitutive exons

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the definition of "constitutive" exons, which were used for computing gene expression levels.

Prior to evaluating gene expression levels, we sought to eliminate minor splice isoforms from the gene models, in order to reduce the level of splicing-related noise in our data. To do this, we combined two approaches: first, we used information on read coverage variation along the gene to eliminate regions with unusually low coverage (likely to belong to minor isoforms), and second, we detected and quantified alternative transcription events based on splice junction frequencies and read coverage variation, and excluded isoforms with low frequency.

Exclusion of low-coverage regions with segclust We observed that the per-base read coverage varies significantly along a single gene, or even along a single exon (illustrated in Supplementary Note Figure 5). This heterogeneity can be partly explained by artifacts such as variable sequencing (fragmentation bias) or mapping efficiency (related for example to a particular nucleotide composition, or to segmental duplications). However, it is likely that most of the variation in read coverage stems from genuine biological phenomena: alternative transcription and splicing, resulting in an unequal representation of the different exon segments in the isoform repertoire. Regardless of the underlying reasons, identifying and removing regions with unusually low read coverage is necessary in order to minimize the level of noise when performing comparisons of gene expression levels.

There are multiple patterns of read coverage variation, even along a single exon block (Supplementary Note Figure 5); the automatic identification of segments with singular read coverage is thus not a trivial matter. To attain this goal, we took advantage of the existence of statistical methods for the identification of copy-number-variable regions from microarray comparative genomic hybridization data[11,12]. The approach chosen here uses a dynamic-programming/expectation-maximization algorithm to identify homogeneous segments in the read coverage pattern, and provides a heuristic for the selection of the optimal number of segments[12].

For this analysis, we used as an input the set of extended annotations, that include all transcribed islands connected to the genes, before the precise exon definition step. We preferred not to restrict this analysis to the “refined” annotation set (that includes a precise definition of exon coordinates and removes potential retained introns), because our procedure for precise exon definition is solely based on splice junctions coordinates, and thus may misclassify as “intron” exonic regions for which the splice junction detection was inefficient. We used the alignment of unambiguously mapping reads obtained with TopHat to compute the per-base read coverage for each sample in our dataset, and we combined all the samples available for a species to get the total coverage (tc). We further transformed tc on a log2 scale, with the formula $log_2(tc + 1)$ (we added an offset of 1 to preserve the information for regions with 0 coverage). The log_2 transformation was applied in order to respect the requirement for a Gaussian distribution of the signal, imposed by the segmentation algorithm.

To apply this segmentation approach on our data, we used its implementation in the R[13] package segclust. We analyzed each transcribed block separately, and we set the maximum number of segments to be evaluated at 7 (since this parameter greatly impacts the computation time, we were obliged to set a restrictive threshold). Blocks that were shorter than 50bp and that had a maximum coverage below 4 reads per base were not evaluated. In addition, to increase computation speed for relatively long transcribed blocks (>500bp), we computed the mean read coverage on 5bp (or 10bp for exon blocks

>100bp) non-overlapping sliding windows, and applied the algorithm on this restricted set of read coverage values.

To define "constitutive" exon parts, we computed the mean per-base read coverage on each segment determined by `segclust`, and evaluated the maximum read coverage observed for the segments within the same exon block. We then discarded (or termed "non-constitutive") those segments that had a mean read coverage below one third of the maximum. The 1/3 threshold is arbitrary; several other thresholds were tested (1/4,1/5), leading to similar results (not shown).

The results of `segclust` are exemplified in Supplementary Note Figure 5. We observed that boundaries of the Ensembl-annotated exons are often (but by no means in all cases) nicely confirmed by this procedure. As a validation of the `segclust` results, we analyzed the proximity between the boundaries of the segments and the coordinates of the TopHat-validated splice junctions. We found that 77 to 83% of the segments have at least one boundary within 10 bp of a splice junction extremity, and this proportion is significantly higher than the one observed for the entire set of transcribed blocks (52 to 64%, Supplementary Note Table 6). Given that the `segclust` segmentation was done without any knowledge of splice junction coordinates, this significant increase in the proximity to splice junctions is an indicator of the accuracy of the procedure.

Quantification of alternative transcription events The procedure described above defines regions with unusually low read coverage (and thus potentially belonging to rare isoforms), by using only one source of information: the variation of the per-base read coverage within exon blocks. As implemented here, this method does not allow defining as "non-constitutive" entire exon blocks, as the segmentation is done within each exon. This is however necessary, for example in the case of "skipped" (or "cassette") exons with low inclusion frequency. As extending the above approach to analyze the read coverage variation along the entire gene length proved to be computationally heavy (data not shown), we developed an additional method that defines and quantifies alternative transcription events using information on splice junction coordinates and frequencies, and that flags as "non-constitutive" those exons (or exon parts) that belong to low-frequency isoforms.

We used as an input the set of exon blocks derived from our annotation extension procedure. For those exon blocks that were divided into "constitutive" and "non-constitutive" segments with `segclust`, we treated each segment as an independent exon block. For the splicing information, we used the coordinates of the validated splice junctions (see section 1.4), as well as the number of reads that support these junctions in each sample. We focused on the detection of 4 types of alternative splicing events: skipped (or "cassette") exons, retained introns, alternative 5' splice sites and alternative 3' splice sites. The detection and the quantification of these types of alternative splicing with RNASeq data was described recently[14]; we followed the principles described in this publication, as detailed below:

- An exon block with coordinates $i - j$ is termed "skipped" if there exists a splice junction with coordinates $x - y$, where $x < i$ and $y > j$, x and y found within other exon blocks of the same gene, and the strand of the splice junction is the same as the strand of the gene.

- An exon block with coordinates $i - j$ is said to contain a retained intron if there exists a splice junction with coordinates $x - y$, where $x > i$ and $y < j$, and the strand of the splice junction is the same as the strand of the gene.
- An exon block with coordinates $i - j$ is said to have several alternative 5' splice sites in one of the following cases:

1. if the strand of the gene is positive, and there are at least two positive-strand junctions with coordinates $x_1 - y_1$ and $x_2 - y_2$, such that $x_1 > i$, $x_1 \leq j$, $y_1 > j$, $x_2 > i$, $x_2 \leq j$ and $y_2 > j$, with y_1 and y_2 found within other exon blocks of the same gene.

2. if the strand of the gene is negative, and there are at least two negative-strand junctions with coordinates $x_1 - y_1$ and $x_2 - y_2$, such that $x_1 < i$, $y_1 \geq i$, $y_1 < j$, $x_2 < i$, $y_2 \geq i$ and $y_2 < j$, with x_1 and x_2 found within other exon blocks of the same gene.

- An exon block with coordinates $i - j$ is said to have several alternative 3' splice sites in one of the following cases:

1. if the strand of the gene is negative, and there are at least two negative-strand junctions with coordinates $x_1 - y_1$ and $x_2 - y_2$, such that $x_1 > i$, $x_1 \leq j$, $y_1 > j$, $x_2 > i$, $x_2 \leq j$ and $y_2 > j$, with y_1 and y_2 found within other exon blocks of the same gene.

2. if the strand of the gene is positive, and there are at least two positive-strand junctions with coordinates $x_1 - y_1$ and $x_2 - y_2$, such that $x_1 < i$, $y_1 \geq i$, $y_1 < j$, $x_2 < i$, $y_2 \geq i$ and $y_2 < j$, with x_1 and x_2 found within other exon blocks of the same gene.

For the retained introns and skipped exons, we quantified their inclusion frequency through the ratio $r = \frac{meancov_{e/i}}{meancov_{gene}}$, with $meancov_{e/i}$ the mean per-base coverage of the intron/exon, and $meancov_{gene}$ the mean per-base read coverage of all the exon blocks in the gene. For alternative 5' (or 3') splice sites, the inclusion frequency was defined as follows: $f = \frac{rc_j}{\sum_i (rc_i)}$, with rc_j the read count of the splice junction that defines the splice site j, and $\sum_i (rc_i)$ the sum of the read counts for all of the splice junctions that define alternative 5' (or 3') splice sites in the same exon block.

As for the `segclust` approach, we set a threshold on the minimum expression level of a gene (average per-base read coverage over all exon blocks of at least 3) before attempting to distinguish constitutive and non-constitutive exons. The level of expression was computed using the unambiguously mapping reads provided by TopHat, as described above (section 1.3); before computing the expression level, we excluded the exon blocks that are part of more than one annotated gene. We then excluded (or termed "non-constitutive") the exon blocks (or parts of exon blocks) that had an inclusion frequency below 0.15.

This procedure was applied independently on each sample, and the excluded blocks obtained for all samples of a single species were finally collapsed into a single set.

Statistics for constitutive exon regions To validate of our approach, we compared the coordinates of the "constitutive" exons with other annotation datasets. We first compared the "constitutive" exons with the extended exon blocks defined with our annotation procedure (that include all of the transcribed islands connected to Ensembl-annotated genes, and thus may include retained introns). We found that the "constitutive" exon parts represent between 47% and 66% (depending on the species) of the length of the extended exon blocks, for protein-coding genes (Supplementary Note Table 7). The variation among species seems to be dependent on the read coverage: the species for which we have the best sequencing depth (mouse, human and chimpanzee) are the ones for which the "constitutive" fraction is the lowest. This is not unexpected, since with more sequencing depth we can more easily include (relatively unfrequent) retained introns in the extended annotations, and those regions will be removed when defining "constitutive" exons.

The extent of the overlap between "constitutive" exons and Ensembl-annotated exons also varies extensively among species, between 67.7% (for the opossum) and 92.7% (for human) of the total length of the "constitutive" exons (note that for this comparison, we only considered Ensembl transcripts annotated as "protein-coding", excluding those annotated as "retained_intron","nonsense-mediate decay" etc., as they are likely to represent low-frequency isoforms). Conversely, the proportion of the length of Ensembl-annotated exons that is termed "constitutive" with our procedure varies between 72.5% in human and 87.3% in platypus (Supplementary Note Table 7). This variation might be explained by the quality of existing genomic annotations: for well-studied species, such as human, the existing genome annotations also include low-frequency isoforms, and thus the proportion of "constitutive" exons is relatively low.

We next studied the intersection between the "constitutive" exons and our refined exon boundaries (defined in section 1.4). We found that a high proportion of "constitutive" exons (between 93% and 97%) are indeed defined as exons with our splice-junction-based refinement procedure (Supplementary Note Table 7). The proportion is not 100% because we preferred to include in our constitutive exon set those (apparent) retained introns that correspond to relatively high frequency isoforms, rather than excluding *a priori* all retained introns. Indeed, these transcribed regions might also correspond to (or include) genuine exons for which no splice junctions could be detected.

Conversely, between 67% and 76% (depending on the species) of the total length of the refined exons was termed "constitutive" with our procedure (Supplementary Note Table 7). This fraction is lower than the one computed for Ensembl-annotated exons, for all species. This result is again expected, since for most species Ensembl annotations are likely to exclude low-frequency isoforms, but these isoforms are often detected in our RNASeq data and are thus included in our "refined" annotation dataset.

We further analyzed the influence of the annotation dataset on the estimation of the gene expression level. To do this, we computed the gene expression level as the mean per-base read coverage, averaged over all exonic positions, after a log_2 transformation ($exp = log_2(coverage + 1)$; an offset of 1 was imposed to keep information for genes with 0 coverage). As previously, we used only unambiguously mapping reads (see section 1.3) to evaluate the read coverage. Only protein-coding genes were analyzed, and we further restricted the dataset to those genes that have at least 5 Ensembl-annotated exons and at least 5 "constitutive" exon blocks. Exon blocks smaller than 20bp were removed from the annotations before computing the expression levels. We then compared the mean expression level as computed with Ensembl exons and with "constitutive" exons. We found that the expression level is higher when computed on "constitutive" exon blocks (Supplementary Note Table 8, Supplementary Note Figures 6 to 8). This is expected, given that the goal of our "constitutive" exon definition was to remove exonic regions that correspond to low-frequency isoforms. Furthermore, we computed the variance of the expression levels of exon blocks of a same gene. We found that the variance is higher when computed on Ensembl annotations than with "constitutive" exons (Supplementary Note Table 8, Supplementary Note Figures 6 to 8); this shows that our procedure was successful in increasing within-gene homogeneity for expression level estimation.

Finally, we compared the extent of sequence conservation between Ensembl-annotated exons, "constitutive" and "non-constitutive" exons. As previously (section 1.4), we used the PhastCons scores for vertebrate species sets, downloaded from the UCSC Genome Browser. We found that the "non-constitutive" exons are the least conserved, for all species (Supplementary Note Figure 9). Ensembl-annotated exons have slightly higher levels of sequence conservation than "constitutive" exons; the difference between the two is statistically significant (Wilcoxon rank sum test, p-value $< 10^{-3}$), but the levels of conservation remain comparable between the two annotation sets.

1.6 Orthologous gene sets and exon alignment

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the assembly of the two sets of 1-1 orthologous genes (for all amniotes and for primates), and about the procedure that was used to extract perfectly aligned exon sequences, which were used as controls for gene expression analyses.

To compare gene expression levels between species, we relied on the assignment of orthology relationships between gene families provided in Ensembl 57[15]. We restricted our analysis to those gene families that are perfect 1-1 orthologues, *i.e.* there is a 1-1 orthology relationship between any two species in our dataset. For the entire dataset (9 species - for the bonobo, we used chimpanzee annotations), we thus extracted 5,636 families of 1-1 orthologues. In addition, we extracted the set of 1-1 orthologues for the 5 primate species (13,277 gene families).

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 different sequences. To correct for this potential bias, we aimed to construct a set of constitutive, perfectly aligned exon regions - using this dataset to compute expression level would minimize sequence differences between species. To do this, we aligned the cDNA sequences of the orthologous gene families using TBA[16]. We filtered these alignments to extract perfectly aligned blocks of sequence (no gaps were permitted), that corresponded to exon parts considered as "constitutive" in all species.

With this procedure, we obtained a total length of aligned constitutive exons of 4.1 Mb for the 5,636 genes set, which represents between 22 and 38% of the total constitutive exon length of each species (Supplementary Note Table 9). For the primate dataset, we obtained 17.4 Mb of constitutive, perfectly aligned exons, corresponding to 44-58% of the total constitutive exon length (Supplementary Note Table 9).

We redid the main gene expression analyses using this set of constitutive aligned exons; the conclusions remain unchanged (see section 1.10 below).

1.7 Detection of new multi-exonic transcribed loci

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the detection of new multi-exonic transcribed loci with our RNA-Seq data.

In addition to extending and refining the coordinates of Ensembl-annotated genes, we also endeavoured to define and characterize new transcribed loci, not included in the existing annotations. That such loci must exist is evident from the heterogeneity of the genomic annotations of the species in our dataset: for example, in Ensembl release 57, the human annotations contain 22,320 protein-coding genes, 3,517 long non-coding RNAs and 9,456 pseudogenes, whereas the platypus annotations consist of 17,951 protein-coding genes, 0 long non-coding RNAs and 547 pseudogenes. While part of the discrepancy is undoubtedly a biological reality, the unequal annotation effort is likely responsible for most of this variability among species. We thus decided to use our transcriptome data to extend the existing repertoire of multi-exonic transcribed loci.

The principle of the method used here is simple: the transcribed islands can be considered as vertices in a graph, and the splice junctions that connect different transcribed islands are the edges of the same graph. With these conventions, the multi-exonic transcribed loci correspond to the connected components of the graph, *i.e.* those subsets of vertices in which any two vertices are connected with each other (directly or indirectly, through a more complex path), and where there are no connections with vertices outside of the set.

To apply these principles, we first constructed a unique set of transcribed islands for each species. To do this, we extracted the transcribed islands defined for each sample, from the unambiguously mapping reads extracted from the TopHat read alignment (section 1.3), and we merged their coordinates into a single set of islands. We also constructed a global set of splice junctions, by combining all the splice junctions determined with TopHat for each sample, again based on the unambiguously mapping reads (section 1.3), as well as the sets of validated junctions for Ensembl-annotated genes determined previously (section 1.4). We consider that two islands are connected if there is a splice junction that starts within the boundaries of one island and ends within the boundaries of the other. With this information, we constructed two graphs: one where the islands are connected with each other with positive-strand splice junctions, and one where the islands are connected with negative-strand splice junctions (some islands can appear in both graphs, if they have both types of splice junctions). We then extracted the connected components of each graph with a breadth-first recursive algorithm[17]. For each multi-exonic transcribed locus detected, we applied our procedure for refinement of exon coordinates (section 1.4).

We detected between ≈44,000 and ≈60,000 (depending on the species) multi-exonic transcribed loci (Supplementary Note Table 10). The total length (including both exons and introns) covered by these loci varies between ≈750 Mb (in Platypus and Chicken) and ≈1,600 Mb (in Human). Within this dataset, we then searched specifically for loci that fall outside of the boundaries of Ensembl-annotated genes (all gene categories were included, and the gene boundaries were derived from our refined annotations). We found between ≈11,000 and ≈22,600 intergenic multi-exonic loci, depending on the species (Supplementary Note Table 10). The length covered by the intergenic loci represents between 5% (for Human) and 19% (for Platypus) of the total length of all multi-exonic transcribed loci.

We must emphasize that one multi-exonic transcribed locus is not the equivalent of one gene. Indeed, one gene can be divided into several multi-exonic loci, if the RNASeq read coverage is not deep enough to allow the detection of all splice junctions. To estimate the magnitude of this detection insufficiency, we analyzed the overlap between annotated multi-exonic genes and the previously defined loci (Supplementary Note Table 11). We found that, while $> 85\%$ of all multi-exonic protein-coding have at least a partial overlap with our multi-exonic transcribed loci, only 31 - 59% are perfectly found in the second dataset (i.e. all exons are present in the same multi-exonic locus). In addition, we observed that the intergenic transcribed loci have on average fewer exons *per* locus than the loci that overlap with known genes (Supplementary Note Table 10), which is another indication that they might in fact be parts of genes, rather than full-length genes.

As done previously for the new exons added to Ensembl-annotated genes (section 1.4), we analyzed the sequence conservation profile along exons and introns of the intergenic multi-exonic loci, and compared it to the profile observed for Ensembl exons (Supplementary Note Figures 10 to 12). Again, we find that similar profiles for Ensembl exons and for exons of intergenic loci: the conservation score is much higher within the exon than on the intron, with a peak in the ≈ 10 positions surrounding the exon boundaries. The mean PhastCons score is much lower for the exons of intergenic loci than for Ensembl-exons, as observed previously for the exons that we added to known protein-coding genes. This is not unexpected, especially since genome annotations often rely (at least in part) on projecting the annotations of other species on the genome in question, and thus are necessarily biased towards highly conserved regions. Another potential explanation for this difference of sequence conservation may reside in the expression level of the new exons and loci. Indeed, highly expressed genes are known to evolve slowly, at least at the protein level[18], and the new exons and loci that we added to the annotations are generally expressed at lower levels than Ensembl-annotated genes (Supplementary Figure 1, main text). We wanted to verify whether the expression level difference might suffice to explain the different extents of sequence conservation. To do this, we estimated the global exon expression level through the mean per-base read coverage (all samples confounded, and with a log2 transformation - $log_2(readcoverage + 1)$ - that preserves null values), and we divided the exons into 5 equal-size classes based on their expression level. We then computed the mean PhastCons score for each expression class, separately for Ensembl-annotated exons, new exons added to known genes and exons of intergenic loci. As expected, the expression level is positively correlated with the level of sequence conservation, for all three classes of exons; however, even at similar expression levels, new exons still have lower sequence conservation than Ensembl-annotated exons (Supplementary Note Figure 13).

1.8 Evaluation of gene expression

This section is an extension of the Methods and of the Methods Summary in the main text. It provides additional information about the procedure used for the final estimation of gene expression levels.

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 Figure 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 etc.). Specifically, we derived all possible read sequences for each annotation (∼150 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[5]. 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[5]. 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 number of mismatches were chosen (in the case of identical numbers of mismatches spliced reads were favored). Reads that map equally well to different genomic loci (e.g., 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 based on these ratios (*i.e.*, loci receive unique reads in proportion to their unique read mapping ratios). If two or several loci have identical sequences (i.e., they have no uniquely mapping reads), reads are distributed evenly among these copies — if these copies are all multi-exonic. However, in the case of multi-exonic "parental" genes and their identical retroposed gene copies, reads are assigned exclusively to the parental genes, given that the majority of retrocopies (in particular recent ones) are likely to be nonfunctional or at least expressed at very low levels[8]. Consistently, a representative analysis of genomic read coverage and expression levels resulting from our mapping procedure reveals that retroposed gene copies (which may represent a particularly strong confounding factor with respect to highly expressed housekeeping genes that produce many retrocopies[8]), overall only receive relatively few reads and have significantly lower expression levels than their parental genes from which they derive, especially in the case of retropseudogenes (Supplementary Note Figure 15). In summary, while our mapping approach is overall similar to the only previous procedure that takes into account duplicate gene copies[19], the more detailed consideration of intronless and intron-containing gene copies in our method (*e.g.*, by not only distributing reads based on the number of unique reads but also on the basis of the unique read coverage among copies; penalization of identical retrocopies) should provide for larger numbers of correct read assignments and hence more reliable estimates of expression levels. Indeed, the advantage of assessing the expected uniquely mappable area and unique read coverage per genomic element using artificial reads prior to final read mapping was demonstrated in a recent study[20].

Expression levels and normalization Based on the final read assignments described in the previous section, we calculated standard RPKM (reads per kilobase of exon model per million mapped reads) expression values (that were then log_2 transformed) for our orthologous gene set. To render the data comparable across species and tissues, we then normalized these expression values by a scaling procedure (Supplementary Note Figure 16). Specifically, among the genes with expression values in the inner quartile range, we identified the (1000) 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. We note that other normalization procedures resulted in very similar distributions.

As a further control of the normalization procedure, we analyzed the between-samples coefficient of variance (CV, defined as the ratio of the standard deviation over the mean) of the gene expression levels, before and after normalization. We analyzed separately housekeeping and non-housekeeping genes. The definition of human housekeeping genes was taken from She *et al.*[21], and is based on expression patterns estimated with microarrays for 42 normal tissues. By using this independent dataset, we avoid circularity issues, given that our normalization procedure also integrates a definition of housekeeping genes. We filtered this dataset to extract only genes which are present in RefSeq, and which are not annotated as "pseudogene'. We computed the CV for three datasets: 1) for the entire set of human protein-coding genes; 2) for the set of 5,636 amniote 1-1 orthologues and 3) for the set of 13,277 primate 1-1 orthologues. The CV was computed independently for each gene, among all available samples (*i.e.* all human samples for the first dataset, all amniote samples for the second dataset and all the primate samples for the third dataset).

As expected, we observed that the normalization procedure resulted in a significant reduction of the between-samples CV (Supplementary Note Figure 17). The difference between the CV computed before and after normalization is statistically significant for the three datasets (Wilcoxon rank sum test for paired data, p-value $< 10^{-10}$). We also observe that after normalization the CV is significantly lower for housekeeping genes than for non-housekeeping genes, for all three datasets (Wilcoxon rank sum test, p-value $< 10^{-10}$).

Moreover, we analyzed a set of 20 low-variance housekeeping genes, as defined by She *et al.*[21]. Only 17 of these genes could be found in Ensembl, after removing genes annotated as "pseudogene" or "processed_transcript". Although the very small sample size for this dataset prevents statistical testing, it appears that these independently-defined low-variance genes also have low CV in our dataset after the normalization procedure, slightly lower than the bulk of housekeeping genes (Supplementary Note Figure 17).

1.9 Gene expression phylogenies

This section is an extension of the Methods and of the Methods Summary and presents additional analyses for section "Mammalian gene expression phylogenies" in the main text.

We constructed expression trees with the neighbor joining approach, based on pairwise distance matrices between samples. The distance between samples was computed as 1-ρ, where ρ is Spearman's correlation coefficient (this measure was used because it is insensitive to outliers and potential data normalization inaccuracies); Euclidean distances were used as a control (Supplementary Figure 3). The neighbor-joining trees were constructed using functions in the `ape` package[22] 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.

In the gene expression phylogenies presented here, we have treated separately the individuals coming from the same species. The within-species variability in gene expression levels is thus directly represented in the trees. Nevertheless, we performed an additional jackknife-type resampling analysis to verify whether the robustness of the species branching patterns might be influenced by the inclusion of specific individuals in the distance matrices. To do this, we constructed all possible combinations of samples, wherein exactly one individual was considered for each species. We then computed the pairwise distance matrices for these resampled datasets (the distance was computed as 1-ρ, where ρ is Spearman's correlation coefficient) and constructed neighbor-joining trees. The consensus species trees are shown in Supplementary Note Figure 18 (the number of possible combinations are denoted N and are showed next to each organ tree). We found that the branch patterns are generally robust to individual selection, with notable exceptions within the ape group, where we inferred multifurcations for brain, liver and heart. Note that the same exceptions were also observed in our bootstrap analyses, where we found that the branching orders within the ape group were not robust. However, for the other three organs, the branching pattern of all species is perfectly consistent with the tree topology obtained with the entire set of samples which include all individuals from the same species (see Supplementary Figure 2 and Figure 1 in the main text).

1.10 Gene expression analyses with different annotation sets

This section presents additional analyses for section "Mammalian gene expression phylogenies" in the main text.

Most of our gene expression analyses were performed using the set of constitutive exons that we defined. To verify that this choice of annotation set did not bias in any way our conclusions, we redid the expression analyses using Ensembl annotations. For these verifications, we estimated the gene expression level through the mean per-base read coverage, averaged over the entire exonic (or constitutive) length of the gene, and we compared gene expression levels between samples and species with Spearman's correlation coefficient. The read mapping used for these controls correspond to the unambiguously mapping reads extracted from the read alignments provided by TopHat (section 1.3). We found that the correlation coefficients (ρ) are very similar when using Ensembl or constitutive exons (Supplementary Note Figure 19). In 75% of the comparisons we found higher or equal ρ values for constitutive exons than for Ensembl exons. This result confirms that our constitutive exon definition has succeeded in removing, at least in part, the (splicing-related) noise in the gene expression estimation.

Next, we constructed gene expression trees using the neighbour-joining approach, on distance matrices derived from Spearman's correlation coefficient (distance = 1-ρ - see also main text). We found that the tree topologies are highly similar for Ensembl exons and for constitutive exons (Supplementary Note Figures 20 to 22). The only differences in topology are found for heart and liver, where trees constructed with Ensembl exons appear to support a grouping between Opossum and Platypus, while the constitutive exons trees are in agreement with the known species phylogeny.

Finally, we compared the total tree lengths for the two annotation sets (Supplementary Note Figure 23). We found that the tree lengths are slightly lower for constitutive exons than for Ensembl exons, for all tissues. This is of course expected given our previous observation that correlation coefficients are higher when computed on constitutive exons. We note that tree lengths are nevertheless very similar for the two annotation sets; it is thus unlikely that the choice of the annotation set will affect our conclusions on the speed of gene expression evolution.

We also constructed trees for our annotation sets in which exon sequences are perfectly aligned, without gaps, among the ten amniote species (Supplementary Note Figure 24) or six primate species (Supplementary Figure 5, see section 1.6 for details). Branch lengths in these trees are very similar to the corresponding trees that are based on entire constitutive exon sequences of 1–1 orthologous genes (see Figure 1b and Supplementary Figure 2, main text for amniote trees; Supplementary Figure 5 for primate trees). Thus, the conclusions based on the amniote (all constitutive exons) trees (Figure 1b main text), such as the longer branches of apes and platypus relative to that of rodents, are robust to any potential effects introduced by exon sequence differences between species. Note that we chose to present and discuss the "all constitutive exon trees" in the main text with respect to amniotes (rather than those based on perfectly aligned exons), as the total exon length and hence total number of mapped RNA-Seq reads is significantly higher for the underlying exon set, which renders these trees more robust (as indicated by the generally higher bootstrap support). Given that the perfectly aligned exon set for primates is based on a substantially larger set of orthologues and therefore is very robust, the tree based on this exon set is presented and discussed in the main text.

1.11 Transcription modules

This section is an extension of the Methods, Methods Summary, and of section "Modular expression change and phenotypic evolution" in the main text. It provides additional information about the procedure used to detect transcription modules.

Transcription modules[23] were identified with the Iterative Signature Algorithm (ISA)[24], as implemented in the isa2 BioConductor package[25]. The normalized RPKM expression values of the 5,636 and 13,277 gene families constituted the input of the algorithm. Specifically, since the ISA works best with normally distributed data, the RPKM values were first transformed using the inverse hyperbolic sine transformation: $E_t = \log(E_r + \sqrt{E_r \cdot E_r + 1})$, where $E_r \cdot E_r$ denotes element-wise multiplication.

To allow summing up gene expression levels, we scaled them across samples, in order to obtain a mean of zero and a standard deviation equal to one. The centering of the data essentially corresponds to working with the relative expression changes of a gene, instead of the absolute ones. Unlike for previous analyses of expression data using the ISA, which were done with microarray data, we did not renormalize the data for each sample, because RNA-seq allows for direct comparison of gene expression levels within each sample.

The ISA is designed to identify, in an unsupervised manner, sets of genes that exhibit coherent expression patterns over subsets of samples from large sets of expression data. The algorithm starts with a so-called “seed” of random samples. It then selects all genes that are significantly over- or under-expressed across these samples. Subsequently all samples are scored by the weighted average expression levels across these genes. Over- and under-expressed genes have positive and negative weights, respectively. Samples receiving scores above a given significance threshold are selected and this double selection procedure is iterated for fixed thresholds until convergence to a fixed set of samples and genes is achieved. Such a combined and stable set is referred to as a transcription module. The final scores attributed to the samples are shown in Figure 3a of the main manuscript for specific modules. A collection of modules is generated by using a large number of seeds and different combinations of thresholds. Specifically, for the present analysis we ran the ISA independently for the amniote and the primate dataset, using the 121 combinations of the threshold values 1, 1.2, 1.4, . . ., 3 for both genes and samples.

We used a robustness measure (defined in Csárdi *et al.*[25]) to eliminate spurious modules. Robustness quantifies the correlation across module genes and across module samples. Modules that had a lower robustness score than that of spurious modules identified from randomized expression data were discarded. Our analysis pipeline resulted in 639 modules in the all-amniote and 197 modules in the primate-specific dataset.

Subsequently, Gene Ontology[26], KEGG[27] pathway, and chromosome enrichment calculations were performed using the hypergeometric test. The enrichment p-values were corrected using the Benjamini-Hochberg method[28] at a false discovery rate cutoff (FDR) of 5%.

Figure 3a in the main manuscript plots (extended) ISA samples scores for specific modules. For each module, the ISA assigns scores (real values between minus one and one) to the constituent genes and samples; other genes and samples have zero score by definition. Scores further away from zero indicate stronger correlation between the gene/sample and the rest of the module; two genes/samples having the same score sign are correlated, opposite signs indicate anticorrelation. The score of a gene exactly equals to its weighted mean expression across the module samples, the weights being the sample scores. Similarly, the score of a sample is the weighted mean of the expression of the module genes, the weights being the gene scores. Sample scores can be extended to samples not included in the module, by calculating the weighted average of the module gene expression for them. The extended sample scores can be used as the description of the population expression level of the module genes.

1.12 Tests for selection on expression levels

This section is an extension of the Methods, Methods Summary, and of section "Selectively driven expression change of individual genes and phenotypic evolution" in the main text. It provides additional information about the procedure used to infer the presence of significant shifts in expression levels for individual mammalian lineages.

An Ornstein Uhlenbeck (OU) process has previously been used to model the evolution of quantitative characters in the presence of natural selection[29,30]. It has been suggested as an appropriate model for the evolution of gene expression levels[31] as stabilizing selection is expected to maintain these levels about an optimum value. We build upon this previous work, extending it to incorporate sources of variation in gene expression level not due to the evolutionary history.

The Ornstein Uhlenbeck evolutionary model Given a phylogeny of known topology and branch lengths, we define r selective regimes acting on the phylogeny, each regime defined by an optimal expression level $\theta \in \{\theta_1, \ldots, \theta_r\}$. We model the selection of gene expression levels to these optima by assigning an OU process to each branch of the phylogeny. The OU processes are defined by parameters α , which characterizes the strength of selection, σ , which characterizes the phenotypic drift, and θ , the gene expression level that confers the optimal fitness. The same α and σ are common to entire phylogeny while each branch is assigned $\theta \in \{\theta_1, \ldots, \theta_r\}$ according to the selective regime operating on it. Let X_i be the state of the OU process at node i. X_i has a normal distribution with expectation and variance given by

$$[X_i] = E[X_a]e^{-\alpha t_{ia}} + (1 - e^{-\alpha t_{ia}})\theta_i \tag{1}$$

$$Var[X_i] = \frac{\sigma^2}{2\alpha}(1 - e^{-2\alpha t_{ia}}) + Var[X_a]e^{-2\alpha t_{ia}} \tag{2}$$

where X_a is the state of the OU process at node a, the parent node of node i, and $t_i a$ is the distance of node i from its parent. Given two nodes i and j with least common ancestor a

$$Cov[X_i, X_j] = Var[X_a]e^{-\alpha(t_{ia}+t_{ja})} \tag{3}$$

Let $\vec{X} = \begin{bmatrix} X_i \\ \vdots \\ X_N \end{bmatrix}$ be a vector describing the state of the OU process at the N terminal taxa. From 1, 2, and 3 it is clear that $\vec{X}$ has a multivariate normal distribution with expectation $E[\vec{X}] = \begin{bmatrix} E[X_i] \\ \vdots \\ E[X_N] \end{bmatrix}$ and covariance matrix entries $V_{ij} \in \mathbf{V}$ being given by 2 and 3. Given the multivariate normality of $\vec{X}$, the log likelihood function of the parameter values given a vector of observations $\vec{x}$ is given by

$$L(\alpha, \sigma, \theta_1, \ldots, \theta_r|\vec{x}) = -\frac{1}{2}log(\mathbf{V}) + -\frac{1}{2}[\vec{x} - E[\vec{X}]]'\mathbf{V}^{-1}[\vec{x} - E[\vec{X}]] \tag{4}$$

and maximum likelihood estimation can be used to estimate the parameters α,σ,θ_1,. . .,and θ_r.

Multiple observations of gene expression at terminal taxa Measurements of gene expression in multiple individuals of the same species are subject to two important sources of variation: differing biological conditions between individuals and measurement error. If multiple measurements at a terminal taxa are represented by a summary statistic such as the mean, this discounts important information present in the data about within species variation of gene expression levels. As a departure from previous methods, we assume observations of multiple individuals of the same species vary as a normal distribution, with a mean given by the underlying OU process and with variance ϵ, incorporating both measurement and biological variation common to each species of the phylogeny. Formally, at any leaf node X_i, we make k_i observations $X_{i1}, X_{i2}, \ldots, X_{ik_i}$ with $X_{ik} = X_i + N(O, \epsilon)$. The vector

$$\vec{X} = \begin{bmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1k_1} \\ \vdots \\ X_{21} \\ \vdots \\ X_{Nk_N} \end{bmatrix}$$

has a multivariate normal distribution with $E[X_{ik}] = E[X_i]$, $Var[X_{ik}] = Var[X_i] + \epsilon$, and $Cov[X_{ik}, X_{jl}] = Cov[X_i, X_j]$ where $i \neq j$. Maximum likelihood estimation of parameters α,σ,θ_1,. . .,θ_r, and ϵ can be performed as before.

Hypothesis testing To ask whether changes in optimal gene expression levels have occurred in particular lineages of the phylogeny, we test the null hypothesis in which all branches share the same optimum parameter θ_0 against the alternative hypothesis that a different optimum $\theta_1 \neq \theta_0$ acts on particular lineages. A Likelihood ratio test is used to assess the appropriateness of each model. As the null hypothesis is nested within the alternative hypothesis, the likelihood ratio statistic has an asymptotic $\chi^2(df = 1)$ distribution.

Evolutionary Hypotheses

Null Hypothesis	Alternative Hypothesis
$H_0 : \theta_0 = \theta_1$	$H_a : \theta_0 \neq \theta_1$

The analyses presented in the main text are based on the known phylogenetic relationships of the studied amniotes. Branch lengths in our analyses were based on divergence time estimates of previous studies[32,33,34,35]. Specifically, we assumed the following branch lengths (in million years): ((((((((hsa: 6, (ptr: 1.3, ppa: 1.3): 4.7): 1, ggo: 7): 7, ppy: 14): 11, mml: 25): 64, mmu: 89): 91, mdo: 180): 20, oan: 200): 110, gga: 310). However, analyses based on alternative proposed divergence estimates (e.g., divergence of therian mammals and platypus: 166 million years; placental mammals-marsupials: 148 million years) provide very similar results (data not shown).

We note that we only performed the test for genes that do not share exons with neighboring genes in any species according to our annotation, in order to avoid confounding effects of expression signals from these adjacent genes. Also, for each organ, tests were further restricted to genes expressed in all samples for that organ, given that the models compared in the test do not allow for genes with expression levels equal to zero in any of the organ samples. Thus, 3,909 genes among the 5,636 amniote orthologs and 9,969 genes among the 13,277 primate orthologs were considered for the test, respectively. Finally, we note that as the results depend on numerical optimization, there is always some chance that individual results may be affected by a failure to identify the global maximum. However, the likelihood ratio tests will tend to become more conservative because of this as optimization errors are more likely to affect the more parameter rich general model rather than the null model.

Lineage-specific selective constraint We hypothesized that gene expression levels experienced more evolutionary constraint in some lineages as compared to others. To test this hypothesis, we estimate for each gene the parameters of a model in which a different α acts on each of the primate, mouse, and platypus lineages. The Mann-Whitney U statistic is calculated for each pairwise comparison of the distribution of estimated α. A one sided p-value is calculated for the alternative hypothesis that the values of α for one lineage are likely to be higher then the values of α for another lineage. The results of these analyses show that α is significantly larger in the mouse lineage than in the primate and platypus lineages for all somatic tissues (largest p-value $< 10^{-11}$), which is consistent with the Neighbor-Joining tree analyses presented in the main text.

Gene Ontology enrichment analysis For a given ontology and evolutionary hypothesis, we ask if any GO terms are more likely to be associated with genes that show high amounts of evidence for this hypothesis. Given g GO terms, we construct sets G_1,. . .,G_g where G_i contains the log likelihood ratios of the genes associated with the i^{th} GO term. The Mann-Whitney U statistic for the i^{th} GO term is calculated for the sample G_i and the background sample $G_0 = \{GO_1, \ldots, GO_g\}$. A one sided p-value is calculated for the alternative hypothesis that the values of G_i are more likely to be higher then those of G_0. The Mann-Whitney U is calculated in R using wilcox.test and the GO term mapping is done in R using the biomaRt package[36]. The results of these GO analyses (overrepresented categories, p-values < 0.02) are shown in Supplementary Tables 27-42.

Lineage-specific expression shifts for housekeeping genes As a control for our test for lineage-specific expression shifts, we verified whether known housekeeping genes are under-represented in the significant test results, as compared to non-house keeping genes. Indeed, housekeeping genes are expected to experience more selective pressure to conserve their expression levels than non-housekeeping genes, given that they are likely involved in essential cellular functions. To perform this test, we used an independent definition of human housekeeping genes, based on expression patterns determined with microarrays for 42 human tissues[21]. We further filtered this dataset to extract only genes with valid RefSeq identifiers, and which were not annotated as "pseudogene" or "processed_ transcript" in Ensembl release 57. We observed a significantly smaller fraction of significant tests for housekeeping genes than for non-house keeping genes (Fisher's exact test, p-value 0.002467 for the amniote dataset and 3.977×10^{-6} for the primate dataset, Supplementary Note Table 12), indicating that our test performs as expected. However, we must note that a non-negligible proportion of housekeeping genes were present in the significant test results. This is not surprising, given that the definition of housekeeping genes that we used here does allow for variation of expression levels between samples, and only requires the genes to be expressed at significant levels in all tissues. Note that for these comparisons we considered only those genes that had non-null expression levels in all samples.

To strengthen these conclusions, we analyzed a list of 20 genes that display the least amount of variation among the 42 tissues, provided by She *et al.*[21], three of which are also used commercially as housekeeping controls for qPCR analyses (GAPDH, ACTB and UBC). As before, we filtered this set for genes with valid RefSeq identifiers, and which were not annotated as "pseudogene" or "processed_transcript" in Ensembl release 57, which left us with 17 genes. From these 17 genes, 4 had 1-1 orthologues in our primate dataset (*CALR, NONO, HNRNPD* and *EIF3H*), but none were present in the 1-1 amniote orthologue dataset. None of the 172 tests performed for the 4 genes was significant at the 0.05 FDR threshold. On average, for the primate dataset, 1.06% (4638 out of 438764) of the tests were significant, and thus for 172 tests the average expectation is of 1.81 significant tests, for 0 observed. The difference in proportions is not significant (Fisher's exact test, p-value 0.43), due to the very low sample size, but the absence of significant expression shifts for these low-variation genes is reassuring for the validity of our method.

1.13 Differential expression between male and female individuals

This section presents additional analyses regarding the detection of sex-differences in gene expression levels.

Our somatic organ data are derived from both male and female individuals and thus offer an opportunity to investigate sex-biased gene expression across amniotes. We screened for statistically significant expression differences between the two sexes in each species. To detect significant expression differences between male and female individuals, we used the DESeq method[37], which is based on read count data and is implemented as an R/Bioconductor package. Note that for species/tissues for which multiple individuals were available from one sex (e.g., human brain), all reads from these individuals were pooled in the framework of our analyses.

We thus identified a total of 4,990 candidate cases. Given that in most cases we only have data for one individual per sex and tissue in each species, we cannot generally distinguish between general inter-individual variation and true sex-biased expression. Therefore, we only report selected cases that consistently differ between sexes across at least two species, can directly be linked to sex-specific functions, are located on hemizygous sex chromosomes, and/or for which sex-biased expression was previously reported in humans or mouse (Supplementary Table 3).

A particularly intriguing sex-biased gene that we identified is found in the egg-laying platypus. Nutritional reserves that are stored in egg yolk are crucial for embryonic development in non-mammalian egg-laying vertebrates. These reserves are nearly entirely derived from vitellogenin, an extremely versatile protein that is produced in the liver. It was previously suggested that monotremes, the only egg-laying mammals, have retained one of three ancestral vitellogenin genes, whereas all other mammals have progressively lost these major egg yolk genes during evolution[38]. We find that the predicted platypus vitellogenin gene is indeed transcribed at very high levels in liver from female platypus, whereas almost no transcription can be detected in male liver (Supplementary Figure 4 and Supplementary Table 3). We observed a similar female biased expression pattern for the orthologous vitellogenin genes in chicken (Supplementary Figure 4). Thus, our results highlight the intriguing mammal-bird/reptile crossover character of monotremes.

1.14 Comparison of expression level estimates with strand-specfic RNA-Seq protocols

This section presents an additional control for expression level estimates, by comparing estimates of gene expression levels between non-strand-specific and strand-specific RNA-Seq protocols.

In the present manuscript, we have used a non-strand-specific RNA-Seq protocol, *i.e.* the sequenced reads can come from either the sense or the anti-sense strand of mRNAs. This protocol was state-of-the-art when our dataset was generated, but a strand-specific protocol has been recently made available by Illumina. In order to verify that our RNA-Seq data provide a correct estimate of the gene expression level, despite the absence of strand information, we used the new strand-specific protocol to re-sequence one of our samples (human brain, from a male individual). We estimated gene expression levels using unambiguously mapping reads obtained with TopHat, on Ensembl-annotated protein coding genes. Here, we measured gene expression levels as the log2-transformed mean read coverage, computed on Ensembl-annotated exons. Only protein-coding transcripts were considered for the annotation. For the strand-specific sample, we took into account only reads that mapped on the sense strand of the genes.

As shown in Supplementary Note Figure 25, the two expression levels estimates correlate very well (Spearman's correlation coefficient $\rho = 0.95$), despite the differences in the library preparation. Moreover, in the strand-specific sample, we find that most of the reads come from the sense strand of the protein-coding genes (Supplementary Note Figure 25) - in total, we find that only 1.2% of the mapped reads come from the anti-sense strand of the genes. This means that the extent of anti-sense transcription is very limited. In Ensembl annotations, as well as in our extended annotations, it is possible for genes to overlap; an overlap is observed for approximately 5% of protein-coding genes. In this cases, strand-specific RNA-Seq data would indeed help to distinguish the expression patterns of the two overlapping genes. Note that for the analysis of the expression patterns of individual genes (such as the search for significant shifts in expression levels) we removed overlapping genes, and thus the lack of strand-specificity does not influence our conclusions.

1.15 Influence of total read coverage on the detection of transcribed protein-coding genes

This section presents an additional analysis of the power of detecting transcribed genes, as function of the total number of mapped RNA-Seq reads.

We next wanted to verify whether the total read coverage that is available for each sample can influence in our conclusions. To do this, we constructed two large RNA-Seq samples by pooling the reads coming from all the individuals, for mouse brain and liver. We thus obtained approximately 55 million mapped reads for each of the organs. We then resampled 5, 10, 20, 30, 40 and 50 million reads from these mapped reads, and computed the number of protein-coding genes that were detected as transcribed (RPKM>0). To estimate the number of genes that would be detected as transcribed if there were no read coverage limitations, we fitted a model of the form $y = \frac{a}{1+(1/(b*x+c)}$ to the observed distribution, where x is the number of mapped reads and y is the number of protein-coding genes detected as transcribed.

The relationship between the number of detected protein-coding genes and the number of mapped reads is summarized in Supplementary Note Figure 26. Given that in our samples the number of mapped reads is generally elevated (greater than 10 millions in 92% of the cases and greater than 15 millions in 52% of the cases), we are confident that these data allow us to get a broad view of protein-coding genes transcription - although of course increasing read coverage can only improve the sensitivity of the analyses.

Furthermore, we note that for the majority of analyses, we focused on two sets of genes: first, a set of 5,636 genes that have 1-1 orthologues in all the species in our dataset, and second, a set of 13,277 genes that have 1-1 orthologues in the primate species. As shown in Supplementary Note Figure 27 for a set of human samples deriving from the 6 different tissues, these genes are more highly expressed than the bulk of protein-coding genes (Wilcoxon rank sum test, p-value $< 10^{-10}$). The power of detection and expression level quantification is thus greater for these two sets of genes than for the other protein-coding genes. It is thus unlikely that our conclusions will be strongly influenced by the inclusion of genes with low expression levels, for which the expression levels estimation is more noisy.

2 Supplementary Note Tables

Sample id	Species	Tissue	Illumina pipeline	Nb. reads[a]	% N bases[b]		% aligned reads[c]	
					Illumina	Ibis	Illumina	Ibis
288	Human	Cerebellum	GA 1.40	23,214,752	1.50%	1.47%	44.2%	50.0%
588	Macaque	Brain	GA 1.40	22,554,234	0.93%	0.89%	47.7%	49.5%
583	Mouse	Cerebellum	GA 1.60	41,340,785	1.95%	1.95%	47.5%	47.9%
475	Opossum	Testis	GA 1.32	15,293,069	0.98%	0.55%	49.4%	51.5%
487	Platypus	Brain	GA 1.51	24,343,340	0.58%	0.54%	52.0%	53.0%
554	Chicken	Liver	GA 1.60	22,542,615	3.33%	3.33%	42.8%	42.9%

[a] Total number of raw reads for this sample.

[b] Percentage of bases with ambiguous calling in the raw reads.

[c] Percentage of reads that align with at most 3 mismatches on the reference genome sequence. The read alignment was performed with bowtie, in the "-v " mode, which does not take into account the quality scores for the read mapping.

Supplementary Note Table 1: Statistics for the performance of the Ibis base caller and comparison with the standard Illumina pipeline, for a subset of 6 samples.

Species	Parental genes	% detected (no masking)	% detected (masking)
Human	2254	55%	64%
Chimpanzee	1993	55%	66%
Gorilla	1852	48%	57%
Orangutan	1796	48%	61%
Macaque	1775	52%	61%
Mouse	2325	60%	69%
Opossum	2049	43%	54%
Platypus	170	20%	21%
Chicken	210	36%	37%

Supplementary Note Table 2: Junction detection sensitivity for parental genes, before and after masking the retrocopies. The numbers represent the proportion of annotated junctions that are detected by TopHat, for one brain sample, from a male individual for each species.

Species	Nb. genes[a]	Nb. annotated[b]	Nb. TopHat[c]	Nb. tot. confirmed[d]
Human	21,337	261,029	193,704 (74%)	208,156 (80%)
Chimpanzee	19,829	199,418	157,583 (79%)	165,317 (83%)
Gorilla	20,803	193,525	139,026 (72%)	144,755 (75%)
Orangutan	20,009	180,719	128,112 (71%)	136,588 (76%)
Macaque	21,905	201,455	146,526 (73%)	151,330 (75%)
Mouse	23,062	222,696	183,439 (82%)	190,734 (86%)
Opossum	19,466	189,389	130,066 (69%)	133,774 (71%)
Platypus	17,951	157,899	90,375 (57%)	94,569 (60%)
Chicken	16,736	156,050	119,161 (76%)	126292 (81%)

[a] Total number of protein-coding genes.

[b] Number of splice junctions annotated in Ensembl.

[c] Number (percentage) of Ensembl-annotated splice junctions that were confirmed with TopHat.

[d] Total number (percentage) of Ensembl-annotated splice junctions confirmed after our multi-splice validation procedure.

Supplementary Note Table 3: Splice junctions: comparison between Ensembl annotations, TopHat results and our multi-splice validation procedure. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Nb. genes[a]	Gene length[b]			Exon length[c]			Splice junctions[d]			NR[e]
		Ensembl	Refined	% increase[f]	Ensembl	Refined	Intersect[g]	Ensembl	Refined	Intersect[h]	
Human	21,337	1,266 Mb	1,300 Mb	3%	76.4 Mb	78.1 Mb	69.7 Mb	261,029	320,437	208,156	225
Chimpanzee	19,829	1,088 Mb	1,169 Mb	7%	49.0 Mb	61.4 Mb	47.3 Mb	199,418	290,738	165,317	199
Gorilla	20,803	883 Mb	977 Mb	11%	47.5 Mb	57.4 Mb	46.3 Mb	193,525	242,054	144,755	163
Orangutan	20,009	946 Mb	1,014 Mb	7%	36.7 Mb	48.0 Mb	36.2 Mb	180,719	205,491	136,588	131
Macaque	21,905	997 Mb	1,103 Mb	11%	43.9 Mb	55.6 Mb	42.4 Mb	201,455	250,951	151,330	156
Mouse	23,062	1,009 Mb	1,056 Mb	5%	67.8 Mb	75.1 Mb	63.0 Mb	222,696	303,467	190,734	278
Opossum	19,466	996 Mb	1,201 Mb	21%	32.3 Mb	53.9 Mb	31.7 Mb	189,389	228,801	133,774	158
Platypus	17,951	376 Mb	477 Mb	27%	23.8 Mb	34.5 Mb	23.4 Mb	157,899	161,307	94,569	137
Chicken	16,736	426 Mb	513 Mb	20%	30.7 Mb	43.6 Mb	29.9 Mb	156,050	206,618	126,292	146

[a] Total number of protein-coding genes, based on Ensembl annotations, release 57. We excluded the genes that are found on haplotypic regions.

[b] Total length (Mb) of the protein-coding genes, as determined from Ensembl annotations and from our refined annotations.

[c] Total exon length, as determined from Ensembl annotations and from our refined annotations. For Ensembl annotations, we excluded transcripts annotated as "retained_intron" before defining the exon blocks and computing the exon length.

[d] Total number of splice junctions within known protein-coding genes, as determined from Ensembl anntations and from our refined annotations.

[e] Number (millions) of unambiguously mapping reads available for each species, used as a basis for the annotation refinement procedure.

[f] Relative increase in gene length.

[g] Total length of the intersection between the two annotation sets.

[h] Number of splice junctions found in both annotation sets.

Supplementary Note Table 4: Comparison between our refined annotations and Ensembl 57 annotations. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Nb. genes [a]	Ensembl exons[b]	Confirmed [c]	Restricted [d]	Extended[e]	Non-confirmed[f]	New exons[g]
Human	21,337	226,127	189,051	12,147	26,419	565	20,875
Chimpanzee	19,829	207,562	170,904	6,233	31,212	671	33,365
Gorilla	20,803	203,334	170,813	7,725	25,614	591	27,092
Orangutan	20,009	196,377	163,946	2,990	29,873	257	19,971
Macaque	21,905	208,028	168,350	9,599	30,230	1,441	32,051
Mouse	23,062	216,899	180,520	8,299	29,622	433	24,546
Opossum	19,466	187,724	142,184	14,425	31,702	1,075	34,539
Platypus	17,951	162,667	127,250	10,296	26,318	486	22,563
Chicken	16,736	167,137	134,505	8,352	24,978	596	29,642

[a] Total number of protein-coding genes, based on Ensembl annotations, release 57. We excluded the genes that are found on haplotypic regions.

[b] Total number of Ensembl-annotated exon blocks in protein-coding genes. Only transcripts annotated as "protein_coding" were considered for the computation of the Ensembl exon blocks (with one exception, see below).

[c] Ensembl-annotated exons for which the exact boundaries were confirmed by our annotations.

[d] Ensembl-annotated exons for which at least one boundary is not included in our annotations.

[e] Ensembl-annotated exons for which at least one boundary is extended in our annotations.

[f] Ensembl-annotated exons that do not overlap at all with our annotations.

[g] Exons in our annotations that do not overlap with Ensembl-annotated exons. For this comparison, all Ensembl-annotated transcripts were considered.

Supplementary Note Table 5: Comparison between exon blocks defined based on Ensembl annotations and based on our refined annotations. Note that the categories "restricted boundaries" and "extended boundaries" are not mutually exclusive, *i.e.* one Ensembl exon can have one boundary restricted and the other one extended in our annotations. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Nb. genes	All transcribed blocks[a]			Low-coverage segments[b]		
		Nb. blocks	Length	Close to splice sites[c]	Nb. segments	Length	Close to splice sites[d]
Human	21,337	236,511	107 Mb	140,174 (59%)	297,020	38 Mb	234,936 (79%)
Chimpanzee	19,829	218,853	87 Mb	121,670 (56%)	300,583	32 Mb	237,648 (79%)
Gorilla	20,803	210,851	81 Mb	110,656 (52%)	271,611	29 Mb	213,039 (78%)
Orangutan	20,009	200,324	66 Mb	113,627 (56%)	238,711	22 Mb	184,156 (77%)
Macaque	21,905	220,957	78 Mb	118,703 (54%)	295,197	27 Mb	236,311 (80%)
Mouse	23,062	217,279	116 Mb	112,276 (52%)	323,142	49 Mb	260,044 (80%)
Opossum	19,466	206,943	73 Mb	124,497 (60%)	280,990	24 Mb	232,945 (83%)
Platypus	17,951	179,330	41 Mb	113,499 (63%)	207,965	10 Mb	166,713 (80%)
Chicken	16,736	182,343	61 Mb	107,494 (59%)	256,135	21 Mb	205,748 (80%)

[a] Statistics for the extended gene models, that incorporate both Ensembl annotations and RNASeq information. Note that this annotation set includes retained introns (hence the increased length as compared to that of refined exon blocks).

[b] Statistics for the segments defined by `segclust` that were classified as "non-constitutive" because of their relatively low read coverage.

[c] Number (percentage) of transcribed blocks for which at least one boundary is within 10 bp of a splice junction extremity.

[d] Number (percentage) of "non-constitutive" segments for which at least one boundary is within 10 bp of a splice junction extremity.

Supplementary Note Table 6: Statistics for the first step in our procedure for definition of constitutive exons: identification of low-coverage transcribed regions with a segmentation/clustering algorithm. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Extended annot[a]	Constitutive exons		Intersect Ensembl[b]			Intersect refined exons[c]		
		Length[d]	% extended[e]	Length	% const. [f]	% Ensembl[g]	Length	% const[h]	% refined [i]
Human	107.5 Mb	55.6 Mb	51.6%	51.4 Mb	92.7%	72.5%	52.6 Mb	94.7%	67.4%
Chimpanzee	87.4 Mb	44.7 Mb	51.2%	38.6 Mb	86.4%	78.8%	42.3 Mb	94.7%	69.0%
Gorilla	80.8 Mb	43.1 Mb	53.3%	37.6 Mb	87.2%	79.1%	40.9 Mb	94.9%	71.2%
Orangutan	66.1 Mb	38.0 Mb	57.5%	30.7 Mb	80.8%	83.6%	36.3 Mb	95.4%	75.7%
Macaque	77.6 Mb	41.6 Mb	53.7%	35 Mb	84.2%	79.8%	39.9 Mb	96.0%	71.8%
Mouse	116.2 Mb	55.0 Mb	47.4%	49.9 Mb	90.7%	77.6%	52.1 Mb	94.7%	69.4%
Opossum	73.0 Mb	41.6 Mb	57.0%	28.2 Mb	67.7%	87.3%	40.4 Mb	97.2%	75.1%
Platypus	40.9 Mb	26.8 Mb	65.7%	19.8 Mb	73.7%	83.0%	26.2 Mb	97.6%	75.9%
Chicken	61.2 Mb	33.1 Mb	54.1%	25.7 Mb	77.6%	83.6%	31.7 Mb	96.0%	72.9%

[a] Total length of the exon blocks determined with our annotation extension procedure, for known protein-coding genes. N.B. these annotations include potential retained introns; they were used as a basis for the definition of the constitutive exons.

[b] Intersection between constitutive exon blocks and Ensembl exon blocks. For Ensembl annotations, only transcripts annotated as protein-coding were taken into account for the definition of the exon blocks.

[c] Intersection between constitutive exon blocks and the coordinates of the exon blocks determined with our annotation refinement procedure.

[d] Total length of the constitutive exon blocks, for the known protein-coding genes of each species (percentage of the total transcribed length).

[e] Percentage of the total length of the extended annotations represented by the constitutive exons.

[f] Percentage of the length of the constitutive exons.

[g] Percentage of the length of the Ensembl-annotated exons.

[h] Percentage of the length of the constitutive exons.

[i] Percentage of the length of the refined exons.

Supplementary Note Table 7: Statistics for constitutive exons: overlap with our extended annotations and with Ensembl annotations. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Nb. of genes[a]	Mean read coverage (log_2 scale)[b]		Variance within gene[c]	
		Ensembl	Constitutive	Ensembl	Constitutive
Human	13,948	1.35	1.55	0.47	0.36
Chimpanzee	12,902	1.75	1.96	0.49	0.38
Gorilla	12,586	1.69	1.96	0.50	0.38
Orangutan	12,277	1.58	1.76	0.44	0.38
Macaque	12,804	1.62	1.82	0.47	0.34
Mouse	13,544	1.91	2.14	0.44	0.32
Opossum	11,882	1.76	1.96	0.37	0.31
Platypus	9,605	1.87	2.32	0.59	0.34
Chicken	10,334	2.06	2.3	0.49	0.39

[a] Number of Ensembl-annotated protein-coding genes that have at least 5 exon blocks in Ensembl annotations, and at least 5 constitutive exon blocks, based on our definition of constitutive regions. For Ensembl annotations, only transcripts annotated as protein-coding were taken into account for the definition of the exon blocks.

[b] Average per-base read coverage, transformed on a log_2 scale, as computed on Ensembl-annotated exon blocks and on our constitutive exon blocks. We compute the average value for each gene; the value presented here is the median over all genes and over all RNASeq samples.

[c] Variance of the per-base read coverage, transformed on a log_2 scale, within exon blocks of the same gene, for Ensembl-annotated exon blocks and for our constitutive exon blocks. The value presented here is the median over all genes and over all RNASeq samples.

Supplementary Note Table 8: Measuring expression levels on Ensembl-annotated exons and on "constitutive" exon blocks. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Constitutive[a]	Constitutive aligned[b]	
		Length	Fraction[c]
Human	18.4 Mb	4.1 Mb	22.4%
Chimpanzee	16.0 Mb	4.1 Mb	25.8%
Gorilla	15.5 Mb	4.1 Mb	26.6 %
Orangutan	14.1 Mb	4.1 Mb	29.3 %
Macaque	14.4 Mb	4.1 Mb	28.5 %
Mouse	18.3 Mb	4.1 Mb	22.4 %
Opossum	15 Mb	4.1 Mb	27.4 %
Platypus	10 Mb	4.1 Mb	37.6 %
Chicken	13.8 Mb	4.1 Mb	29.9 %

Species	Constitutive[d]	Constitutive aligned[e]	
		Length	Fraction[f]
Human	39.2 Mb	17.4 Mb	44.3 %
Chimpanzee	33.7 Mb	17.4 Mb	51.5 %
Gorilla	33.2 Mb	17.4 Mb	52.3 %
Orangutan	29.8 Mb	17.4 Mb	58.2 %
Macaque	30.9 Mb	17.4 Mb	56.2 %

[a] Total length of the constitutive exon blocks for the 5,636 genes that have 1-1 orthologues in all species.

[b] Perfectly aligned constitutive exons.

[c] Fraction with respect to the constitutive exon length of the 5,636 1-1 orthologues genes, within each species.

[d] Total length of the constitutive exon blocks for the 13,277 genes that have 1-1 orthologues in all primate species.

[e] Perfectly aligned constitutive exons.

[f] Fraction with respect to the constitutive exon length of the 13,277 1-1 orthologues genes, within each species.

Supplementary Note Table 9: Statistics for the constitutive aligned exons. Two datasets are presented: one with 5,636 genes that are 1-1 orthologues for all pairs of species in our dataset, and the other with 13,277 genes that are 1-1 orthologues for all primate species. This analysis was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Nb. loci[a]		Total length[b]		Nb. exons[c]		Exonic length[d]	
	All loci	Intergenic	All loci	Intergenic	All loci	Intergenic	All loci	Intergenic
Human	56,706	10,993	1,605 Mb	84.6 Mb	360,109 (6.4)	26,278 (2.4)	107.1 Mb	3.5 Mb
Chimpanzee	60,572	19,668	1,609 Mb	174.1 Mb	359,854 (5.9)	50,846 (2.6)	107.5 Mb	8.4 Mb
Gorilla	56,734	16,844	1,400 Mb	178.0 Mb	318,790 (5.6)	44,515 (2.6)	87.7 Mb	7.6 Mb
Orangutan	49,975	13,697	1,185 Mb	119.7 Mb	270,291 (5.4)	37,220 (2.7)	68.4 Mb	6.2 Mb
Macaque	58,137	19,631	1,451 Mb	176.1 Mb	334,688 (5.8)	52,984 (2.7)	91.3 Mb	9.0 Mb
Mouse	53,496	13,380	1,376 Mb	98.4 Mb	346,455 (6.5)	34,305 (2.6)	117.1 Mb	5.2 Mb
Opossum	44,451	12,784	1,544 Mb	167.8 Mb	284,833 (6.4)	35,650 (2.8)	79.9 Mb	6.7 Mb
Platypus	54,988	23,237	742.4 Mb	143.3 Mb	257,929 (4.7)	62,529 (2.7)	52.6 Mb	10.0 Mb
Chicken	44,599	14,150	752.6 Mb	79.3 Mb	270,281 (6.1)	37,005 (2.6)	72.9 Mb	6.0 Mb

[a] Number of multi-exonic transcribed loci.

[b] Total length (Mb) of the multi-exonic transcribed loci.

[c] Total number of exons for the multi-exonic transcribed loci (mean number per locus). The exons were defined with our procedure for refinement of exon coordinates.

[d] Total exonic length for the multi-exonic transcribed loci.

Supplementary Note Table 10: Statistics for the multi-exonic transcribed loci detected with our RNASeq data. The "intergenic" class corresponds to transcribed loci that do not overlap with any annotated feature (including non-coding RNAs, pseudogenes etc.), as defined in Ensembl release 57. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

Species	Protein-coding genes[a]				Long non-coding RNAs[b]				Pseudogenes[c]			
	Nb.	> 0[d]	≥ 50%[e]	100% [f]	Nb.	> 0	≥ 50%	100%	Nb.	> 0	≥ 50%	100%
Human	19,699	18,053 (92)	17,394 (88)	9,863 (50)	1,048	623 (59)	597 (57)	374 (36)	2,668	697 (26)	544 (20)	229(9)
Chimp	18,651	16,940 (91)	16,421 (88)	9,219 (49)	0	NA	NA	NA	409	111 (27)	101 (25)	62(15)
Gorilla	19,193	16,479 (86)	15,703 (82)	8,717 (46)	0	NA	NA	NA	1,405	204 (15)	179 (13)	104(7)
Orangutan	17,713	15,101 (85)	14,195 (80)	7,596 (43)	0	NA	NA	NA	1,018	143 (14)	120 (12)	69(7)
Macaque	19,818	16,880 (85)	16,256 (82)	9,176 (47)	0	NA	NA	NA	1,756	210 (12)	196 (11)	134(8)
Mouse	20,188	17,713 (88)	17,170 (85)	11,865 (59)	495	395 (80)	359 (73)	182 (37)	1,230	347(28)	305 (25)	190 (15)
Opossum	17,349	15,250 (88)	14,609 (84)	8,696 (50)	0	NA	NA	NA	718	117 (16)	103 (14)	73(10)
Platypus	16,942	14,664 (87)	13,579 (80)	5,156 (30)	0	NA	NA	NA	180	39 (22)	38 (21)	28(16)
Chicken	15,729	14,115 (90)	13,645 (87)	8,479 (54)	0	NA	NA	NA	95	28 (29)	27 (28)	16(17)

[a] Multi-exonic protein-coding genes annotated in Ensembl 57. Genes found on haplotypic regions were excluded.

[b] Multi-exonic long non-coding RNA genes (lincRNAs) annotated in Ensembl 57. Genes found on haplotypic regions were excluded.

[c] Multi-exonic pseudogenes annotated in Ensembl 57. Genes found on haplotypic regions were excluded.

[d] Number (percentage) of annotated genes that have at least one exon block found in multi-exonic transcribed loci.

[e] Number (percentage) of annotated genes that have at least half of their exon blocks found in multi-exonic transcribed loci.

[f] Number (percentage) of annotated genes with perfect overlap with multi-exonic transcribed loci (*i.e.* all exon blocks are found, and all in a single locus).

Supplementary Note Table 11: Intersection between classes of genes annotated in Ensembl 57 and the multi-exonic transcribed loci detected with our RNASeq data. This comparison was performed only for the species for which we applied our annotation refinement procedure, *i.e.* the bonobo was excluded, given that its genome sequence was not publicly available at the time when these analyses were performed.

a) All-amniote dataset

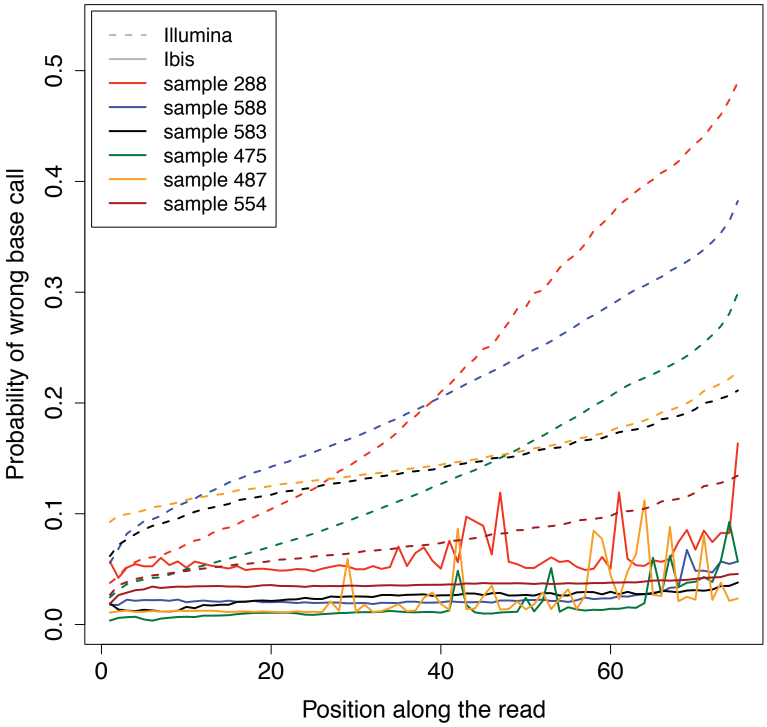
Gene class	Significant tests	Non-significant tests
housekeeping genes	185	125
Non-housekeeping genes	1643	760

b) Primate dataset

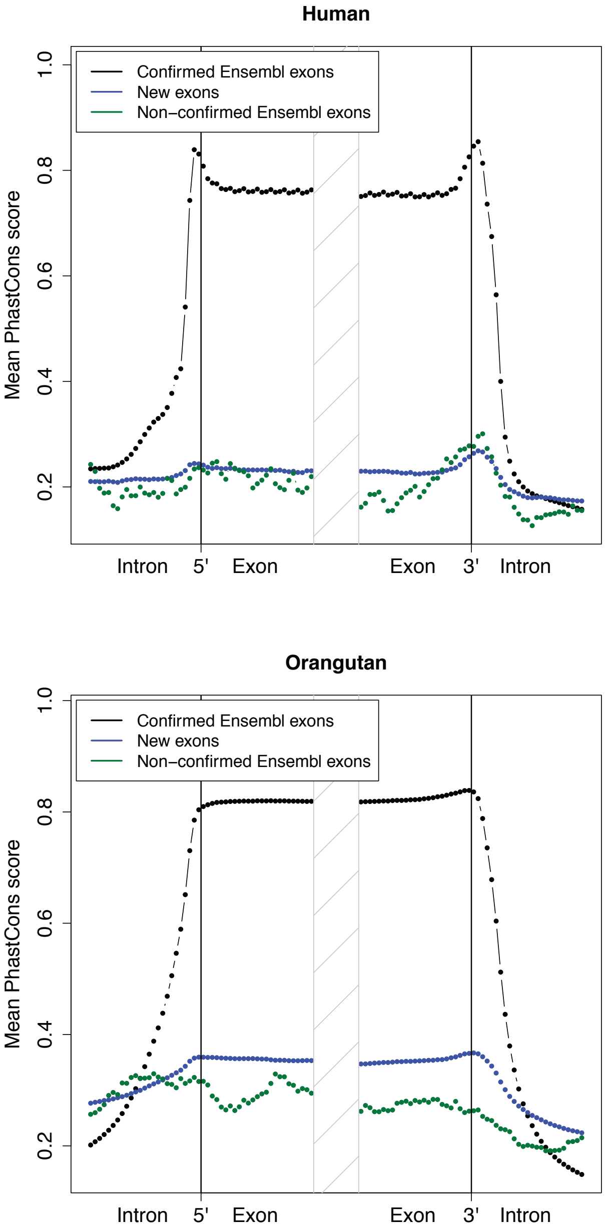
Gene class	Significant tests	Non-significant tests
housekeeping genes	217	575
Non-housekeeping genes	2146	3885

Supplementary Note Table 12: Tests for lineage-specific expression shifts, for housekeeping and non-housekeeping genes. In the "significant tests" column, we count the number of genes for which at least one test was significant (p-value < 0.05 after multiple testing correction with the Benjamini-Hochberg method), in at least one lineage and one tissue. Conversely, the "non-significant tests" column contains only genes for which none of the tests was significant. For these comparisons, we considered only genes which had non-null expression levels in all of the samples (which could thus be tested for all the tissues), and which did not overlap with any other genes in the genome (to avoid biases resulting from the expression patterns of the neighboring genes).

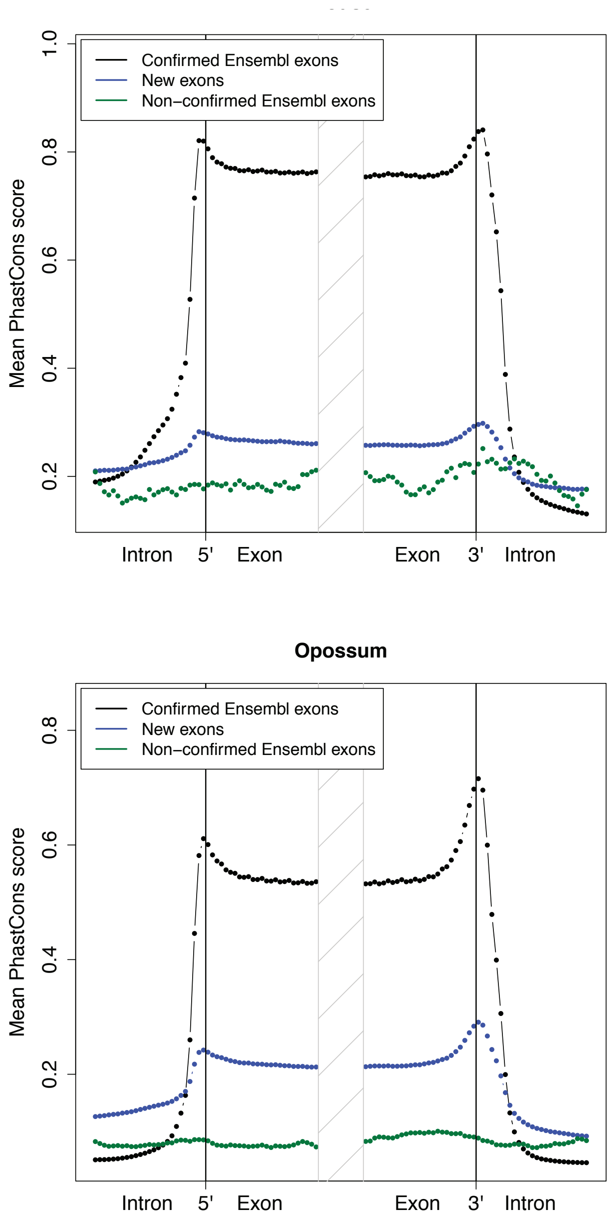
3 Supplementary Note Figures



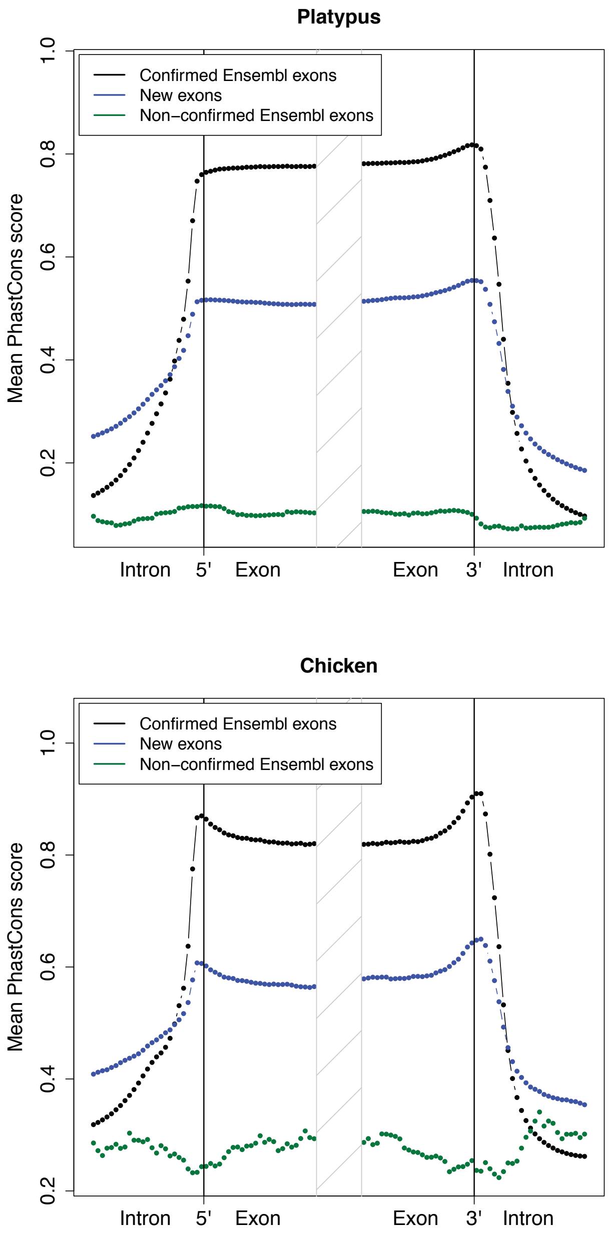
Supplementary Note Figure 1: Variation of the base calling error rate along the read length for a subset of 6 samples, for the Illumina standard base caller and for the Ibis base caller. The probability of error was deduced from the per-base quality score, and then averaged on all reads.



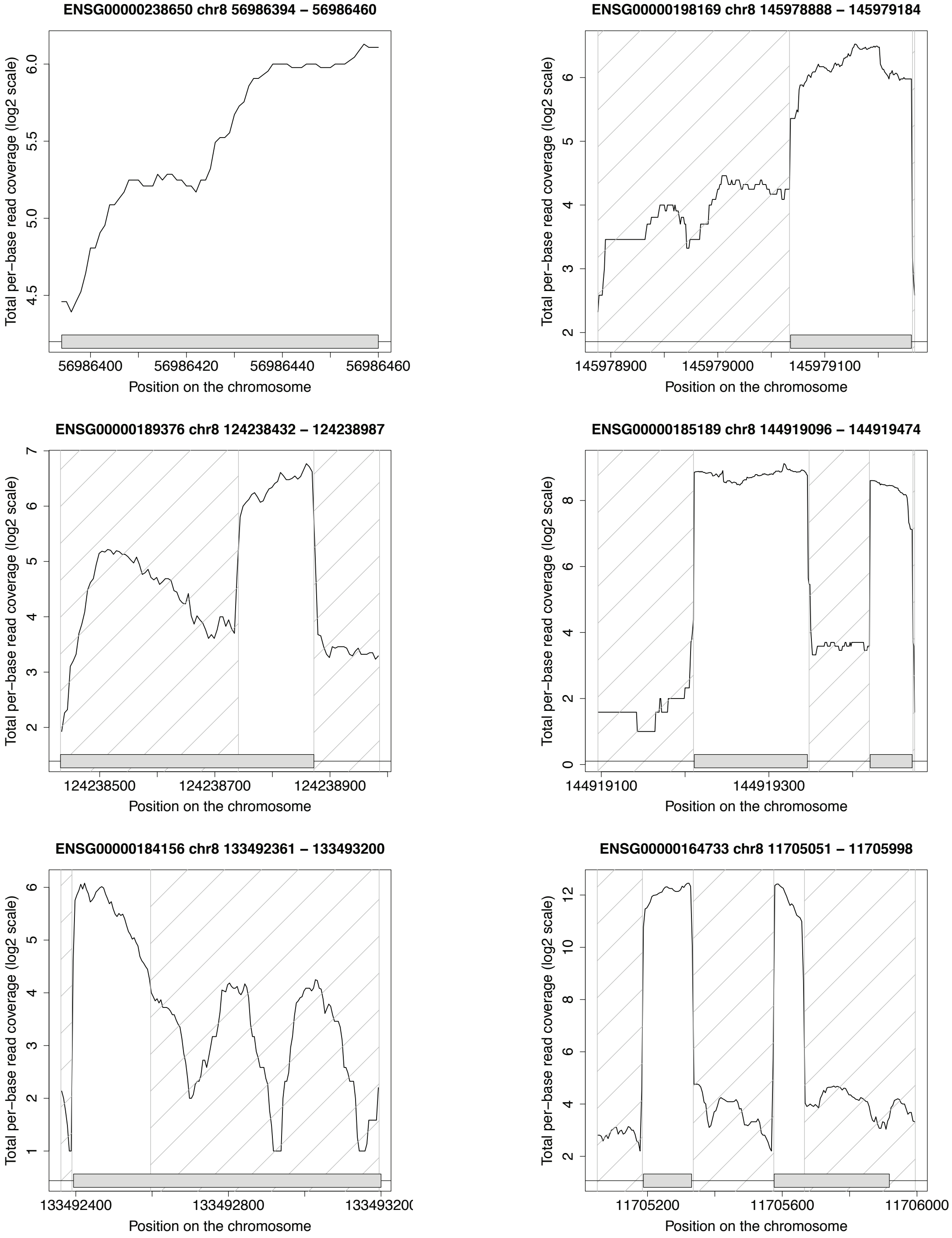
Supplementary Note Figure 2: PhastCons score variation around exon boundaries, for three classes of exons: Ensembl exons confirmed by our annotations (black), Ensembl exons not found in our annotations (green), and new exons added by our annotations (blue). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the human genome; bottom panel: annotations for the orangutan genome.



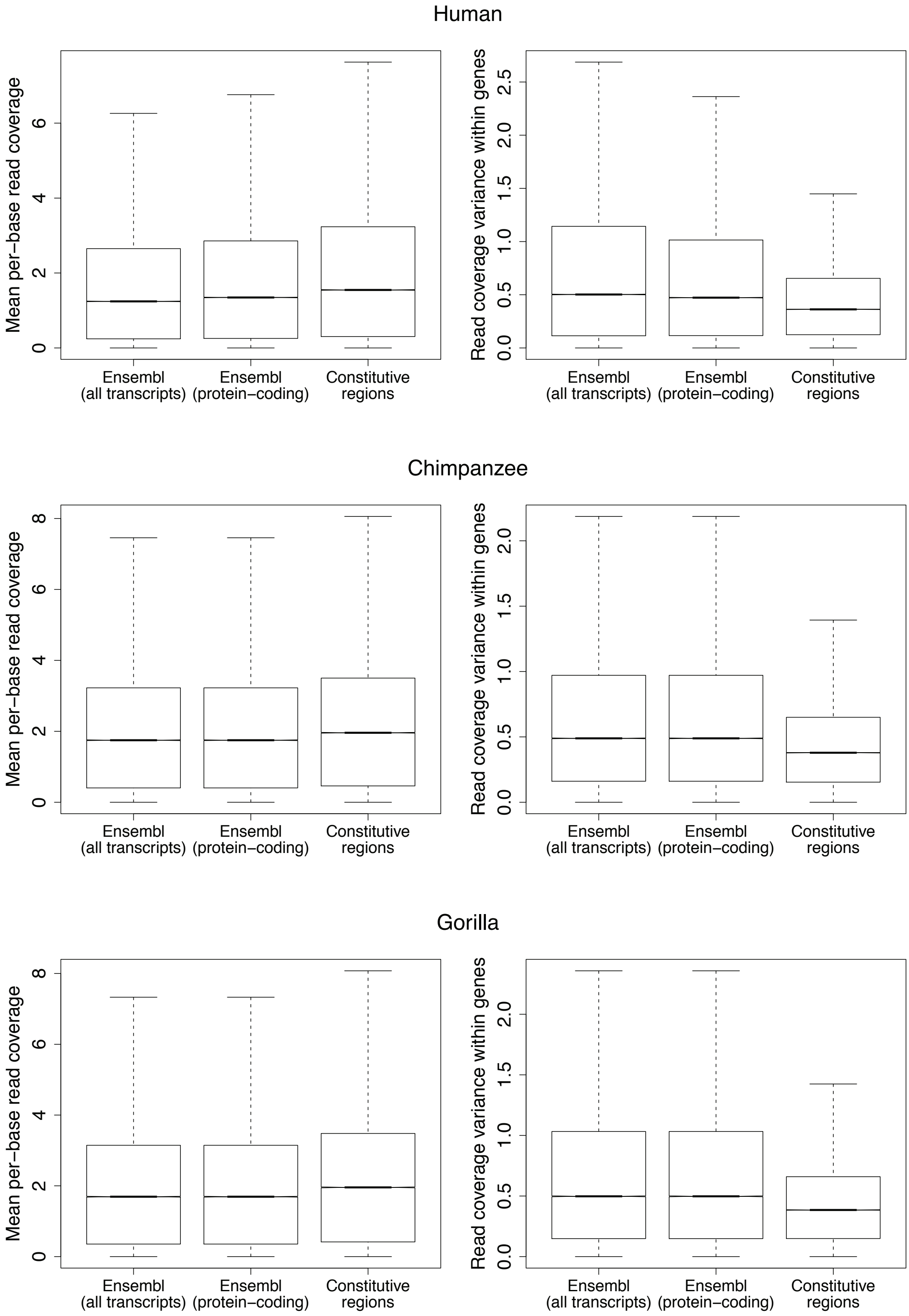
Supplementary Note Figure 3: PhastCons score variation around exon boundaries, for three classes of exons: Ensembl exons confirmed by our annotations (black), Ensembl exons not found in our annotations (green), and new exons added by our annotations (blue). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the mouse genome; bottom panel: annotations for the opossum genome.



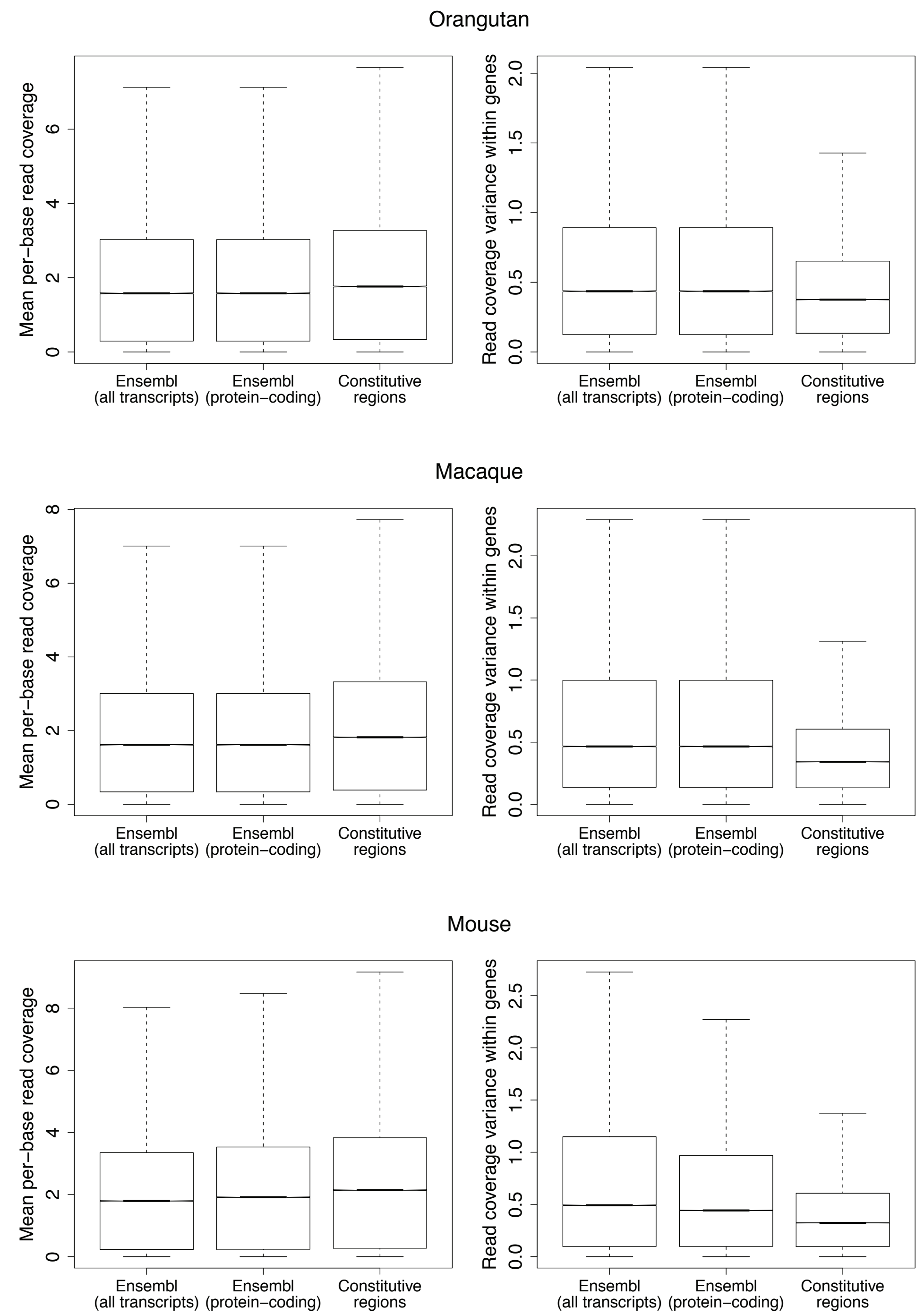
Supplementary Note Figure 4: PhastCons score variation around exon boundaries, for three classes of exons: Ensembl exons confirmed by our annotations (black), Ensembl exons not found in our annotations (green), and new exons added by our annotations (blue). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the platypus genome; bottom panel: annotations for the chicken genome.



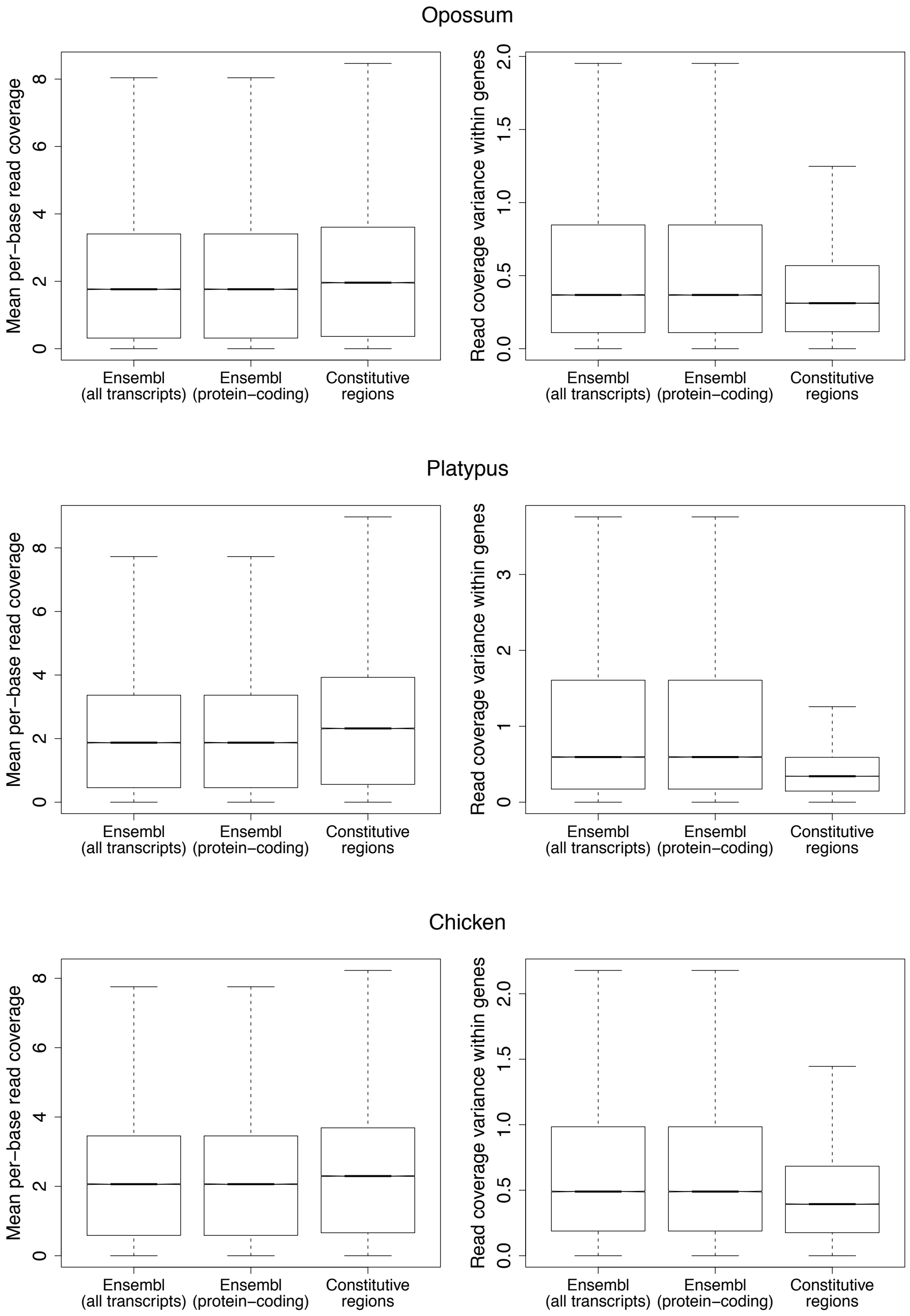
Supplementary Note Figure 5: Examples of read coverage variation along exon blocks. Gray rectangles represent the positions of Ensembl-annotated exon blocks. Hatched rectangles represent regions that were excluded, with our segclust-based approach, before computing gene expression levels.



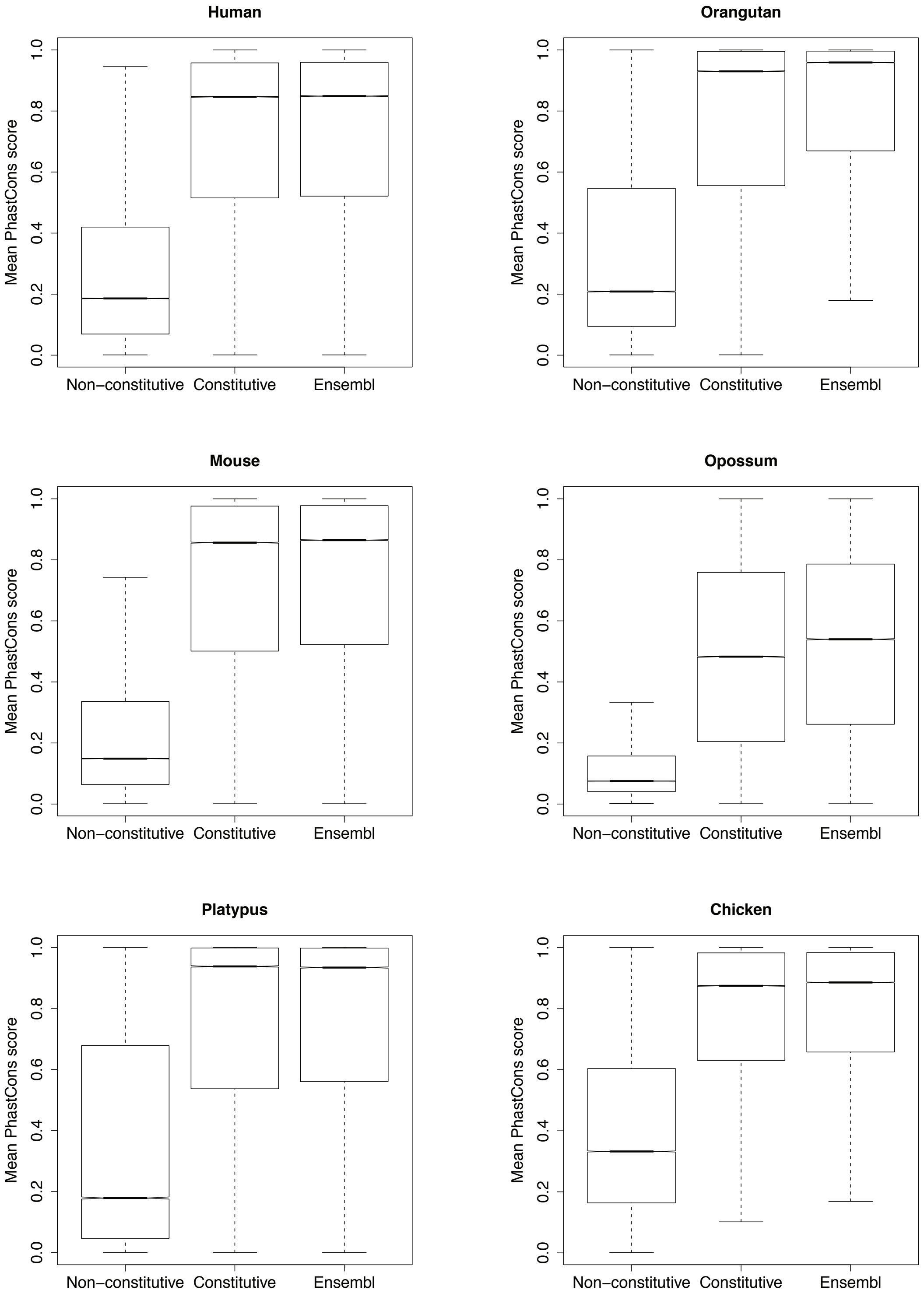
Supplementary Note Figure 6: Distribution of the mean per-base read coverage, and of the within-gene variance of the read coverage, for Ensembl-annotated exons and for "constitutive" exons. All the samples available for one species were combined into a single set before plotting the distributions.



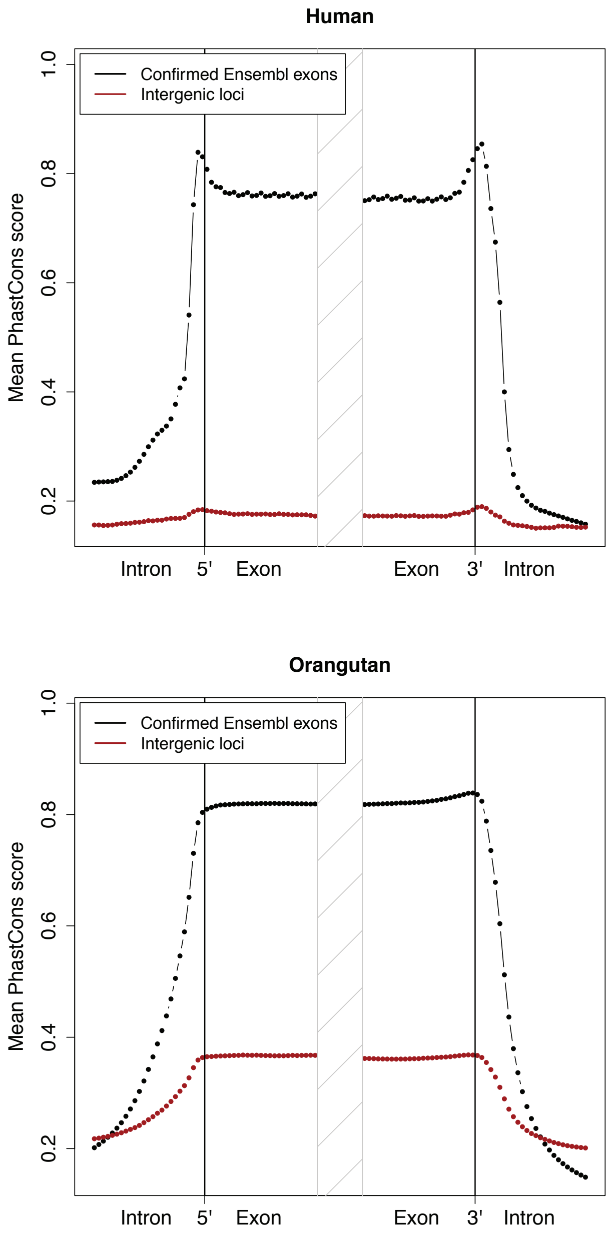
Supplementary Note Figure 7: Distribution of the mean per-base read coverage, and of the within-gene variance of the read coverage, for Ensembl-annotated exons and for "constitutive" exons. All the samples available for one species were combined into a single set before plotting the distributions.



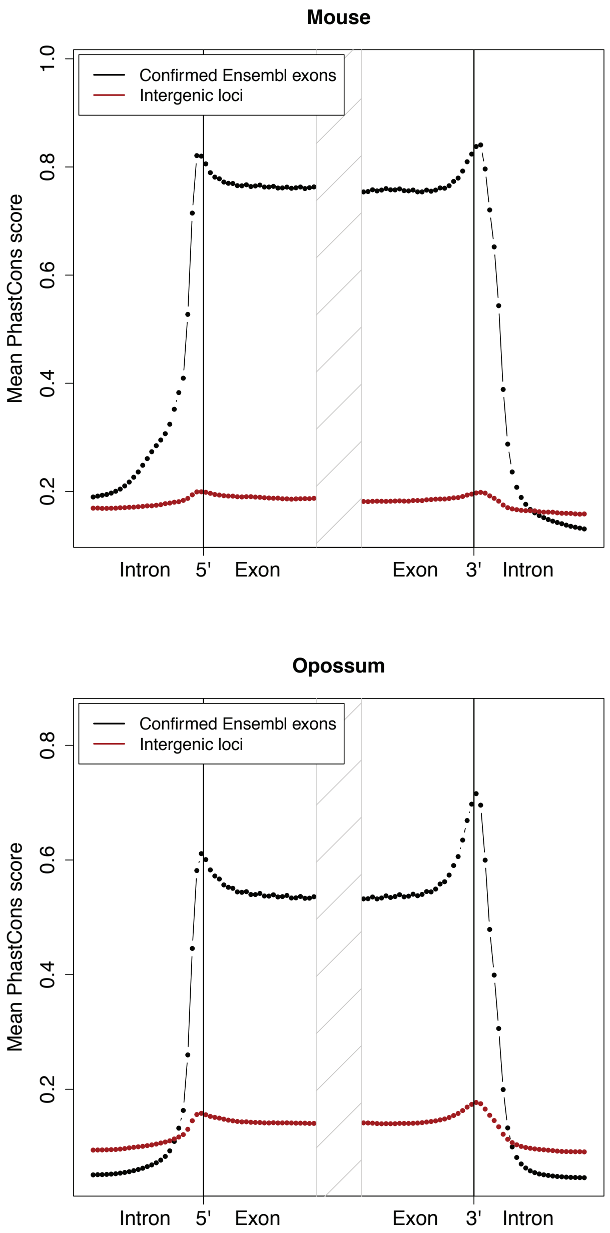
Supplementary Note Figure 8: Distribution of the mean per-base read coverage, and of the within-gene variance of the read coverage, for Ensembl-annotated exons and for "constitutive" exons. All the samples available for one species were combined into a single set before plotting the distributions.



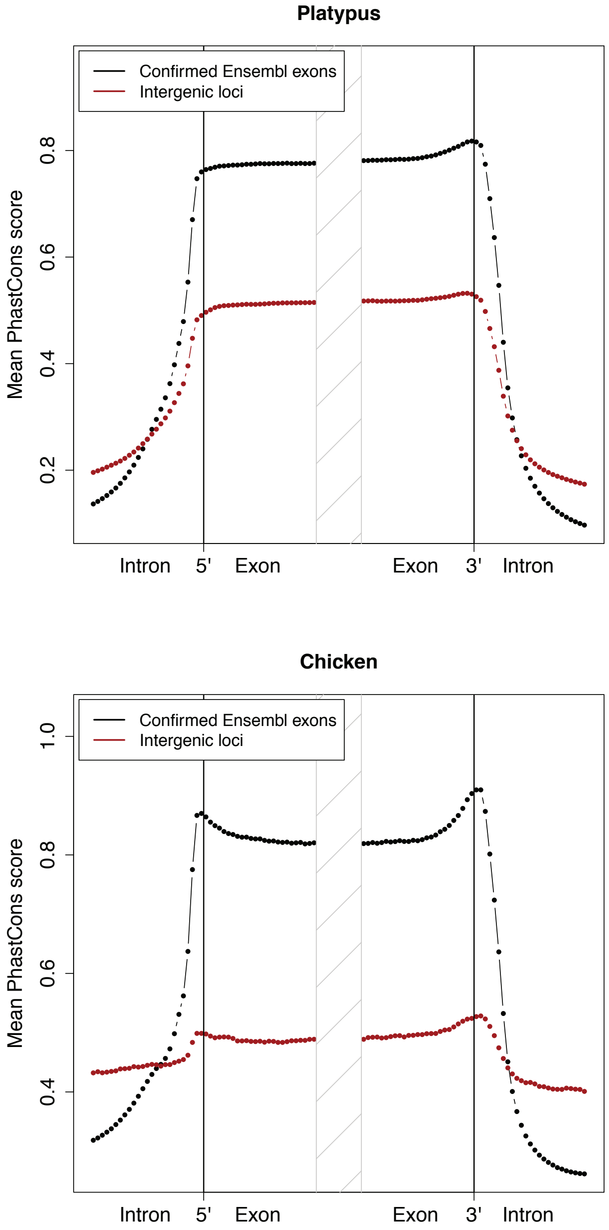
Supplementary Note Figure 9: PhastCons score distribution, for Ensembl-annotated exons,“constitutive” and “non-constitutive” exon blocks.



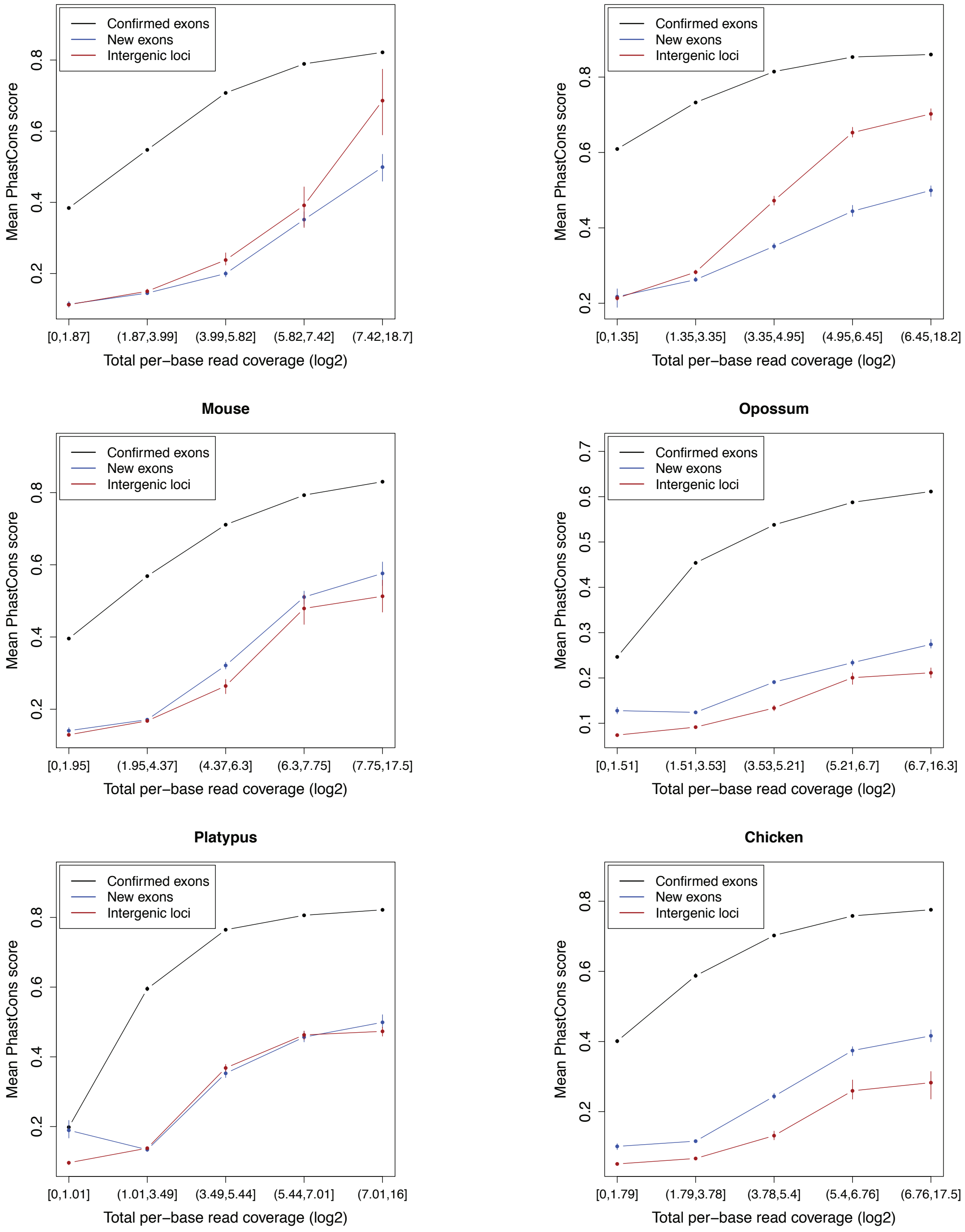
Supplementary Note Figure 10: PhastCons score variation around exon boundaries, for two classes of exons: Ensembl exons confirmed by our annotations (black), and exons of intergenic multi-exonic loci (red). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the human genome; bottom panel: annotations for the orangutan genome.



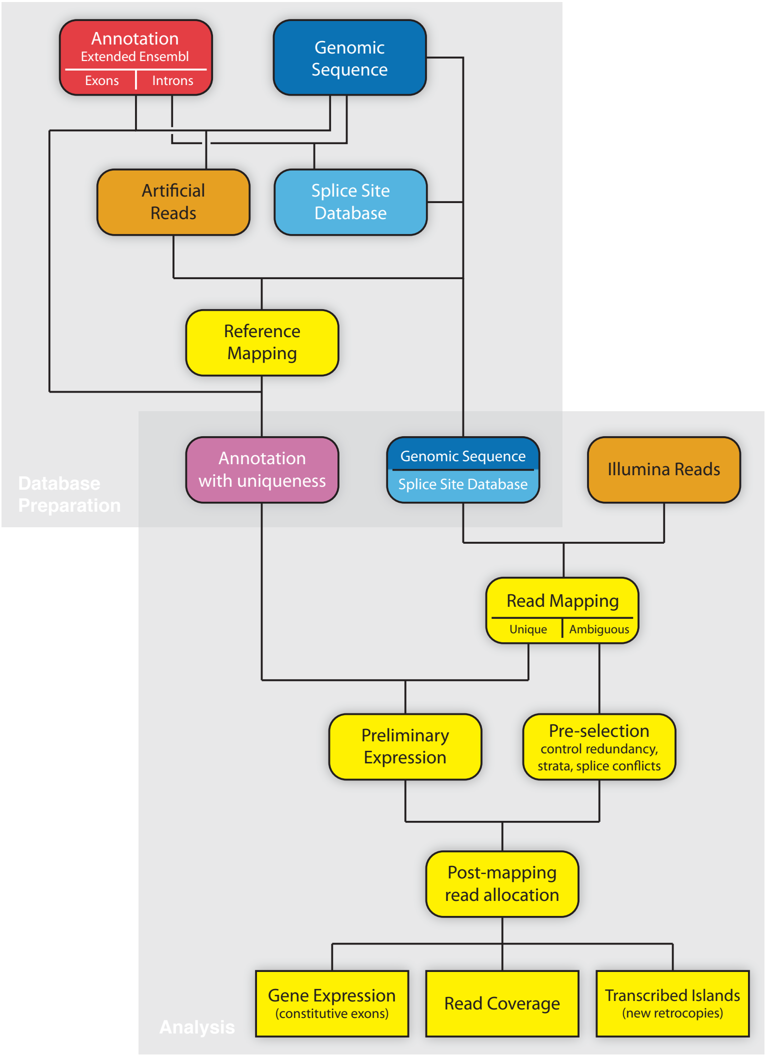
Supplementary Note Figure 11: PhastCons score variation around exon boundaries, for two classes of exons: Ensembl exons confirmed by our annotations (black), and exons of intergenic multi-exonic loci (red). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the mouse genome; bottom panel: annotations for the opossum genome.



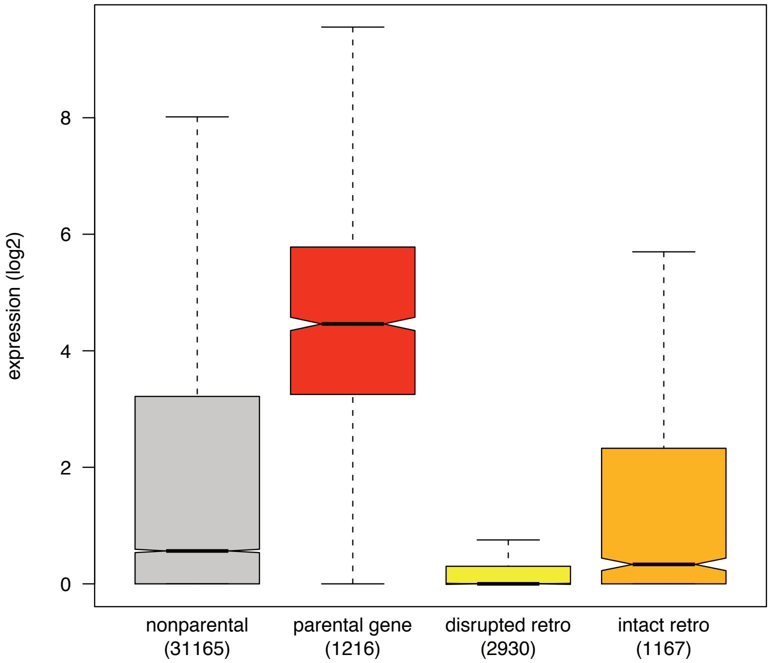
Supplementary Note Figure 12: PhastCons score variation around exon boundaries, for two classes of exons: Ensembl exons confirmed by our annotations (black), and exons of intergenic multi-exonic loci (red). The points represent the mean PhastCons score, averaged over all exons in a class. 4 segments are represented: 1) 25bp upstream of the exon, 2) the first 25bp of the exon, 3) the last 25bp of the exon and 4) 25bp downstream of of the exon. Top panel: annotations for the platypus genome; bottom panel: annotations for the chicken genome.



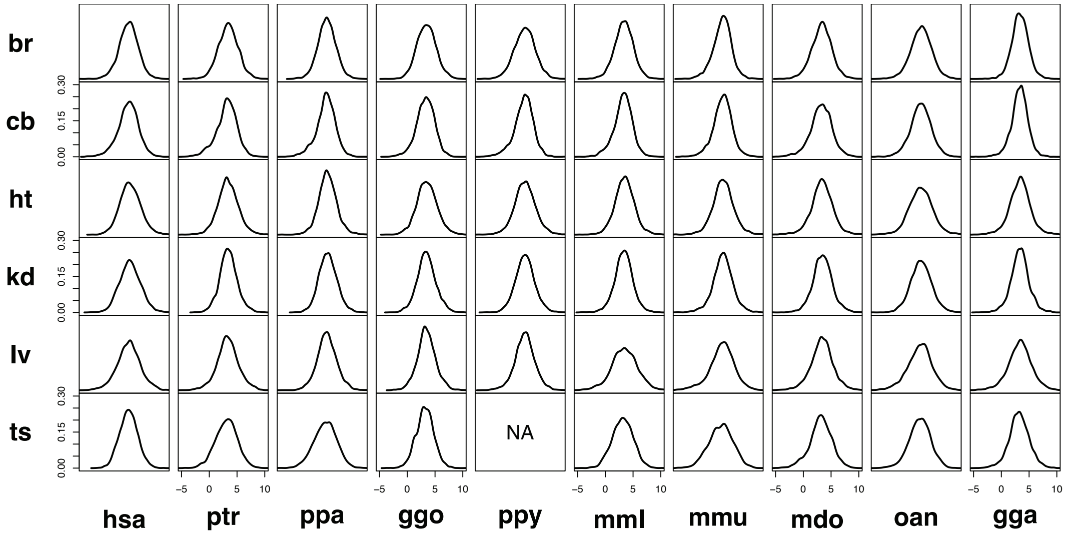
Supplementary Note Figure 13: Relationship between the expression level and the extent of sequence conservation, for three classes of exons: 1) Ensembl-annotated exons confirmed with our annotations, 2) new exons added to Ensembl-annotated genes, and 3) exons of intergenic multi-exonic loci. The expression level was computed as the mean per-base read coverage, all samples confounded, for each exon. Confidence intervals were obtained by bootstrap resampling.



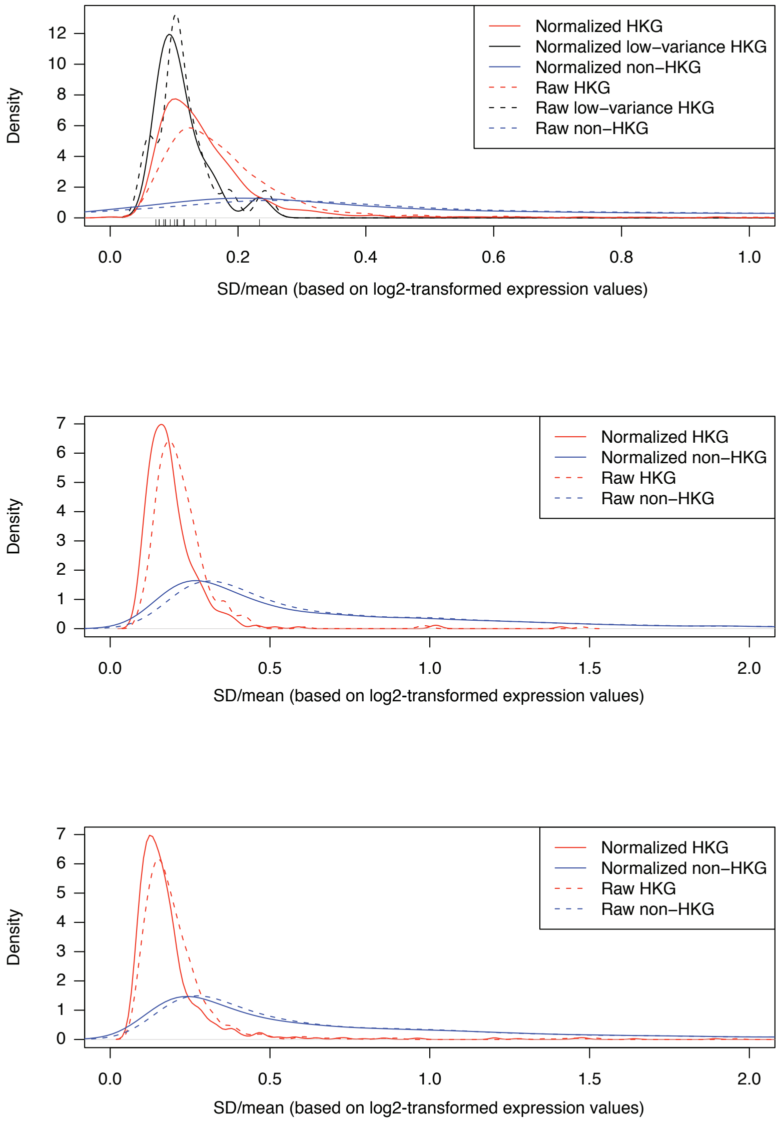
Supplementary Note Figure 14: Overview of the final RNA-Seq read mapping procedure.



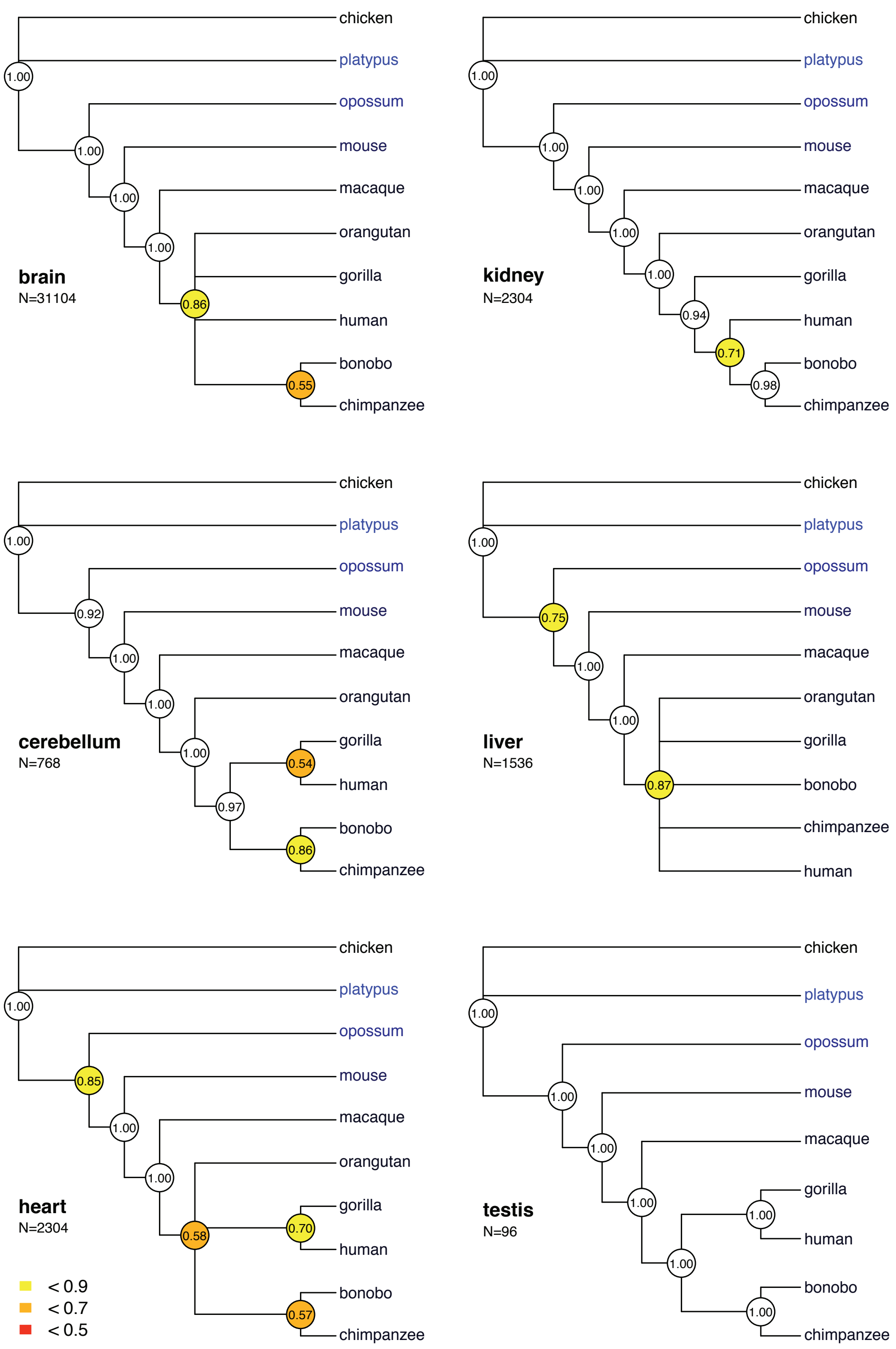
Supplementary Note Figure 15: Expression levels of parental genes and their retroposed copies in the mouse brain (sample ID 670, Supplementary Table 1). Parental gene/retrogene coordinates were established in a previous study[39]. Expression level distributions are shown for multi-exonic genes without retrocopies ("nonparental" genes, grey boxplot), "parental genes" (i.e., genes that gave rise to retrocopies; red), retrocopies with open reading frames disrupted by frame-shift or stop codon mutations (i.e., retropseudogenes; "disrupted retro", yellow), and retrocopies with intact open reading frames ("intact retro", orange). Total numbers of genes/retrocopies are indicated in parentheses.



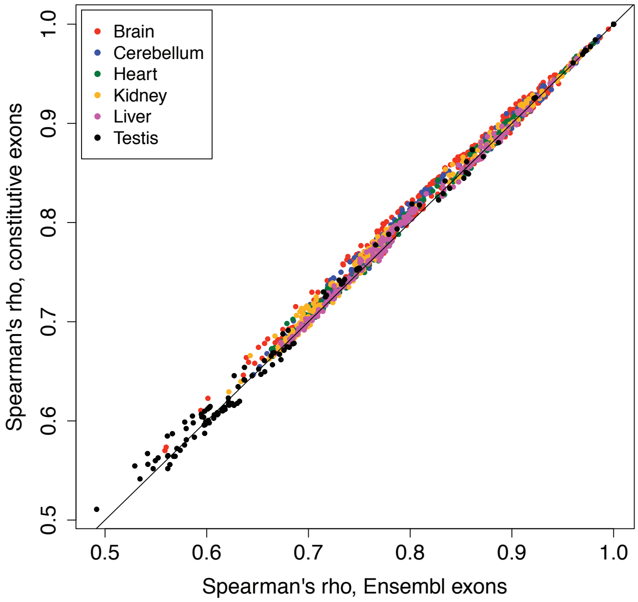
Supplementary Note Figure 16: Distributions of expression levels after normalization (Y-axes: proportion of genes; X-axes: log2-transformed RPKM expression levels).



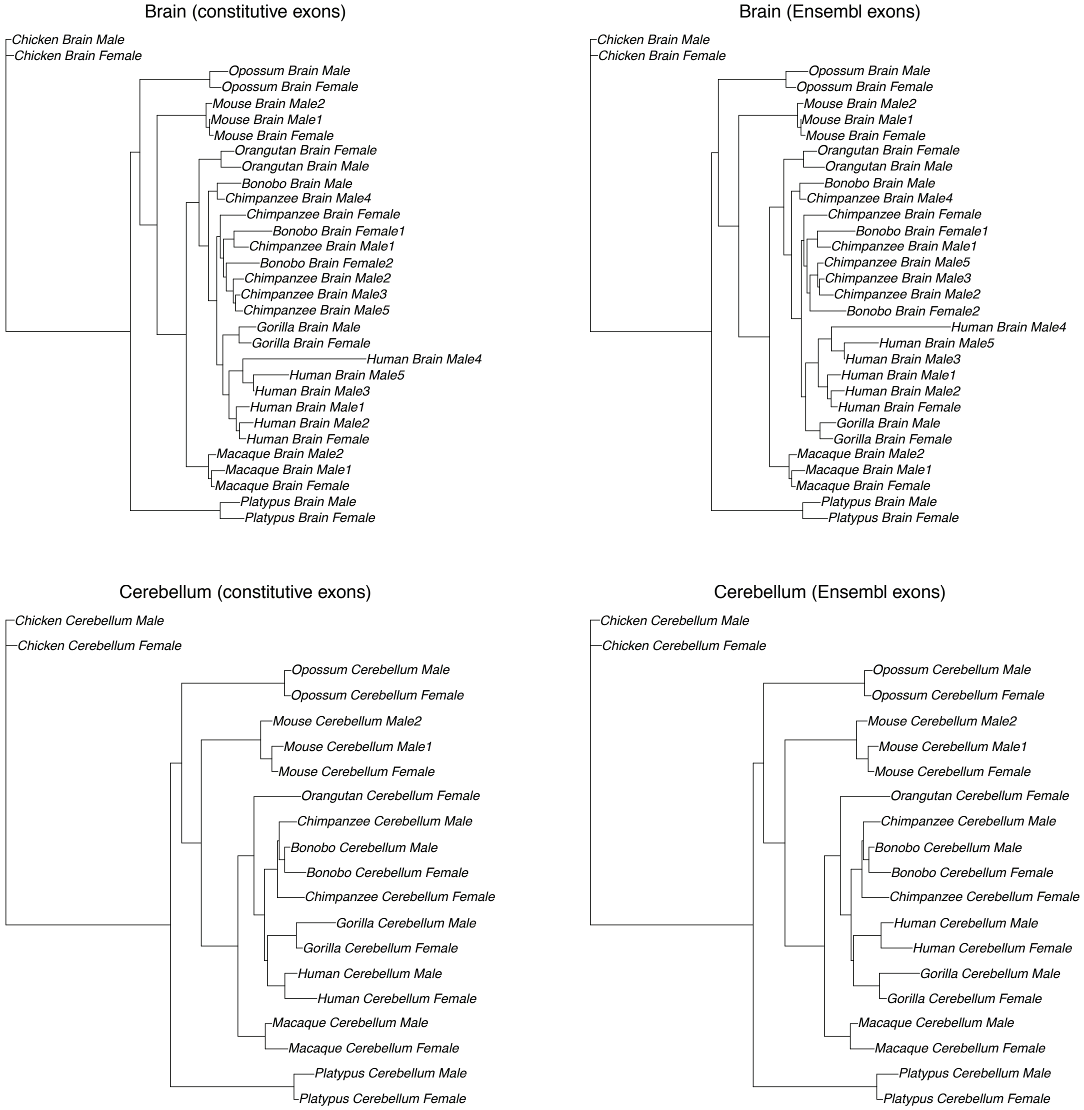
Supplementary Note Figure 17: Distributions of the between-samples coefficient of variance (CV) of gene expression levels, for housekeeping genes (as defined by She *et al*[21]) and for non-housekeeping genes, before and after normalization. Red : housekeeping genes; black : low-variance housekeeping genes; blue : non-housekeeping genes; continuous lines: after normalization; dotted lines: before normalization. Three datasets were analyzed: all human protein-coding genes (top), 5,636 protein-coding genes with 1-1 orthologues in all amniote species (center) and 13,277 protein-coding genes that with 1-1 orthologues in primates(bottom). The CV was computed independently for each gene, among all available samples. The black vertical segments pinpoint the position of the CV values for the low-variance housekeeping genes.



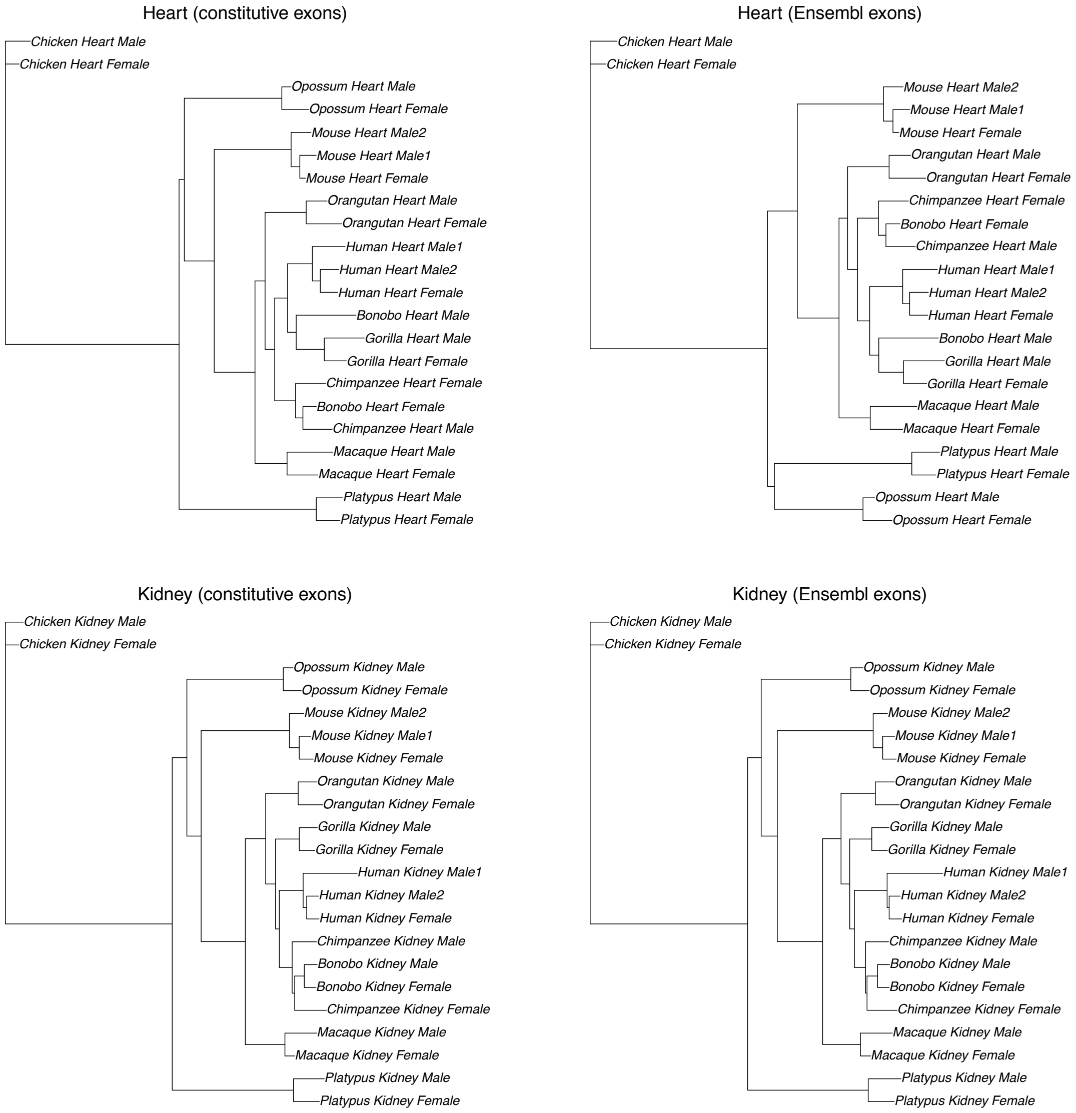
Supplementary Note Figure 18: Robustness of branching patterns in gene expression trees, evaluated with a jackknife-type resampling analysis. In each resampling, one individual was drawn at random from each species. All possible combinations of individuals were analyzed; the number of combinations (N) is shown next to each tree. The numbers represent the proportion of jackknife trees which support the corresponding internal branch.



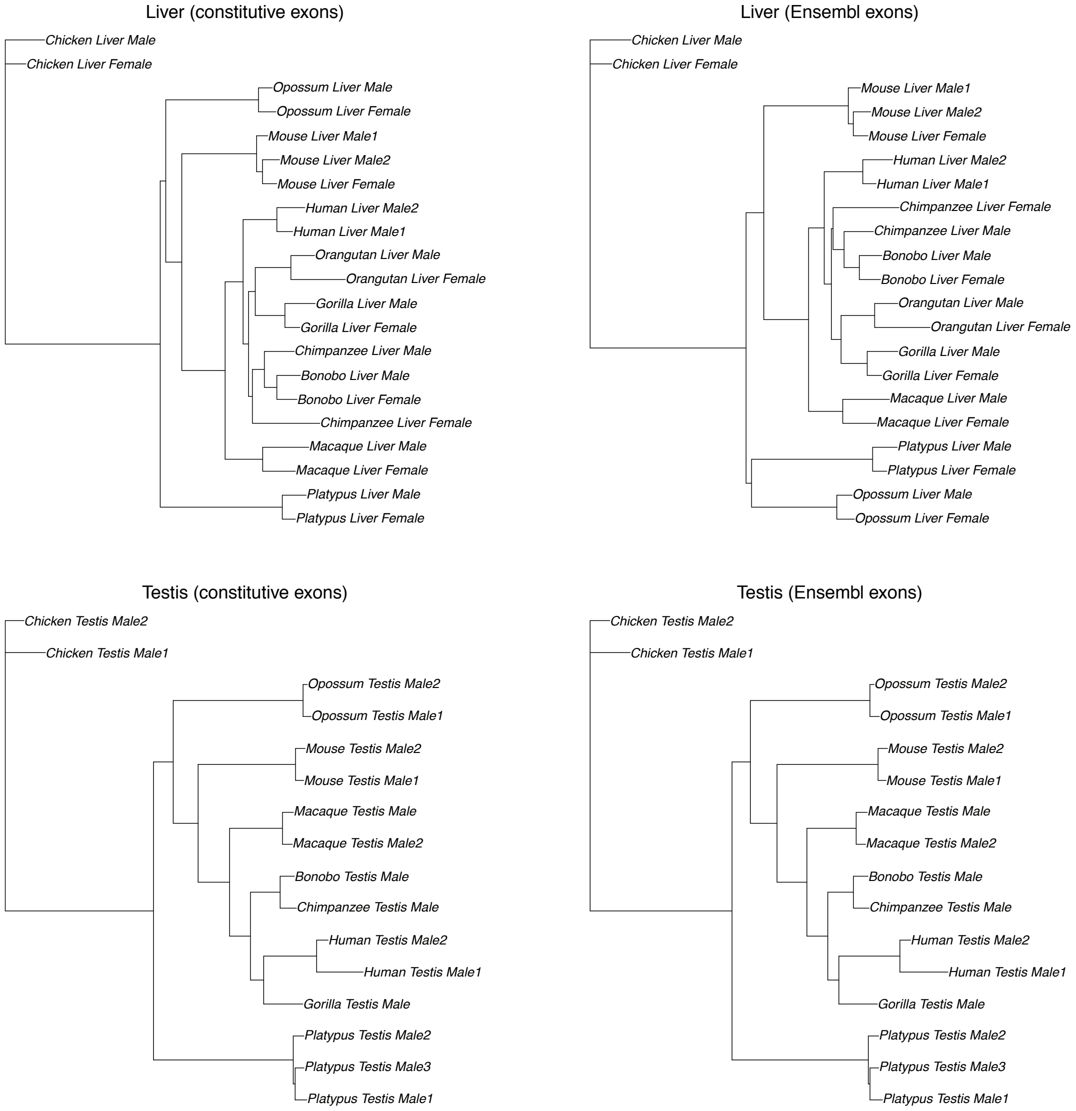
Supplementary Note Figure 19: Comparison between the correlation coefficients obtained when comparing expression levels between samples, for constitutive exons and Ensembl-annotated exons. All comparisons were performed within a single tissue.



Supplementary Note Figure 20: Comparison between expression trees obtained by computing expression levels on constitutive exons (left) and on Ensembl-annotated exons (right). The trees were built with neighbor-joining, on distance matrices derived from Spearman correlation coefficients.



Supplementary Note Figure 21: Comparison between expression trees obtained by computing expression levels on constitutive exons (left) and on Ensembl-annotated exons (right). The trees were built with neighbor-joining, on distance matrices derived from Spearman correlation coefficients.



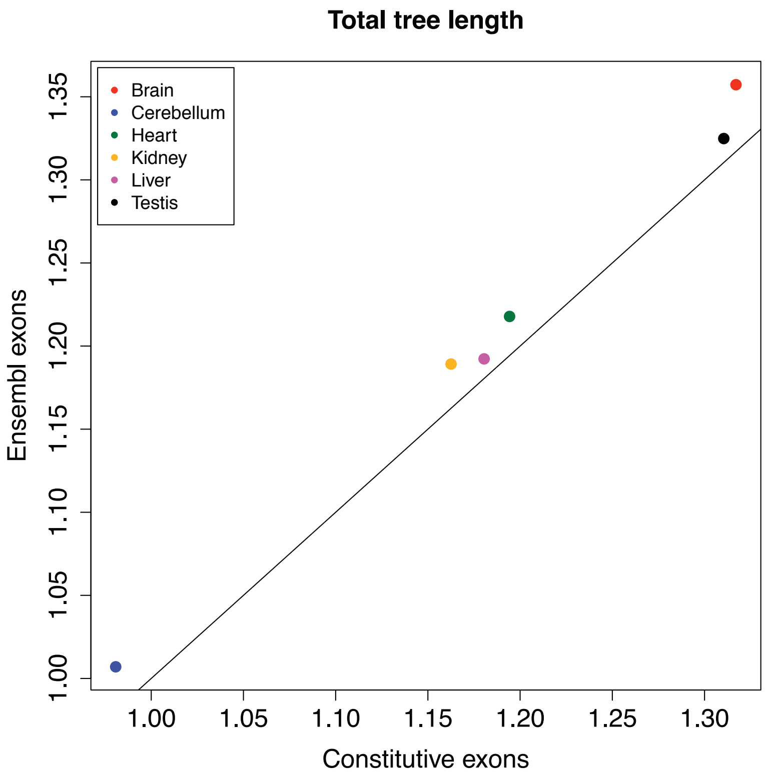
Liver (constitutive exons)

Liver (Ensembl exons)

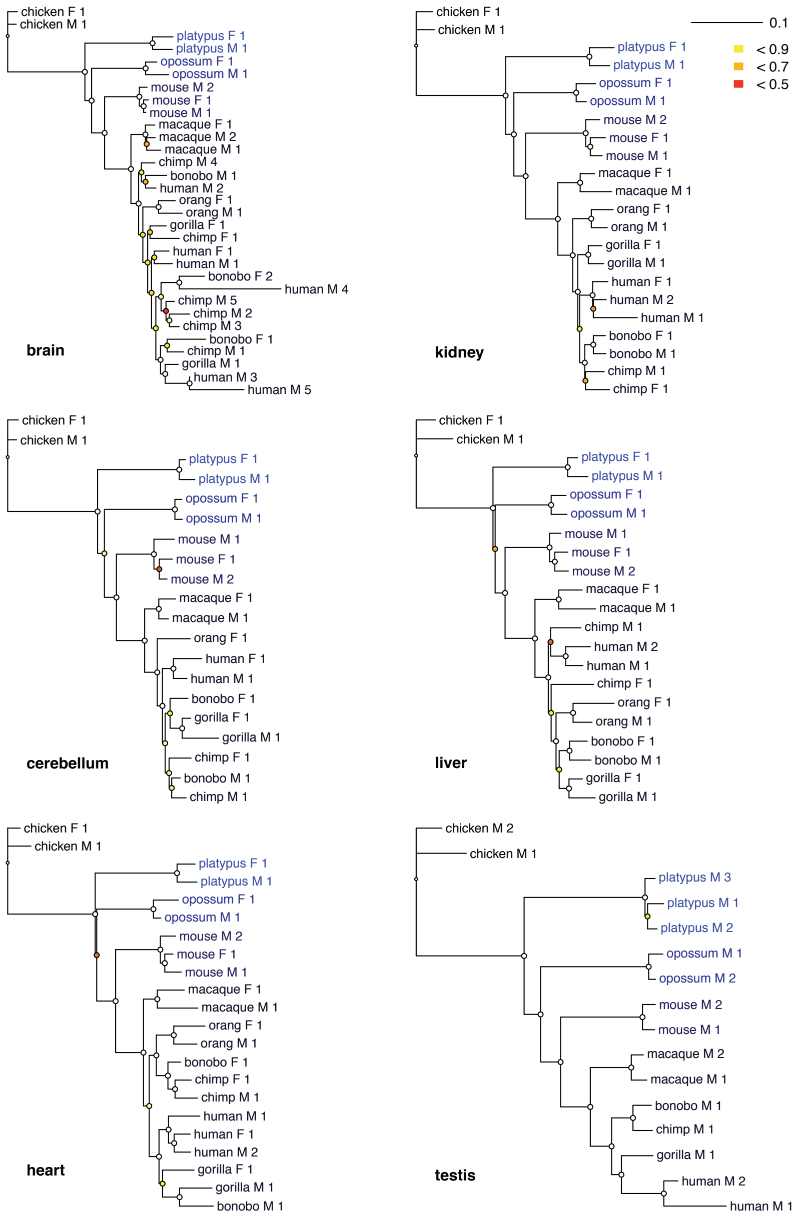
Testis (constitutive exons)

Testis (Ensembl exons)

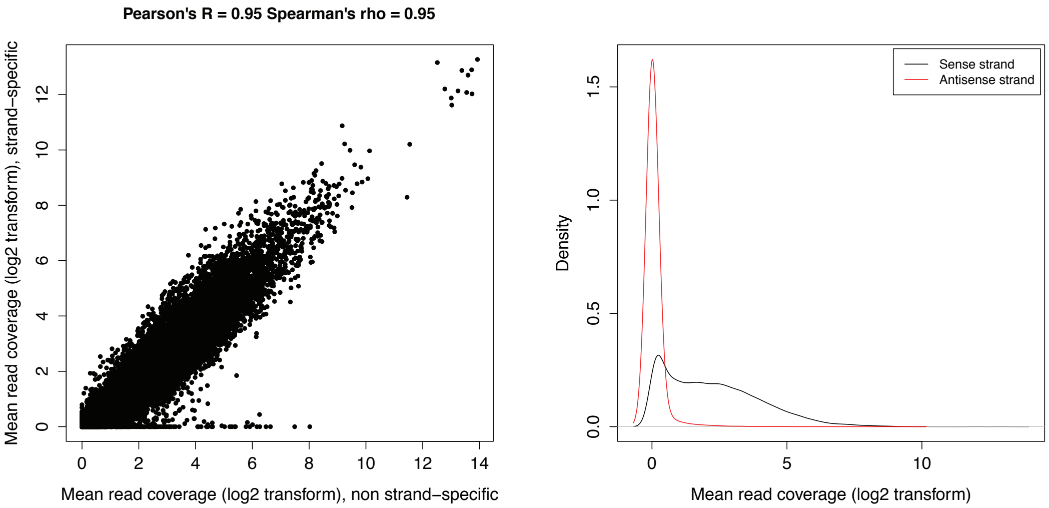
Supplementary Note Figure 22: Comparison between expression trees obtained by computing expression levels on constitutive exons (left) and on Ensembl-annotated exons (right). The trees were built with neighbor-joining, on distance matrices derived from Spearman correlation coefficients.



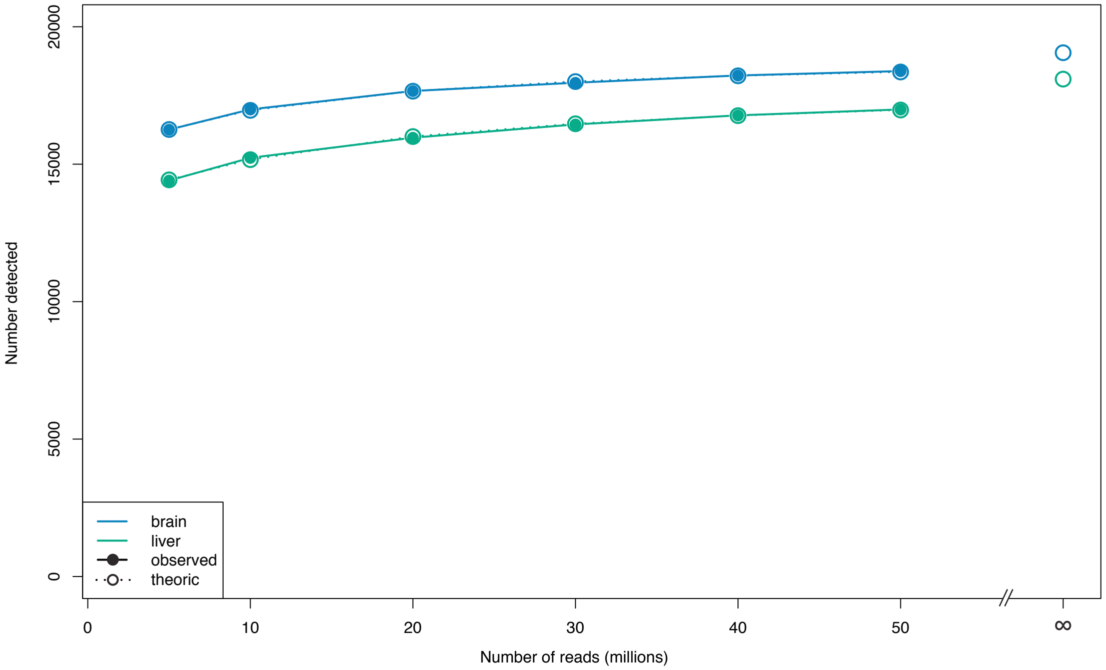
Supplementary Note Figure 23: Total tree length: comparison between Ensembl annotations and our annotations. (N.B.: the tree length is not normalized by the number of taxa.)



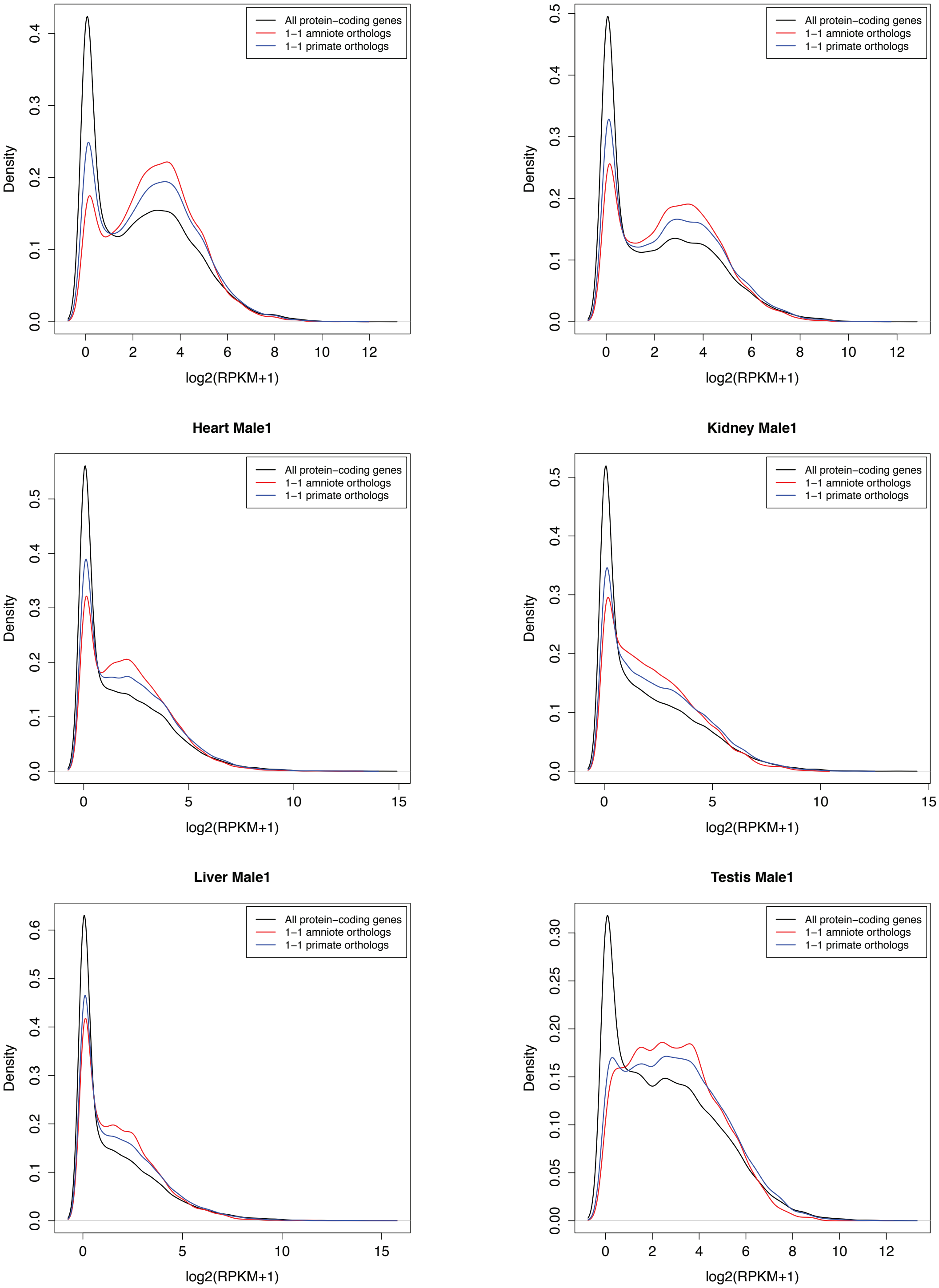
Supplementary Note Figure 24: Neighbor-joining trees based on distance matrices comprising all pairwise expression level distances (1–rho, Spearman's correlation coefficient) for the six different organs. Note that the underlying gene expression values were calculated based on constitutive orthologous exons (from the 5,636 1–1 orthologous genes) that are perfectly aligned (no gaps permitted) between the 10 amniote species studied (see section 1.6 for details). Thus, we rule out any gene expression variation that might arise when gene expression levels are computed for potentially different sequences (e.g., as a result of annotation or biological/genomic differences) among 1–1 orthologs from different species. See Supplementary Figure 2 legend for details regarding the bootstrapping procedure and color-codes.



Supplementary Note Figure 25: Comparison between expression level estimates obtained with a non-strandspecific RNA-Seq protocol (X axis) and with a strand-specific protocol (Y axis), for one human brain sample (male 2). The mean read coverage was computed with unambiguously mapping reads (as determined by TopHat), on Ensembl-annotated exons for protein-coding genes. The read coverage was log2-transformed with the formula $log2(rc + 1)$. Left: scatterplot of the two expression level estimates. Right: distributions of the mean read coverage computed on the sense and antisense strands.



Supplementary Note Figure 26: Effect of the total read coverage on the detection of transcribed protein-coding genes. The saturation curves were plotted using resamplings of 5, 10, 20, 30, 40 and 50 million reads for two mouse tissues (brain and liver). The points represent the numbers of observed (filled circles) and estimated (empty circles) transcribed protein-coding genes, as a function of the total number of mapped reads. The theoretical estimation was done by fitting a model of the form $y = \frac{a}{1+(1/(b*x+c))}$ to the observed distribution. This model was used to estimate the total number of genes that would be detected if there were no read coverage limitations.



Supplementary Note Figure 27: Comparison between the expression levels (RPKM, log2-transformed) of three classes of genes: all protein-coding genes (black), protein-coding genes with 1-1 orthologues in the all-amniote dataset (red) and with 1-1 orthologues in the primate dataset (blue).

References

[1] Khaitovich, P. *et al.* Regional patterns of gene expression in human and chimpanzee brains. *Genome Research* **14**, 1462–73 (2004).

[2] Kircher, M., Stenzel, U. & Kelso, J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**, R83 (2009).

[3] Hubbard, T. J. P. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690–D697 (2009).

[4] Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* (2009).

[5] 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).

[6] Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *www.repeatmasker.org* **000**, 000–000 (1996-2010).

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

[8] Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**, 19–31 (2009).

[9] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050 (2005).

[10] Karolchik, D., Hinrichs, A. S. & Kent, W. J. The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1**, Unit1.4 (2009).

[11] Lai, W. R., Johnson, M. D., Kucherlapati, R. & Park, P. J. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* **21**, 3763–3770 (2005).

[12] 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).

[13] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2010). ISBN 3-900051-07-0.

[14] Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

[15] Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327–335 (2009).

[16] Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708–715 (2004).

[17] Hopcroft, J. & Tarjan, R. Efficient algorithms for graph manipulation. *Communications of the ACM* **16**, 372–378 (1973).

[18] Drummond, D. A. & Wilke, C. O. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352 (2008).

[19] Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5**, 621–628 (2008).

[20] Lee, S. *et al.* Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* (2010).

[21] She, X. *et al.* Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics* **10**, 269 (2009).

[22] Paradis, E., Claude, J. & Strimmer, K. Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289–290 (2004).

[23] Ihmels, J., Bergmann, S. & Barkai, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics* **20**, 1993–2003 (2004).

[24] Bergmann, S., Ihmels, J. & Barkai, N. Iterative signature algorithm for the analysis of large-scale gene expression data. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**, 031902–031902 (2003).

[25] Csárdi, G., Kutalik, Z. & Bergmann, S. Modular analysis of gene expression data with R. *Bioinformatics* **26**, 1376–1377 (2010).

[26] Consortium, T. G. O. Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).

[27] Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).

[28] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 289–300 (1995).

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

[30] Butler, M. & King, A. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Nat* **164**, 683–695 (2004).

[31] Bedford, T. & Hartl, D. L. Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A* **106**, 1133–8 (2009).

[32] Goodman, M. The genomic record of humankind's evolutionary roots. *Am J Hum Genet* **64**, 31–39 (1999).

[33] Janecka, J. E. *et al.* Molecular and genomic data identify the closest living relative of primates. *Science* **318**, 792–794 (2007).

[34] Woodburne, M. O., Rich, T. H. & Springer, M. S. The evolution of tribospheny and the antiquity of mammalian clades. *Mol Phylogenet Evol* **28**, 360–385 (2003).

[35] Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–920 (1998).

[36] Haider, S. *et al.* BioMart Central Portal–unified access to biological data. *Nucleic Acids Res* **37**, W23–7 (2009).

[37] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).

[38] Brawand, D., Wahli, W. & Kaessmann, H. Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol* **6**, e63 (2008).

[39] Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* **6**, e80 (2008).