Title page:

Oxana Sachenkova1, CORE-RIKEN-AUTHORS2 and Lukasz Huminiecki1

1 Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
2 RIKEN Omics Centre, Yokohama, Japan

Corresponding author: Lukasz.Huminiecki@scilifelab.se

Keywords: gene expression, CAGE, evolution, cancer

Title: The structure of animal expression pattern evolution, and gene expression space, revealed by the F5-CAGE encyclopedia of expression patterns

Running Title: F5-CAGE encyclopedia of expression patterns

Abstract:

The F5-CAGE encyclopedia of expression patterns is arguably the most comprehensive and technologically uniform functional genomics dataset generated to date. F5-CAGE includes --- tissues, --- primary cell and --- cancer cell-line samples from human and mouse, and facilitates systematic comparison of gene family expression patters in normal tissues versus cancer. Herein, we apply F5-CAGE towards the investigation of animal expression pattern evolution, focusing on human-mouse comparison. Brain tissues repeatedly exhibit many unique transcriptional features, including clustering into fetal, newborn, and aged adult samples; differential phyloexpression signature; and distinct evolutionary rates in intra- and inter-species comparisons. Analysis of paralog expression pattern divergence suggests global disregulation and devolution of gene expression profiles in caner cell-lines. We compare different strategies for ortholog tissue assignment, a non-trivial problem that has been largely neglected to-date. We investigate strong associations between timing of gene duplication, expression pattern divergence, and tissue-specific domain of duplicated genes. We discuss in detail the cdc42 gene family which includes many tissue-specific genes and dramatic expression pattern shifts, which can be correlated with promoter structure revealed by ENCODE. Finally, tissues, primary cells and cancer cell-lines group into three distinct categories with respect to their rates of expression pattern evolution (static, dynamic, and intermediate), suggesting that conservation of expression profiles is very much a cell- and tissue-type dependent evolutionary force.

(227 words)

Introduction:

Animals are primarily characterized by multicellularity, and existence of distinct tissue- and cell-types. Characterization of the animal expression space in terms of its tissue-specificity, difference in expression between normal tissues, cells in culture, and cancer cells, as well as the timing of emergence of specific cell- and tissue-unique expression programs, is of general interest to broad readership. Herein, we analyze expression pattern evolution and the structure of animal expression pattern space, using the extensive F5-CAGE encyclopedia of vertebrate expression patterns.

The major strength and novelty of the F5-CAGE expression encyclopedia is it's comprehensive and technologically uniform nature, as well as unbiased coverage of the entire genome space, allowing for comparison between tissue samples, primary cell-lines, and cancer cell-lines. F5-CAGE is arguably the most complete functional genomics dataset generated to date. F5-CAGE includes --- tissues, --- primary cell and --- cancer cell-line samples from human and mouse, and will grow to include comparable datasets from three further vertebrate species (rat, dog and chicken). Table 1 briefly summarizes the first release of the F5 CAGE resource.

Expression divergence between protein-coding paralog and ortholog genes has been long studied on the level of amino-acid sequence, however, only recently we have acquired technologies allowing for genome-wide characterization of expression patterns, starting with ESTs, and then microarray data. However, these technologies have fundamental flaws, chiefly tendency for cross-hybridization when studying related sequences of paralogs which are of foremost evolutionary interest. Furthermore, available expression datasets have been limited in scope to a narrow and biased set of somatic tissues, and not allowing for parallel examination of coding and non-coding transcripts (Su, Cooke et al. 2002). Due to these technological limitations and narrow-scope of available datasets, many controversies with regards to animal expression pattern evolution remained unresolved. For example, several authors argued for (Khaitovich, Weiss et al. 2004) and against (Jordan, Marino-Ramirez et al. 2005) the hypothesis of neutral rate of expression pattern evolution, and it still remains controversial what is the appropriate method of calculating expression distances (Pereira, Waxman et al. 2009; Piasecka, Robinson-Rechavi et al. 2012) and clustering expression profiles (Eisen, Spellman et al. 1998; Quackenbush 2001), however fairly early in this debate, we signaled that Person R works best for tissue-specific genes, and that conservation of expression profiles is a complex phenomenon, most likely dependent on tissues and functional classes of genes of interest (Huminiecki and Wolfe 2004).

Here, we use the largest and technologically uniform expression pattern encyclopedia generated to date: F5-CAGE. F5-CAGE is based on the novel CAGE technology (Plessy, Bertin et al. 2010; Salimullah, Sakai et al. 2011). CAGE is less susceptible to cross-hybridization between related paralog sequences, as these tags target very fast diverging promoter regions, instead of conserved transcribed sequences targeted by ESTs of microarrays. CAGE also enables investigation of the genome in a unbiased way, not being limited to a pre-selected protein-coding genes chosen for a given microarray chip platform by the manufacturer.

Crucially, very different trends are observed in terms of expression pattern divergence between tissue samples, and primary cell isolates, and cell lines. Expression divergence between paralogs in cancer cell lines is much higher, indicating that transcriptional programs in many of these samples are disregulated and might not correspond directly to

the *in situ* situation. This matters particularly in view of the fact that the recent catalog of functional DNA elements (ENCODE) was based primarily on immortalized cell line data.

Results:

Principal component analysis (PCA) exploration of the F5 CAGE encyclopedia
The initial exploration of expression pattern diversity in the F5 CAGE encyclopedia using PCA revealed no clear clustering pattern for either PC1/PC2, or PC2/PC3 comparison (data not shown). Sinificantly, no clustering into tissue groups corresponding to ectoderm, mesoderm and endoderm could be observed. However, a PCA analysis focused on brain tissues revealed unexpected clustering into fetal, newborn and aged adult samples (see Unique expression features of brain samples).

Unique expression features of brain samples
In the F5-CAGE encyclopedia, brain samples stand out in many respects, suggesting unique features associated with expression pattern evolution in the CNS. Previously, when comparing expression profiles derived from dbEST, SAGEmap and Affymetrix microarrays, we noted that brain has uniquely complex transcriptional program (Huminiecki, Lloyd et al. 2003), and that brain-expressed genes mostly derive from the 2R-WGD event (Huminiecki and Heldin 2010), with few novel brain-specific genes forming during diversification of mammals. The results presented here, both strengthen and clarify this hypothesis. Brain samples were clear outliers in many different analyses. For example, unlike all other tissues, brain samples exhibited unique clustering by age of donor, forming three clusters which can be attributed to developmental stage: (a) fetal, (b) newborn, and (c) aged adult (Fig1). No other tissue type demonstrated this type of clustering, despite availability of multiple donors and developmental stages for both human and mouse samples, precluding the possibility that observed clustering of brain samples is only due to post-mortem delay, or differential pre-processing of samples during RNA isolation.

Phylogenetic timing of gene duplications using TreeFam8 database
We have previously used TreeFam database (Li, Coghlan et al. 2006) to phylogenetically time gene duplication events (Huminiecki and Heldin 2010). New and much larger release of the database, TreeFam8, was used here. This new database release strongly confirmed previous work, linking taxa Vertebrata and Bilateria with two greatest waves of gene duplications in the history of animal kingdom (Fig---).

Phylogenetic data from TreeFam database was linked with F5-CAGE WP4 expression tables, and most later stages of analysis were performed in R/Bioconductor (2.11), using among others, packages Biodist (correlations), gplot and ggplot (graphics), rtracklayer, TxDb.Hsapiens.UCSC.hg19.knownGene, GOstats.

TODO: F5-CAGE WP4 expression data pipeline ---
(From the main manuscript).

Hierarchical clustering of human tissue samples
When F5-CAGE WP4 expression data are clustered (Spearman correlation), samples cluster primarily depending on cell/tissue type of origin, not sample donor or developmental stage. This indicates that tissue-of-origin differences are much more important than individual variability, as multiple donors are available for a high proportion of samples. Corresponding heatmaps for primary cell-lines, and cancer cell-lines are shown in Figure S1 and S2. Significantly, The cancer cell-line clustering shows a major divide between leukemias and solid tumors.

Assignment of ortholog tissues

We compared ortholog human-mouse tissue clusters obtained by simple name matching (NM-clusters), with those inferred using inter-species hierarchical clustering of samples (ISHC). Table 3 is a summary table, comparing orthology inferences made from the Pearson R based ISHC-clusters (Fig4a) and the NM-clusters. The full F5 CAGE NM-dataset is shown in Table S---. 8 NM-clusters are recovered by the ISHC, while 19 are not. The name clusters which are recovered include: skin, liver, tongue, heart, pancreas, pituitary gland, thymus and “total RNA control”.

Gene duplication timing and spatial expression domain of progeny genes

We defined phyloexpression profiles as strong associations between individual F5 CAGE samples, and gene duplicates derived from certain taxa, and searched for associations between phyloexpression profiles and taxa-specific evolutionary novelties.

This analysis revealed several exceptionally strong tissue/taxon associations (Table 7). For example, cystatins (Baron, DeCarlo et al. 1999) and proline-rich salivary proteins (Amado, Lobo et al. 2010) are known to be present in saliva, and here we show that these proteins are very strongly associated with --- specific gene duplications and strongly expressed in salivary gland samples.

Young duplicates are tissue-specific in their expression domain, while old duplicates are broadly expressed

We observe two broad evolutionary trends related to animal expression patterns: trend for gradual paralog expression pattern divergence, and trend for young genes to be more tissue-specific in their expression domain, than old genes. Figure 6a shows the trend for gradual expression pattern divergence between paralogs in human tissues over evolutionary time. In addition, we observe a very strong trend for young duplicates to be more specific in their spatial expression domain (Figure 6b). Taxon Eutheria (Figure 6a) is a strong outlier to the gradual divergence trend, raising the possibility of unique functional characteristics of genes derived from this taxon. Bilateria is also an outlier, although to a lesser extent. Overall, paralog expression divergence is fast, reaching overall plateau as soon as Amniota. In contrast, the trend for older genes becoming more and more housekeeping, doesn't really reach a plateau until the base of the animal tree of life (taxon Metazoa). When this analysis is extended into human normal and cancer cell lines, it becomes apparent that similar trends can be observed in normal cells, but cancer cell lines differ, as there is no strong signature for young duplicates to be co-expressed (Fig 6c).

Histogram in Fig 6b illustrates the size of the dataset (number of paralog pairs) available for each taxon (with Vertebrates and Bilateria being most numerous).

We then asked what are the functional characteristics of genes lying at the extremes of these two trends. Supplementary tables xxx1 and xxx2 and xxx3 and xxx4, summarize the results of GO term and PFAM domain enrichment, for all taxa. Table xxx1 and xxx2 relate to the expression divergence trend (highly and lowly correlated extremes of the distribution, respectively), while xxx3 and xxx4 related to the breadth of expression trend (housekeeping and tissue-specific extremes of the distribution respectively). The number of pairs available for some taxa may too low to reach definite conclusions, but several observations at the strict p=0.00001 cut-off are worth mentioning:
(a) Eutheria, highly significant association of top correlated (0.75 quantile) genes (Figure 6a pointer) with cellular macromolecular complex assembly (GO:0034622), chromatin assembly or disassembly (GO:0006333), nucleosome assembly (GO:0006334), DNA packaging (GO:0006323); (b) association of PF00125 (Histone) domain with human-specific highly housekeeping genes; (c) associations of the following cellular location (CC) GO terms with broadly expressed duplicates mapping to Homo/Pan/Gorilla and Catarrhini: extracellular region (GO:0005576), extracellular space GO:0005615; (d) under-

representation of intracellular (GO:0005622), cell (GO:0005623), cytoplasm (GO:0005737), organelle (GO:0043226) among broadly expressed duplicates mapping to Catarrhini.

Evolutionary history of mammalian tissues.
To illustrate to the potential of phylogenetically-aware expression pattern profiling, we show Figures focusing on two examples. Figure 8 focuses on brain samples, while Figure 9 focuses on the reproductive track. Few new brain genes have been created since 2R-WGD, but many genes specific to the reproductive track have emerged in mammals. In addition, Table 6 --- Table 7--- describe strongest associations between taxa of gene duplications, and expression patterns in F5 CAGE.

Example gene family with dynamic pattern of expression pattern evolution
TreeFam family TF101109 is a family of small signaling G proteins (more specifically GTPases), and members of the Rac subfamily of the family Rho family of GTPases. Rho family of GTPases appear to regulate a diverse array of cellular events, including the control of cell growth, cytoskeletal reorganization, and the activation of protein kinases. The cdc42 gene family includes many tissue-specific genes and dramatic expression pattern shifts. Figure 11 illustrate expression patterns of this family in human tissue, while supplementary figures (S10 and S11) provide information about expression in primary cells, and cancer cell-lines.

Expression of cdc42 family in human tissues illustrate a novel phenomenon of mutually exclusive expression, where Ras–related C3 botulinum toxin substrate 3 (Rac3: GeneID 5881) is very highly expressed in five fetal tissues from which RHOG (GeneID 391), and ras–related C3 botulinum toxin substrate 2 (RAC2: GeneId 5880) are excluded. The five tissue in question include: fetal parietal lobe, fetal temporal lobe, fetal duodenum, fetal occipital lobe, and fetal brain pool, indicating that following 2R-WGD Rac3 neofuncionalized to play a role embryonic CNS. In contrast, RHOG is highly expressed in adult corpus callosum, where the other two genes are not detectable. Furthermore, a cluster of tissues associated with the immune and circulatory systems (thymus adult, blood adult, tonsil adult, appendix adult, thymus fetal, vein adult, spleen adult, lymph node adult, spleen fetal), express RHOG and RAC2, but not RAC3. An examination of the cdc42 TreeFam family tree (http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF101109), suggests that Rac2 and Rac3 are 2R-ohnologs, while divergence between Rac2/3 ancestral gene and RHOG predates the origin of animals (taxon of duplication = Eukaryota). This suggests a very clear example of expression pattern neofunctionalization following Rac2/3 gene duplication, where RAC2 retained most of the ancestral expression characteristics (shared by RHOG), while RAC3 acquired new expression sites in fetal brain, while loosing expression in most other adult and fetal tissues.

For clarity, it should be mentioned that RHOJ and RHOQ are very divergent unrelated members of the family which appear to show little evidence for expression in human tissues. Rac1 and cdc42 are also highly divergent genes (taxon of duplication Eukaryota), which appear nevertheless similar in their expression characteristics in human tissues.

It is also interesting to compare expression features of the cdc42 family across human tissues with those observed in normal cell lines and cancer cell lines. Here RAC3 appears weakly expressed in unrelated cancer cell-lines, in congruence with it's emerging characteristics as a tightly regulated tissue-specific gene, restricted in its expression mostly to embryonic brain tissues. In contrast, RHOG and RAC2 are widely expressed in most cancer cell lines, and no clear pattern of mutually exclusive expression can be seen. In normal cell lines, RHOG and RAC2 are also widely expressed, while RAC3 is limited in its expression.

Discussion:

To examine overall variability of expression between different tissues in the Fantom5 dataset, we used principal component analysis (PCA) but no clear overall pattern of clustering was revealed. This underlines intrinsic variability of expression patterns in different lineages, and argues against well controlled global blocks of transcriptional regulation common to sets of tissues, for example for those derived from the three distinct embryonic layers (ectoderm, endoderm, and mesoderm).

Brain has many unique expression features, such as high number of genes specific to this tissue (Huminiecki, Lloyd et al. 2003) most of which originated from the 2R-WGD event (Huminiecki and Heldin 2010), which prompted us to examine brain samples separately by PCA (Figure 1). Interestingly, Brawand et al. (Brawand, Soumillon et al. 2011) reported lack of clear separation of brain tissue by PCA, in a limited sample of six organs. However, our dataset is vastly more complete than that used by Brawand et al., and includes samples from multiple donors and developmental stages.

The most striking feature of our brain-focused PCA analysis (Figure 1) is that brain libraries appear to form three distinct clusters (a) fetal, (b) neonatal, and (c) somewhat dispersed adult cluster. Intuitively, this makes sense if one considers lengthy brain development and maturation: brains of newborns are arguably the most immature of all babies organs. No similar clustering by developmental stage could be observed for two other human tissues examined in detail: lung, or liver, where samples from multiple donors and developmental stages were available in the F5 dataset (data not shown).

This initial PCA analysis was followed by hierarchical clustering of human brain samples (dendrogram shown in Figure S---), which confirmed the existence of age-related clusters, described in detail in Table 2.

Figure S1 shows subsequent comparative-expression-PCA (CE-PCA) analysis performed jointly for both human and mouse brain samples (where human and mouse data were linked using ortholog genes, following the same principle applied to the ISHC). In the CE-PCA, human and mouse samples form two strongly demarcated clusters. Insufficient mouse data are available at this stage to draw firm conclusions (for example, there are no mouse fetal F5-CAGE samples, and mouse data are somewhat biased towards cerebellum and visual cortex samples). However, given available data, no developmental stage related clustering can be observed in mouse (Figure S1), suggesting that murine brain development is different in that respect from hominid.

Previous studies comparing evolutionary profiles between species assumed a very simple model for assignment of ortholog tissues: automatically assuming that samples with matching names in different species are ortholog tissues. Here, we suggest that this is over-simplification of a complex problem. Firstly, expression pattern evolution is fast: lineage-specific expression pattern shifts and tissue-specific evolutionary novelties put into question the very assumption of the existence of ortholog tissues. Conceptually it may simply be wrong to assume that human brain corresponds to mouse brain, or that human stomach corresponds to mouse stomach, as the behavioral and ecological differences between these two species are very substantial. Secondly, differences in pre-processing of samples and anatomical differences between species may make it difficult to isolate RNAs in a reproducible fashion from species with dramatically different body size, and different anatomy (especially in case of tissues with complex anatomy, such as the CNS, or reproductive system).

To investigate the problem of ortholog tissue assignment more systematically, we compared ortholog clusters derived through (a) simple name matching procedure (NM-clusters), and the ortholog-based ISHC procedure. Conceptually, NM-clustering procedure

may be regarded as equivalent to a data-mining approach utilizing supervised learning. On the other hand, the ISHC procedure, just like hierarchical clustering on which it is based, is an unsupervised learning approach (Table 3).

The NM-clusters which are recovered through the ISHC include: skin, liver, tongue, heart, pancreas, pituitary gland, thymus and “total RNA control”. However, twice as many NM-clusters (namely 19) were not recovered. This effect seems robust to alterations of the ISHC procedure. We have experimented with other expression distance measures, and ISHC based on whole family-averaging rather than ortholog genes, but not higher rates of NM-cluster recovery could be achieved (data not shown).

Histogram in Fig4b contains stack histogram with Pearson R distributions, obtained during the course of the ISHC procedure, for the two intra-species comparisons (hs, mm), and the inter-species comparison (hs-mm). Inter-species distances (hs-mm) are somewhat higher than intra-species distances (hs and mm): 0.78, 0.68 and 0.72 are the respective means.

A number of observations followed on comparison of several sample types (tissues, primary cells, and cancer cell-lines) available for human in the F5-CAGE dataset. For example, one of the most broadly relevant conclusions of this work, is that closely related paralogs are not correlated in their expression in cancer-cell lines, in contrast to tissues and primary cell-lines. In other words, while recent closely related paralogs tend to be expressed in the same tissue, they are not co-expressed in the same cancer cell lines. Assuming that co-expression of closely related paralogs is conditioned by the structure of their promoter regions, this suggest that normal conditions of promoter regulation do not exist in cancer cell lines. The effect cannot be attributed to cell culture conditions alone, as paralog co-expression signature is seen in primary cell culture samples.

It has been proposed previously that primary mutations and expression pattern changes in cancers are accompanied by a wide range of secondary changes, however, the findings presented here are perhaps the most comprehensive demonstration of this expression pattern disregulation in cancer. We propose a novel term: devolution of expression patterns, to suggest that normal evolutionary constraints on expression patterns do not exist in cancer.

In a more detailed follow up analysis focused on expression patterns of entire gene families, we have compared two expression characteristics of TreeFam families in tissues, primary cells, and cancer cell-lines: average expression, and preferential expression measure (Table 4, Table 5 and supplementary Table S---). Data in Table S--- can be queried in multiple additional ways, depending on the biological question that is of most interest to the reader. For example, one can identify families preferentially expressed in both tissues, and primary cells, versus cancer cell-lines; preferentially expressed in tissues versus primary cells and cancer cell-lines; tissue-specific in tissues but not in cancer; tissue-specific in tissues and primary cells but not cancer cell-lines, etc, etc.

To examine overall differences in average gene family expression characteristics between F5-CAGE tissues, primary cells, and cancer cell-lines, we performed multivariate analysis of seven variables associated with TreeFam families in the F5-CAGE dataset (Figure 12): number of human family members (family_size), average expression in human tissues (AVG_tissues), average expression in cells (AVG_cells), average expression in cancer (AVG_cancer), preferential expression measure in tissues (PEM_tissues), preferential expression measure in cells (PEM_cells), and preferential expression measure in cancer (PEM_cancer). While average expression values correlate strongly between the three sample types, preferential expression measure in tissues does not correlate with those calculated for cells and cancer cell lines. This makes sense if one considers that tissues are mixtures of different cell-types, and suggests that fundamentally

different regulatory phenomena are responsible for tissue-specific expression versus cell-line specific expression.

Finally, we have used self-organizing map (SOM) as a multivariate data-mining approach to cluster gene families into groups of similar expression characteristics in human tissues, primary cells, and cancer cell lines (Figure 13, TableS_SOMa and TableS_SOMb). The rectangular 6 by 6 SOM proved efficient in clustering TreeFam familyA families into groups with similar expression characteristics in the three human sample types. Several, clusters (i.e. prototypes) warrant particular attention (**prototype No**: *size*): (**1**: *2*) high average expression in all three sample types (SH3 domain-binding glutamic acid-rich-like protein, tumor rejection antigen); (**4**: *3*) high expression in tissues but low in cells and cancer cell lines (natriuretic peptide precursor A/B, quinoid dihydropteridine reductase, kinesin family member 5); (**30**: *5*) high PEM in cells and cancer cell lines, all other variables low (superoxide dismutase 2, RAB7, ATP synthase, RAN, protein disulfide isomerase family A); (**35**: *4*) high PEM in tissues, lower in cells and cancer cell lines, all other variables low (damage-specific DNA binding protein, replication factor C, cleavage and polyadenylation specific factor 3, replication protein A1); (**36**: *2*) very high PEM in cells and cancer cell-lines, all other variables low (heat shock 27kDa protein, and superoxide dismutase families).

It is interesting to ask if changes in expression patterns, in particular dramatic shifts in expression patters, and mutually exclusive expression are correlated with promoter structure and transcription binding patterns. We have therefore, examined promoters of the three cdc42 family genes (RAC2, RAC3, and RHOG) using pooled ENCODE data for transcription factor binding sites (Tfbs) from ENCODE cell lines (see methods). Table 8 lists all Tfbs linked to these promoters, and a subset of the the dataset with the strongest signal (score > 750, overall score varies between 0-1000). Table 9 shows Jaccard index values for pairwise comparisons between all family members. We have also examined promoter regions of these three genes manually with the UCSC genome browser featuring ENCODE data. Narrowly expressed, RAC3 (promoter region at chr17: 79988532 - 79990531) features one strong binding site for TF ZBTB7A (ZBTB7A_(SC-34508); cluster score (out of 1000): 1000; K562 cell line; chr17:79989226-79989435), a weak and narrow DNASEI site (score 24), and a weak H3K27AC signal. In contrast, liberally expressed RAC2 (Promoter = chr22: 37639306-37641305) and RHOG (Promoter = chr11:3861214-3863213) have broad and high-scoring DNASEI sites (DNASEI scores of 121 and 136 respectively), and very strong H3k27AC signals. RAC2 and RHOG have many Tfbs within narrow 500 bps window of their TSSes, with strong NFKB signature for both RAC2 (cluster score (out of 1000): 852; chr22:37640075-37640448; cell line GM12891), and RHOG (cluster score (out of 1000): 716; chr11:3862446-3862758; cell line GM12891). This suggests a scenario where ancestral NFKB binding site was replaced by ZBTB7A in RAC3.

Indeed, published literature and genomics data available for NFKB and ZBTB7A, give further credence to the evolutionary scenario correlating replacement of these two Tfbs, with expression pattern shift from broad expression pattern preferentially associated with immune and circulatory systems (RAC2 and RHOG), to narrow expression associated with fetal brain (RAC3):

(a) ZBTB7A (http://www.factorbook.org/mediawiki/index.php/ZBTB7A) stands for Homo sapiens zinc finger and BTB domain containing 7A (a gene originally called Pokemon, a name which was withdrawn for obvious reasons under a threat of legal action), with highest expression in GNF Expression Atlas 1 associated with “whole brain”. Interestingly, ZBTB7A promoter region itself (chr19:4,066,216-4,068,615) features three ENCODE ZBTB7A_(SC-34508) binding sites, in addition to Egr-1, SUZ12, and a CTCF binding site. ZBTB7A had been shown to be an oncogenic transcription factor (Maeda,

Hobbs et al. 2005; Maeda, Hobbs et al. 2005), and has been implicated in glioma (Rovin and Winn 2005). ZBTB7A can act as a transcriptional repressor or activator depending on the promoter context.

(b) NFKB (http://www.factorbook.org/mediawiki/index.php/NFKB) is very widely expressed in animal cells and well-known to play a role in immune and stress responses 16724054.

An interesting sideline to the examination of cdc42 family, is that the gene tree derived from F5-CAGE expression clustering, bears little similarity to the TreeFam phylogenetic tree for this family, suggesting that expression patterns cannot be readily used as phenotypic markers for phylogenetic inference. This conclusion of limited phylogenetic signal contained in expression pattern data, is also in some ways a logical extension of Figure 6a, which shows little differentiation in paralog expression patterns can be seen for taxa older than Amniota, and high variability in expression distances seen for younger taxa.

Future work: we are currently using the ENCODE dataset in conjunction with F5-CAGE to examine the over-all correlation between promoter structure as defined by ENCODE Tfbs and DNASEI footprints and expression patterns, in particular in gene duplicates. However, this analysis will be finalized when full ENCODE dataset, including mouse data, is released in the future.

We are also currently investigating the relationship between TSS divergence between paralogs and their expression pattern divergence. Previous published work suggested that sharing TSS increases expression correlation (Park and Makova 2009), however that study used out-dated methodology for gene family identification and alignment (CLUSTALW), out-dated gene expression datasource, and was limited only to the ENCODE pilot regions (Birney, Stamatoyannopoulos et al. 2007).

Finally, as the assignment of ortholog tissues in an unsupervised approach is based on matching expression patterns of gene orthologs, this question is closely related to the broader issue of ortholog conjecture (Studer and Robinson-Rechavi 2009). Recently, there has been a heated debate with reference to the ortholog conjecture, in particular with reference to the use of gene ontology data for tests of the hypothesis (Nehrt, Clark et al. 2011; Altenhoff, Studer et al. 2012; Thomas, Wood et al. 2012). We are currently testing ortholog conjecture using F5-CAGE expression data, hoping that our expression encyclopedia will be a much better proxy of gene function between species.

Data access:

--- consortium provides details

Methods:

F5-CAGE WP4 expression tables

TreeFam8 database
The release 8 of the TreeFam database (2012.02.10) includes 79 species (based on Ensembl v.54). There are 1,539,621 genes in total in 16,064 different TreeFam families.

Expression distances and hierarchical clustering
Expression distances were calculated using Bioconductor package bioDist Release (2.11).

Supercomputer resources
The Swedish National Infrastructure for Computing (SNIC) coordinates and develops high end computing capacity for Swedish research. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

ENCODE data
Multi-cell-line clustered Tfbs data (file wgEncodeRegTfbsClusteredV2.bed, called from henceforth TfbsClusteredV2, published by the ENCODE consortium in 2012) were used to analyze promoter regions of gene of interest within 500 bps window (-/+ 250 bps from the TSS). TfbsClusteredV2 includes 2.7 million peaks, which combine data from the Myers Lab at the HudsonAlpha Institute for Biotechnology and by the labs of Michael Snyder, Mark Gerstein and Sherman Weissman at Yale University; Peggy Farnham at UC Davis; and Kevin Struhl at Harvard, Kevin White at The University of Chicago, and Vishy Iyer at The University of Texas Austin. TfbsClusteredV2 includes data for 148 TFs including CTCF and PolII.

Promoter regions of several genes were also inspected manually using the UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly, which included tracks with the identical subset of ENCODE data.

Data access:

Acknowledgments:

Disclosure declaration:

The authors declare no competing interests.

FIGURE LEGENDS:

Fig1. Brain samples group into age-related clusters.
Human brain samples cluster in a PCA analysis (PC1 versus PC2) into three age related clusters. Samples are shown using colored marks (adult - red, fetal - green, newborn - blue). These clusters are illustrated using oval shapes with boundaries hued in respective colors. Few samples are outliers to the overall PCA clustering pattern, for example: putamen, cerebellum, spinal cord, medulla oblongata. Interestingly, most outlying samples correspond to the brain stem, suggesting different developmental patterns for forebrain, midbrain and hindbrain.

Fig2. Hierarchical clustering of human tissue samples (Spearman correlation).
CNS samples cluster together, forming two distinct sub-clusters: (a) medulla oblongata, spinal cord, brain glands, and eye cluster; (b) other brain tissues. The single outlying substantia nigra sample may be a technical problem since other substantia nigra samples cluster more closely with the CNS tissues. Several additional clusters are clearly visible in human tissues, and have been annotated at the dendrogram level.

Fig3. Hierarchical clustering of mouse samples (Spearman correlation).
Overall, except for brain samples which form distinct clusters in both species, system-by-system logic of clustering is not very well-mirrored between human and mouse, despite the same correlation coefficient used (Spearman).

Fig4. Inter-species hierarchical clustering of samples (ISHC).
Panel (**a**) shows the tree for inter-species hierarchical clustering of samples (ISHC). Panel (**b**) shows histogram with Pearson R distribution (all-against-all sample comparisons) for hs (human), mm (mouse), and hs-mm (human-mouse).

Fig5. Paralog expression divergence between in normal versus cancer cells.
Paralog expression divergence offers unexpected evidence for global transcriptional disregulation in cancer cell lines. Panel (a) human tissues, panel (b) primary cells, panel (c) cancer cell lines. Expression distances are calculated using Pearson R.

Fig6. **Expression pattern divergence bar-plot between paralogs.**

Fig7. Young genes are tissue-specific, old genes are housekeeping.
As organisms grew in complexity and new tissues formed, new genes became more and more tissue-specific in their epression patterns. However, there are exceptions to this trend: in particular, taxa Eutheria and Deuterostomia, suggesting unique evolutionary features of genes associated with these taxa.

Fig8. Evolutionary history of mammalian tissues: brain samples.

Fig9. Evolutionary history of mammalian tissues: reproductive system.

Fig10. Expression pattern evolution rates vary widely between tissues
Tissues can be subdivided into three groups of differing expression pattern evolution rates, as calculated by Eucledian distance between paralog pairs. The three groups are: (a) dynamic (for example, consistently including thumus, adipose, liver, pancreas and blood), (b) intermediate and (c) static. Brain samples are split between clusters (b) and (c).

Fig11. Evolution of expression patterns in the cdc42 family
Panel (a) shows heatmap of expression patterns for the cdc42 family in human tissues. Panel (b) shows heatmap of the values for Jaccard-index for ENCODE Tfbs in pairwise comparisons. Table in panel (c) shows actual values of the Jaccard-index.

Fig12. Comparative multivariate analysis of gene family expression patterns
To examine overall differences in average gene family expression characteristics between F5-CAGE tissues, primary cells, and cancer cell-lines, we performed multivariate analysis of the following 7 variables associated with TreeFam families in the F5-CAGE dataset (Figure 12a): number of human family members (family_size), average expression in human tissues (AVG_tissues), average expression in cells (AVG_cells), average expression in cancer (AVG_cancer), preferential expression measure in tissues (PEM_tissues), preferential expression measure in cells (PEM_cells), and preferential expression measure in cancer (PEM_cancer).

Fig13. Self-organizing map (SOMs) identifies clusters of families with similar characteristic expression characteristics in the three human sample types
6 by 6 rectangular SOM (R package Kohonen) was used to classify 682 TreeFam manually-annotated familyA families into groups with similar size, average expression and preferential expression in the three human sample types. Classification of all families into a prototype number is given in TableS_SOMa (prototypes are numbered/ordered left to right first, and bottom to top second). “Counts plot” in panel (a) illustrates final prototype sizes, while panel (b) illustrates average variable values for all TreeFam families clustering into a given prototype. Exact prototype sizes are as follows (in ascending numerical order): 2, 7, 8, 3, 3, 5, 7, 12, 16, 7, 10, 6, 13, 9, 28, 9, 23, 16, 12, 37, 55, 4, 19, 22, 28, 26, 56, 38, 8, 5, 57, 78, 32, 14, 4, 2. Prototype 32 is the most populated one, with 78 TreeFam families (top row, 2nd from the left). Exact average variable values in each prototype are given in TableS_SOMb.

FIGURES:

TABLES:

Table 1. The structure of the F5 encyclopedia of expression patterns.

Table 2. Brain samples in human cluster according to developmental stage.

Table 3. Two strategies for ortholog tissue assignment.
Comparison of two different approaches for ortholog tissue identification: “name matching” (NM) and inter-species hierarchical clustering. Eight name clusters are also simple inter-species hierarchical clusters (signified by “YES” in the last column and green color in the first). However, many name clusters are not recovered as inter-species hierarchical clusters (Fig4) (signified by “NO” in the last column), while two are split and difficult to classify (signified by “Not certain” in the last column).

Table 4. Gene families with differential tissue-specific expression in tissues versus primary cells and cancer cell lines.

Table 5. Families with differential average expression in tissues versus primary cells and cancer cell lines.

Table 6. Strongest associations between timing of gene duplication and expression site in F5 CAGE

Table 7. Example genes behind strongest associations between timing of gene duplication and expression site in F5 CAGE

Table 8. cdc42 family ENCODE Tfbs in a 500 bp window (-/+ 250 bps from the TSS).

Table 9. Tfbs Jaccard index for the cdc42 family ENCODE Tfbs

Table 3. Comparison of ortholog human-mouse tissue clusters obtained by simple name matching (NM-clusters), with those inferred using inter-species hierarchical clustering of samples (ISHC).

Name matching (NM) cluster	No. of human samples in the NM cluster	No. of mouse samples in the name cluster	NM cluster recovered by the ISHC procedure
lung	3	14	NO
colon	3	1	Not certain
diencephalon	1	2	NO
skin	3	5	YES
kidney	2	10	NO
liver	2	16	YES
uterus	2	2	NO
tongue	3	1	YES
stomach	1	10	NO
ovary	1	3	NO
aorta	1	1	NO
heart	3	15	YES
prostate	1	1	NO
vagina	1	1	NO
spleen	2	6	NO
placenta	1	2	NO
pancreas	1	12	YES
testis	2	11	NO
small intestine	2	1	Not certain
medulla oblongata	3	2	NO
adrenal gland	1	7	NO
spinal cord	4	1	NO
pituitary gland	1	8	YES
thymus	2	14	YES
hippocampus	3	2	NO
cerebellum	3	38	NO
RNA	2	2	YES
eye	6	9	NO
epididymis	1	3	NO
Total: 29 clusters	Total: 58 samples	Total: 186 samples	Total: 8-YES, 19-NO

Table 4. Gene families with differential tissue-specific expression in tissues versus primary cells and cancer cell lines.
Average PEM values, across all genes in a given family and all samples in a given category, are given in: T - tissues, PC - primary cells, CCL - cancer cell lines. Top ten families, with at least ten members, for each of the four differentially expressed categories are given.

top 10 PEM high in cancer				
TF101527	Eukaryotic translation initiation factor 4 gamma	1.0	16.2	24.0
TF106303	translocation protein isoform 1	9.6	18.5	18.5
TF105310	wingless-type MMTV integration site family	7.4	12.9	12.9
TF105321	glutathione S-transferase A/M/P1	8.3	12.1	12.1
TF106495	Rho guanine nucleotide exchange factor (GEF) 1	11	9.8	10.3
TF105128	dual specificity phosphatase 3/14/18/19/21/26	7.4	10.0	10.0
TF105122	dual specificity phosphatase 1/2/4-7/9/10	9.7	8.9	8.9
TF106481	Trinucleotide repeat containing 9 (TNRC9)/Langerhans cell protein (LCP1)/ Granulosa cell HMG box protein 1 (GCX1)/Thymocyte sele	5.6	8.8	8.8
TF105351	p21 (CDKN1A)-activated kinase 1-3	5.8	8.8	8.8
TF106499	gonadotropin-releasing hormone receptor/arginine vasopressin receptor	5.1	8.5	8.5
top 10 PEM high in tissues				
TF106108	fucosyltransferase 8 (alpha (1	6.0	0.0	0.0
TF106311	N-acetyltransferase 1/2 (arylamine N-acetyltransferase)	2.7	1.5	1.5
TF105318	glutathione peroxidase	6.0	3.8	3.8
TF101524	Eukaryotic translation initiation factor 4A	2.8	1.8	1.8
TF106173	histone deacetylase 6/histone deacetylase 10	2.0	1.4	1.4
TF101007	Cyclin G/I	4.2	3.2	3.2
TF101031	Cyclin-dependent kinase-like 1/2/3	3.0	2.4	2.4
TF101069	Cell division cycle associated protein 4	3.5	2.9	2.9
TF101128	RAD6 homolog	2.0	1.6	1.6
TF101155	cytoplasmic linker associated protein	2.7	2.3	2.3
top 10 PEM high in cancer, low in tissues				
TF101527	Eukaryotic translation initiation factor 4 gamma	1.0	16.2	24.0
TF105042	heat shock 70kDa protein 2/6/7	2.2	5.1	5.1
TF106303	translocation protein isoform 1	9.6	18.5	18.5
TF105750	stomatin (EPB72)-like 2	3.2	6.2	6.2
TF105310	wingless-type MMTV integration site family	7.4	12.9	12.9

TF106002	epidermal growth factor receptor / v-erb-b2 erythroblastic leukemia viral oncogene	2.7	4.7	4.7
TF106450	REST corepressor 12/3	2.1	3.5	3.5
TF106499	gonadotropin-releasing hormone receptor/arginine vasopressin receptor	5.1	8.5	8.5
TF101080	Septin 6/8/10/11	4.1	6.6	6.6
TF105317	glypican family	3.8	6.1	6.1
top 10 PEM high in tissues, low in cancer				
TF106108	fucosyltransferase 8 (alpha (1	6.0	0.0	0.0
TF106311	N-acetyltransferase 1/2 (arylamine N-acetyltransferase)	2.7	1.5	1.5
TF105318	glutathione peroxidase	6.0	3.8	3.8
TF101524	Eukaryotic translation initiation factor 4A	2.8	1.8	1.8
TF106173	histone deacetylase 6/histone deacetylase 10	2.0	1.4	1.4
TF101007	Cyclin G/I	4.2	3.2	3.2
TF101031	Cyclin-dependent kinase-like 1/2/3	3.0	2.4	2.4
TF101069	Cell division cycle associated protein 4	3.5	2.9	2.9
TF101128	RAD6 homolog	2.0	1.6	1.6
TF101155	cytoplasmic linker associated protein	2.7	2.3	2.3

Table 5. Families with differential average expression in tissues versus primary cells and cancer cell lines.
Fold difference given in the “fold” column. Average TPM values, across all genes in a given family and all samples in a given category, are given in: T - tissues, PC - primary cells, CCL - cancer cell lines. Top ten families, with more than two human members, for each of the four differentially expressed categories are given.

high in CCL, low in T		**fold**	**T**	**PC**	**CCL**
TF106434	Ubiquitin-like	18.8	1.0	10.9	18.8
TF101116	Ubiquitin-conjugating enzyme E2 C	13.7	3.3	18.5	45.2
TF105231	kinesin family member 18A	11.9	1.5	6.3	17.5
TF105232	kinesin family member 20A/23 (MKLP1)	11.0	2.8	12.7	30.9
TF101001	Cyclin B	10.3	6.7	30.1	69.5
TF105331	aurora kinase	9.6	0.7	2.6	6.9
TF101002	Cyclin A	9.4	2.4	9.5	22.6
TF101021	Cyclin-dependent kinase 1/2/3	9.0	2.8	9.4	25.1
TF101170	F-box only protein 5	8.1	2.1	5.5	17.3
TF101076	Cell division cycle associated 7	7.5	3.3	6.9	24.9
high in T, low in PC and CCL		**fold**	**T**	**PC**	**CCL**
TF105403	A kinase (PRKA) anchor protein 3/4	63.4	1.4	0.0	0.0
TF105451	retinol dehydrogenase 8 (all-trans)	9.4	0.2	0.0	0.0
TF101036	Cyclin-dependent kinase 5 activator	5.6	36.6	2.5	4.1
TF101074	F-box/WD-repeat protein 7	4.9	17.0	1.6	1.9
TF105225	kinesin family member 5 (KHC)	3.3	131	15.6	24.2
TF106489	Patched	3.0	2.9	0.3	0.7
TF106496	Adenomatous polyposis coli	2.7	21.9	3.7	4.5
TF105285	flavin containing monooxygenase	2.4	4.1	1.0	0.7
TF105395	integrin beta 1 binding protein 3	2.3	21.4	4.5	4.7
TF105424	dual oxidase	2.3	4.5	1.3	0.7
high in PC and CCL, low in T		**fold**	**T**	**PC**	**CCL**
TF106434	Ubiquitin-like	29.7	1.0	10.9	18.8
TF101116	Ubiquitin-conjugating enzyme E2 C	19.3	3.3	18.5	45.2
TF105231	kinesin family member 18A	16.2	1.5	6.3	17.5
TF105232	kinesin family member 20A/23 (MKLP1)	15.5	2.8	12.7	30.9

TF101001	Cyclin B	14.8	6.7	30.1	69.5
TF101002	Cyclin A	13.4	2.4	9.5	22.6
TF105331	aurora kinase	13.3	0.7	2.6	6.9
TF101021	Cyclin-dependent kinase 1/2/3	12.3	2.8	9.4	25.1
TF101142	Cyclin-dependent kinases regulatory subunit	10.7	16.6	55.1	123
TF101170	F-box only protein 5	10.7	2.1	5.5	17.3
high in T, low in CCL		**fold**	**T**	**PC**	**CCL**
TF105403	A kinase (PRKA) anchor protein 3/4	98.9	1.4	0.0	0.0
TF105451	retinol dehydrogenase 8 (all-trans)	13.6	0.2	0.0	0.0
TF101036	Cyclin-dependent kinase 5 activator	9.0	36.6	2.5	4.1
TF101074	F-box/WD-repeat protein 7	8.9	17.0	1.6	1.9
TF105424	dual oxidase	6.7	4.5	1.3	0.7
TF105569	Zinc finger protein 106 homolog	6.2	36.7	12.2	5.9
TF105285	flavin containing monooxygenase	6.0	4.1	1.0	0.7
TF105225	kinesin family member 5 (KHC)	5.4	131	15.6	24.2
TF106496	Adenomatous polyposis coli	4.9	21.9	3.7	4.5
TF105395	integrin beta 1 binding protein 3	4.5	21.4	4.5	4.7

Table 8. cdc42 family ENCODE Tfbs in a 500 bp window (-/+ 250 bps from the TSS).

No	EntrezID	Name, TSS location	All Tfbs	Strong Tfbs
1	23433	ras homolog family member Q,RHOQ, 46769867,chr2	Promoter=chr2:46769617-46770116, includes: HA-E2F1, Pol2, ELF1_(SC-631), ZNF263, ZEB1_(SC-25388), Sin3Ak-20, CCNT2, Nrf1, E2F1, c-Myc, E2F6_(H-50), E2F6, Egr-1, **ZBTB7A**_(SC-34508), TAF1, EBF	Promoter=chr2:46769617-46770116, includes (at 0.75 quantile cut-off): HA-E2F1, ZNF263,
2	391	ras homolog family member G,RHOG, -3848208,chr11	Promoter=chr11:3861964-3862463, includes: HA-E2F1, CCNT2, Pol2, HEY1, ELF1_(SC-631), Pol2-4H8, GABP, PU.1, Egr-1, HA-E2F1, **NFKB**	Promoter=chr11:3861964-3862463, includes (at 0.75 quantile cut-off): ,
3	57381	ras homolog family member J,RHOJ, 63671145,chr14	Promoter=chr14:63670852-63671351, includes: KAP1, c-Jun, c-Fos, JunD, GATA-2, CTCF, HDAC2_(SC-6296), Rad21, p300, ELF1_(SC-631), Pol2(b), SRF, Pol2	Promoter=chr14:63670852-63671351, includes (at 0.75 quantile cut-off): c-Jun, Pol2(b),
4	5879	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1),RAC1, 6414126,chr7	Promoter=chr7:6413876-6414375, includes: TFIIIC-110, TBP, Pol2, RPC155, BDP1, HA-E2F1, YY1, YY1_(C-20), HMGN3, E2F4, p300, E2F1, TAF1, c-Myc, GABP, Egr-1, ELF1_(SC-631), NANOG_(SC-33759), CCNT2, Pol2, Pol2-4H8, HEY1, YY1_(C-20)	Promoter=chr7:6413876-6414375, includes (at 0.75 quantile cut-off): TBP, HA-E2F1, YY1, YY1_(C-20), E2F1, GABP,
5	5880	ras-related C3 botulinum toxin substrate 2 (rho family, small GTP binding protein Rac2),RAC2,-37621312,chr22	Promoter=chr22:37640056-37640555, includes: Pol2-4H8, TAF1, HEY1, **NFKB**, Pol2, POU2F2, Oct-2, Sin3Ak-20, c-Fos, TBP, GABP, ELF1_(SC-631), ETS1, E2F6_(H-50), PU.1, c-Myc, Max, Egr-1, IRF1, PAX5-C20, Pbx3, EBF1_(C-8), **ZBTB7A**_(SC-34508), TCF12	Promoter=chr22:37640056-37640555, includes (at 0.75 quantile cut-off): **NFKB**, Pol2,
6	5881	ras-related C3 botulinum toxin substrate 3 (rho family, small GTP binding protein Rac3),RAC3, 79989532,chr17	Promoter=chr17:79989282-79989781, includes: ETS1, Sin3Ak-20, **ZBTB7A**_(SC-34508), Egr-1, SRF	Promoter=chr17:79989282-79989781, includes (at 0.75 quantile cut-off): **ZBTB7A**_(SC-34508),
7	998	cell division cycle 42 (GTP binding protein, 25kDa),CDC42, 22379120,chr1	Promoter=chr1:22378870-22379369, includes: TFIIIC-110, Nrf1, Pol2, E2F6_(H-50), IRF1, RFX5_(N-494), ELF1_(SC-631), SP1, GABP, p300, PU.1, JunD, TBP, **NFKB**, HMGN3, E2F4, PAX5-C20, CCNT2, USF-1, USF1_(SC-8983), Egr-1, c-Myc, GTF2F1_(RAP-74), YY1_(C-20), YY1, c-Jun, PAX5-N19, Sin3Ak-20, Pol2(b), eGFP-JunD, **ZBTB7A**_(SC-34508), NRSF, Pol2-4H8, TAF1, SIX5, ZEB1_(SC-25388), Pol2(phosphoS2), HEY1, EBF1_(C-8), TCF12, POU2F2, Oct-2	Promoter=chr1:22378870-22379369, includes (at 0.75 quantile cut-off): Nrf1, Pol2, ELF1_(SC-631), PU.1, HMGN3, eGFP-JunD, Pol2-4H8, TAF1, HEY1,

	RHOQ	RHOG	RHOJ	RAC1	RAC2	RAC3	CDC42
RHOQ	1.00	0.24	0.07	0.28	0.25	0.17	0.23
RHOG	0.24	1.00	0.10	0.35	0.31	0.07	0.21
RHOJ	0.07	0.10	1.00	0.10	0.09	0.06	0.12
RAC1	0.28	0.35	0.10	1.00	0.25	0.04	0.34
RAC2	0.25	0.31	0.09	0.25	1.00	0.16	0.43
RAC3	0.17	0.07	0.06	0.04	0.16	1.00	0.07
CDC42	0.23	0.21	0.12	0.34	0.43	0.07	1.00

Table 9. Jaccard index for the cdc42 family

Bibliography

Altenhoff, A. M., R. A. Studer, et al. (2012). "Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs." PLoS computational biology **8**(5): e1002514.

Amado, F., M. J. Lobo, et al. (2010). "Salivary peptidomics." Expert review of proteomics **7**(5): 709-721.

Baron, A., A. DeCarlo, et al. (1999). "Functional aspects of the human salivary cystatins in the oral environment." Oral diseases **5**(3): 234-240.

Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.

Brawand, D., M. Soumillon, et al. (2011). "The evolution of gene expression levels in mammalian organs." Nature **478**(7369): 343-348.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences of the United States of America **95**(25): 14863-14868.

Huminiecki, L. and C. H. Heldin (2010). "2R and remodeling of vertebrate signal transduction engine." BMC Biol **8**: 146.

Huminiecki, L., A. T. Lloyd, et al. (2003). "Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases." BMC Genomics **4**(1): 31.

Huminiecki, L. and K. H. Wolfe (2004). "Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse." Genome Res **14**(10A): 1870-1879.

Jordan, I. K., L. Marino-Ramirez, et al. (2005). "Evolutionary significance of gene expression divergence." Gene **345**(1): 119-126.

Khaitovich, P., G. Weiss, et al. (2004). "A neutral model of transcriptome evolution." PLoS Biol **2**(5): E132.

Li, H., A. Coghlan, et al. (2006). "TreeFam: a curated database of phylogenetic trees of animal gene families." Nucleic Acids Research **34**(Database issue): D572-580.

Nehrt, N. L., W. T. Clark, et al. (2011). "Testing the ortholog conjecture with comparative functional genomic data from mammals." PLoS computational biology **7**(6): e1002073.

Park, C. and K. D. Makova (2009). "Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes." Genome Biology **10**(1): R10.

Pereira, V., D. Waxman, et al. (2009). "A problem with the correlation coefficient as a measure of gene expression divergence." Genetics **183**(4): 1597-1600.

Piasecka, B., M. Robinson-Rechavi, et al. (2012). "Correcting for the bias due to expression specificity improves the estimation of constrained evolution of expression between mouse and human." Bioinformatics **28**(14): 1865-1872.

Plessy, C., N. Bertin, et al. (2010). "Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan." Nature methods **7**(7): 528-534.

Quackenbush, J. (2001). "Computational analysis of microarray data." Nature reviews. Genetics **2**(6): 418-427.

Salimullah, M., M. Sakai, et al. (2011). "NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes." Cold Spring Harbor protocols **2011**(1): pdb prot5559.

Studer, R. A. and M. Robinson-Rechavi (2009). "How confident can we be that orthologs are similar, but paralogs differ?" Trends in genetics : TIG **25**(5): 210-216.

Su, A. I., M. P. Cooke, et al. (2002). "Large-scale analysis of the human and mouse transcriptomes." Proc Natl Acad Sci U S A **99**(7): 4465-4470.
Thomas, P. D., V. Wood, et al. (2012). "On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report." PLoS computational biology **8**(2): e1002386.

Figure 1

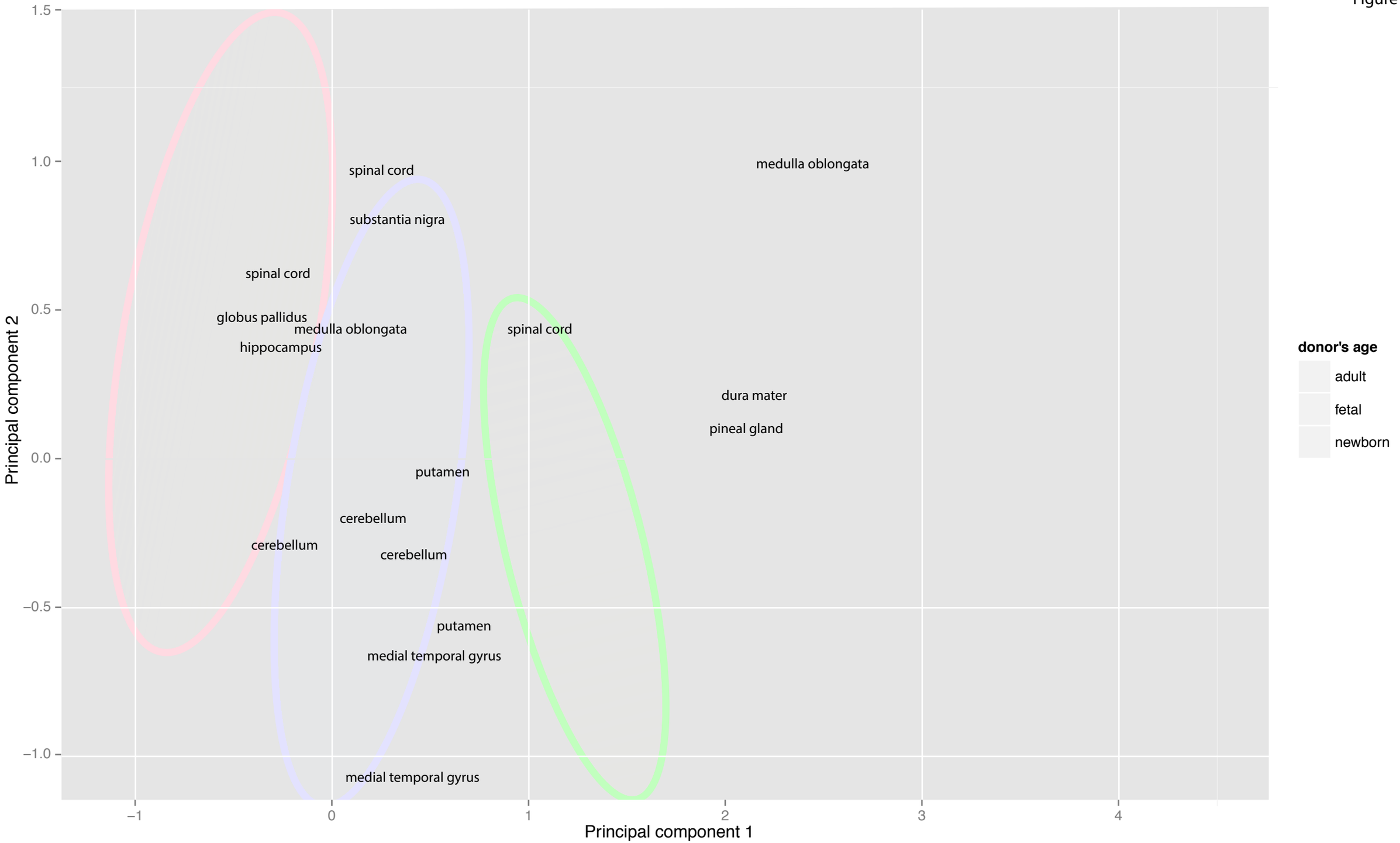

1.5
1.0
0.5
0.0
−0.5
−1.0
Principal component 2
−1
0
1
2
3
4
Principal component 1
spinal cord
substantia nigra
medulla oblongata
spinal cord
globus pallidus
medulla oblongata
hippocampus
spinal cord
dura mater
pineal gland
putamen
cerebellum
cerebellum
cerebellum
putamen
medial temporal gyrus
medial temporal gyrus
donor's age
adult
fetal
newborn

Figure 2

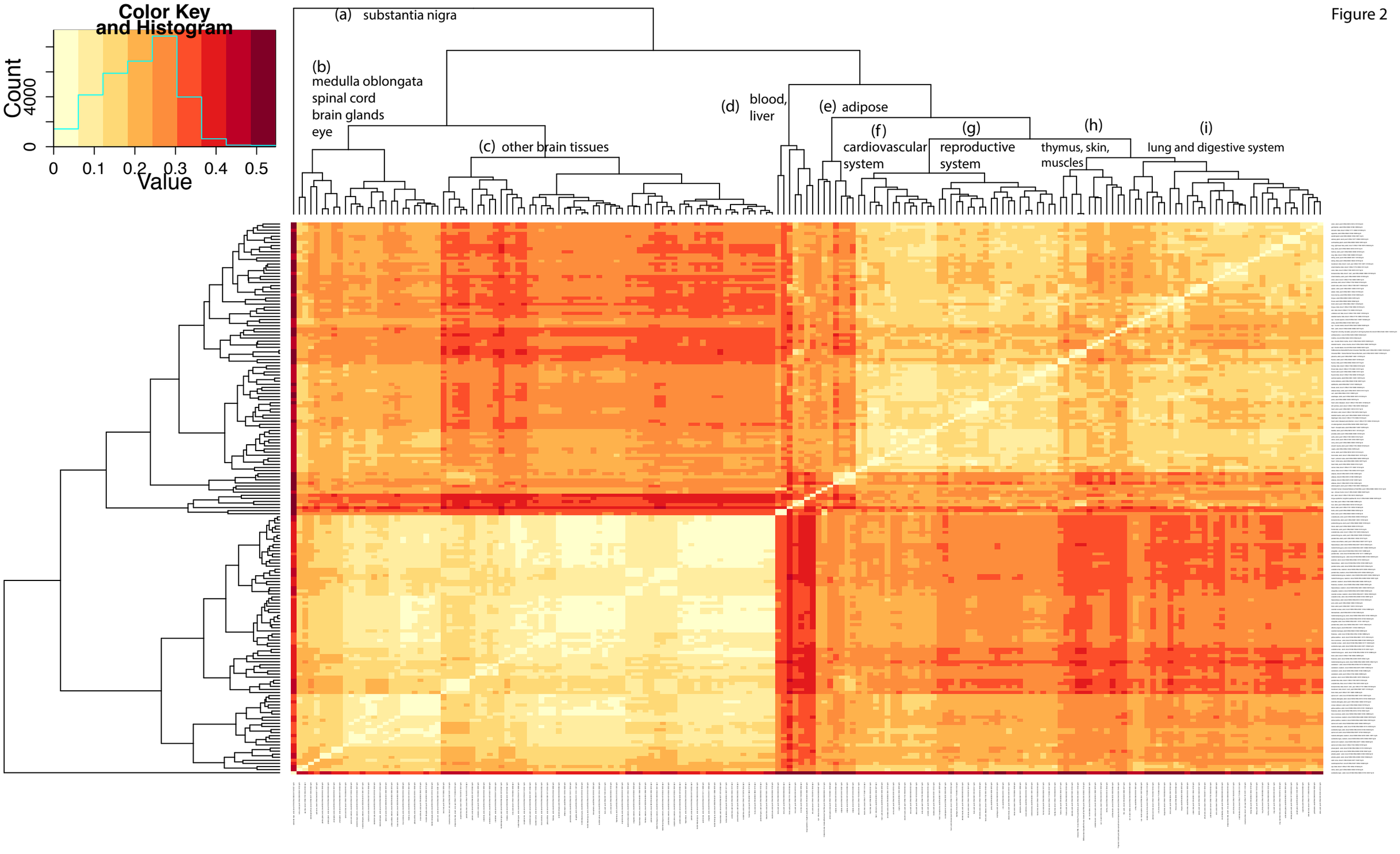

Color Key and Histogram
Count
4000
0
0
0.1
0.2
0.3
0.4
0.5
Value
(a) substantia nigra
(b)
medulla oblongata
spinal cord
brain glands
eye
(c) other brain tissues
(d) blood, liver
(e) adipose
(f) cardiovascular system
(g) reproductive system
(h) thymus, skin, muscles
(i) lung and digestive system

Figure 3

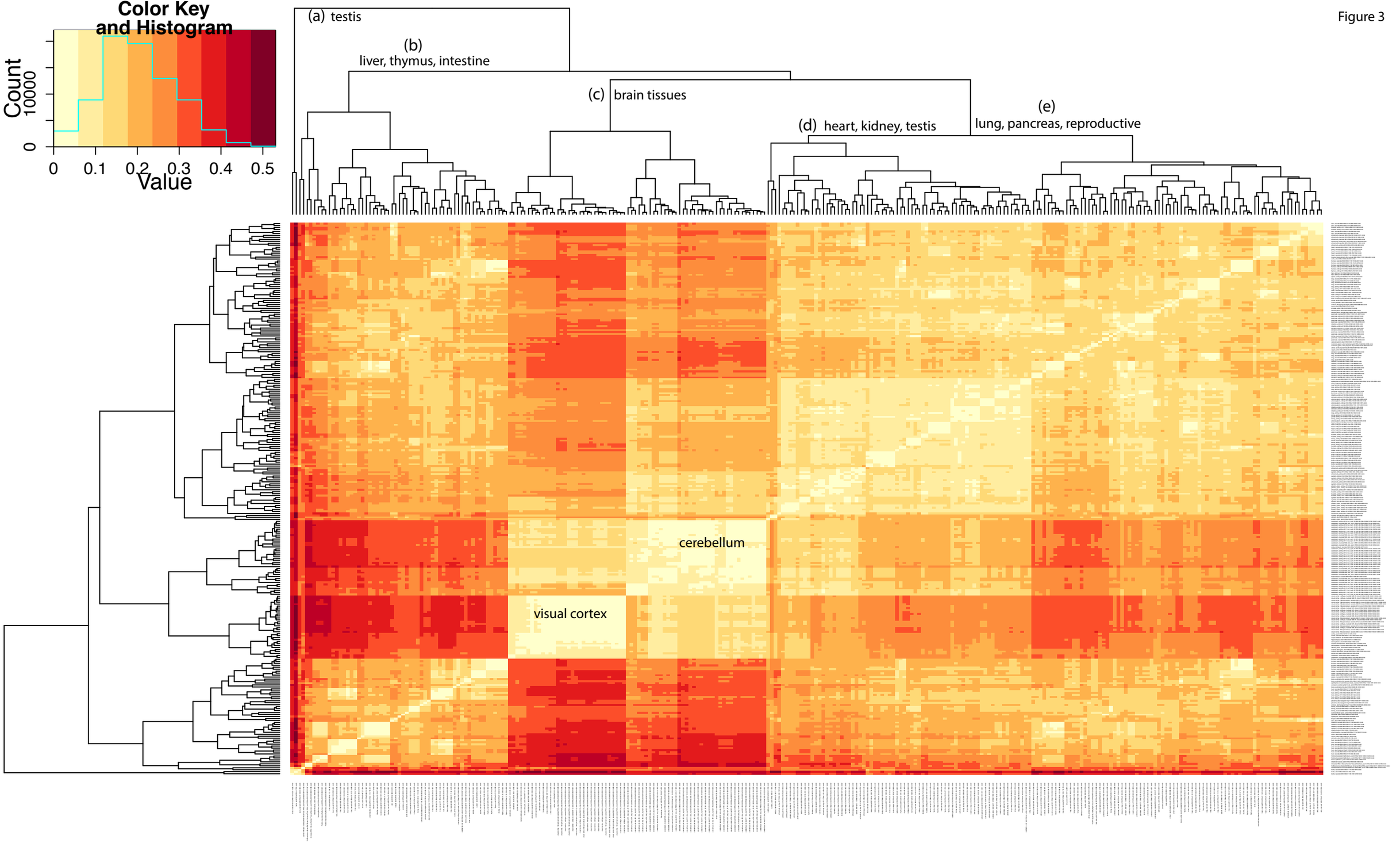
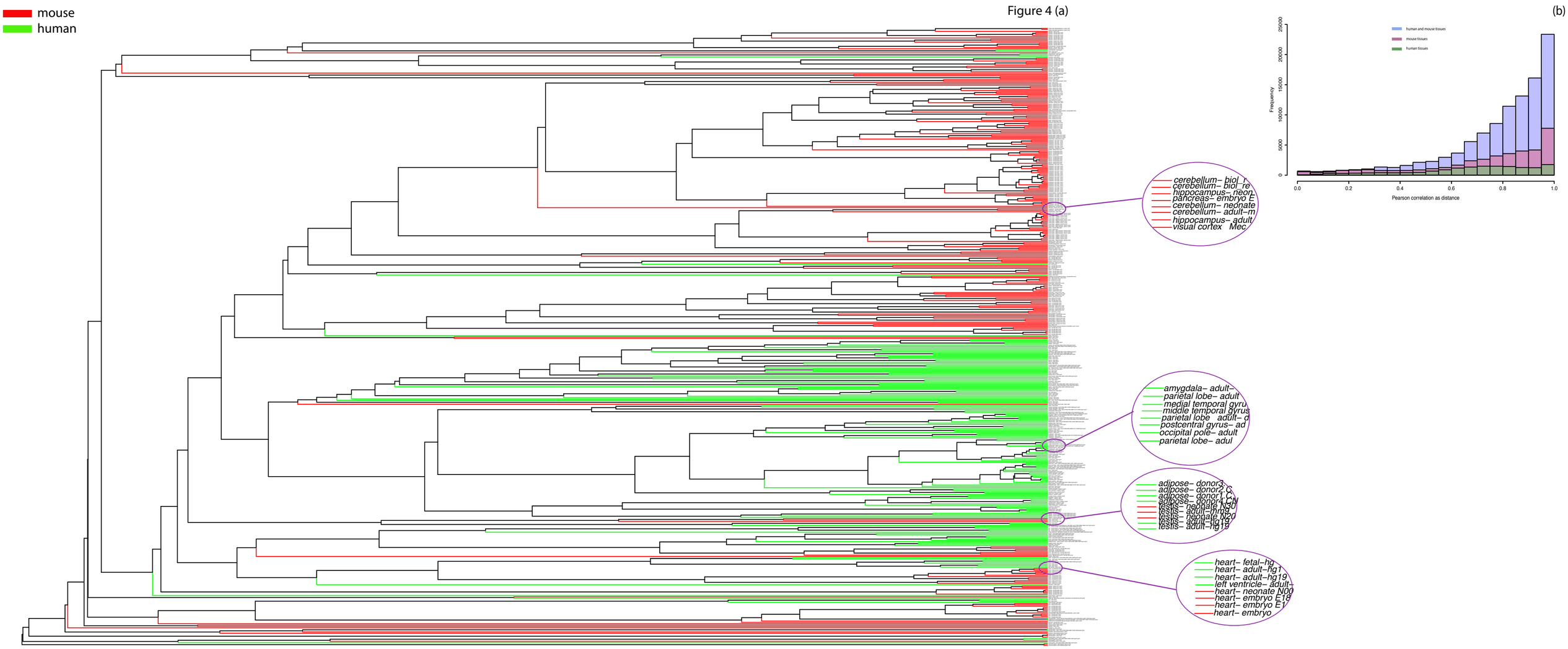

Color Key and Histogram
Count
10000
0
0
0.1
0.2
0.3
0.4
0.5
Value
(a) testis
(b)
liver, thymus, intestine
(c) brain tissues
(d) heart, kidney, testis
(e)
lung, pancreas, reproductive
cerebellum
visual cortex

Figure 4 (a)


mouse
human
cerebellum– biol_1
cerebellum– biol_re
hippocampus– neon
pancreas– embryo E
cerebellum– neonate
cerebellum– adult–m
hippocampus– adult
visual cortex Mec
amygdala– adult–
parietal lobe– adult
medial temporal gyru
middle temporal gyrus
parietal lobe adult– d
postcentral gyrus– ad
occipital pole– adult
parietal lobe– adul
adipose– donor3
adipose– donor2 C
adipose– donor1 C
adipose– donor4 CN
testis– neonate N30
testis– adult–hrs9
testis– neonate N20
testis– adult–hg19
testis– adult–hg19
heart– fetal–hg
heart– adult–hg1
heart– adult–hg19
left ventricle– adult–
heart– neonate N00
heart– embryo E16
heart– embryo E1
heart– embryo

(b)

human and mouse tissues
mouse tissues
human tissues
Frequency
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
1.0
Pearson correlation as distance

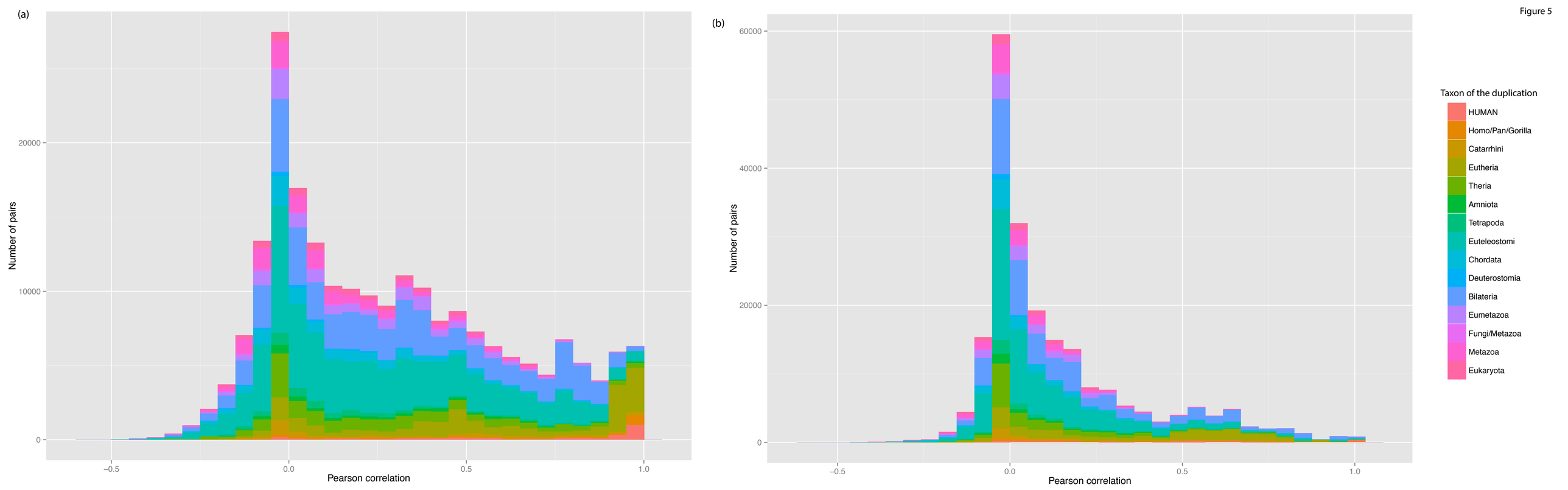

(a)
Number of pairs
20000
10000
0
−0.5
0.0
0.5
1.0
Pearson correlation
(b)
Number of pairs
60000
40000
20000
0
−0.5
0.0
0.5
1.0
Pearson correlation
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 6

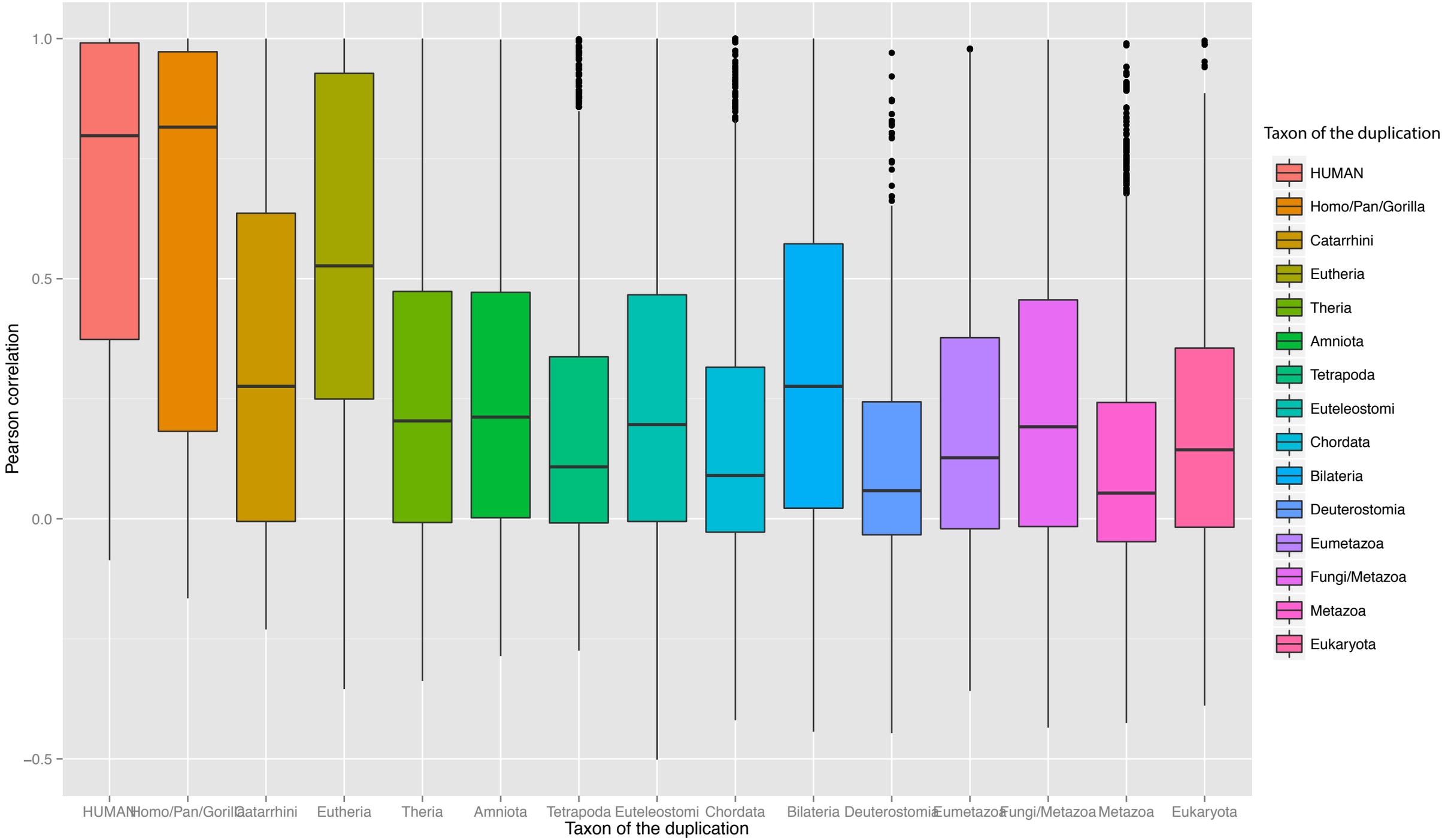

1.0
0.5
0.0
−0.5
Pearson correlation
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Bilateria
Deuterostomia
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
Taxon of the duplication
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Bilateria
Deuterostomia
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 7

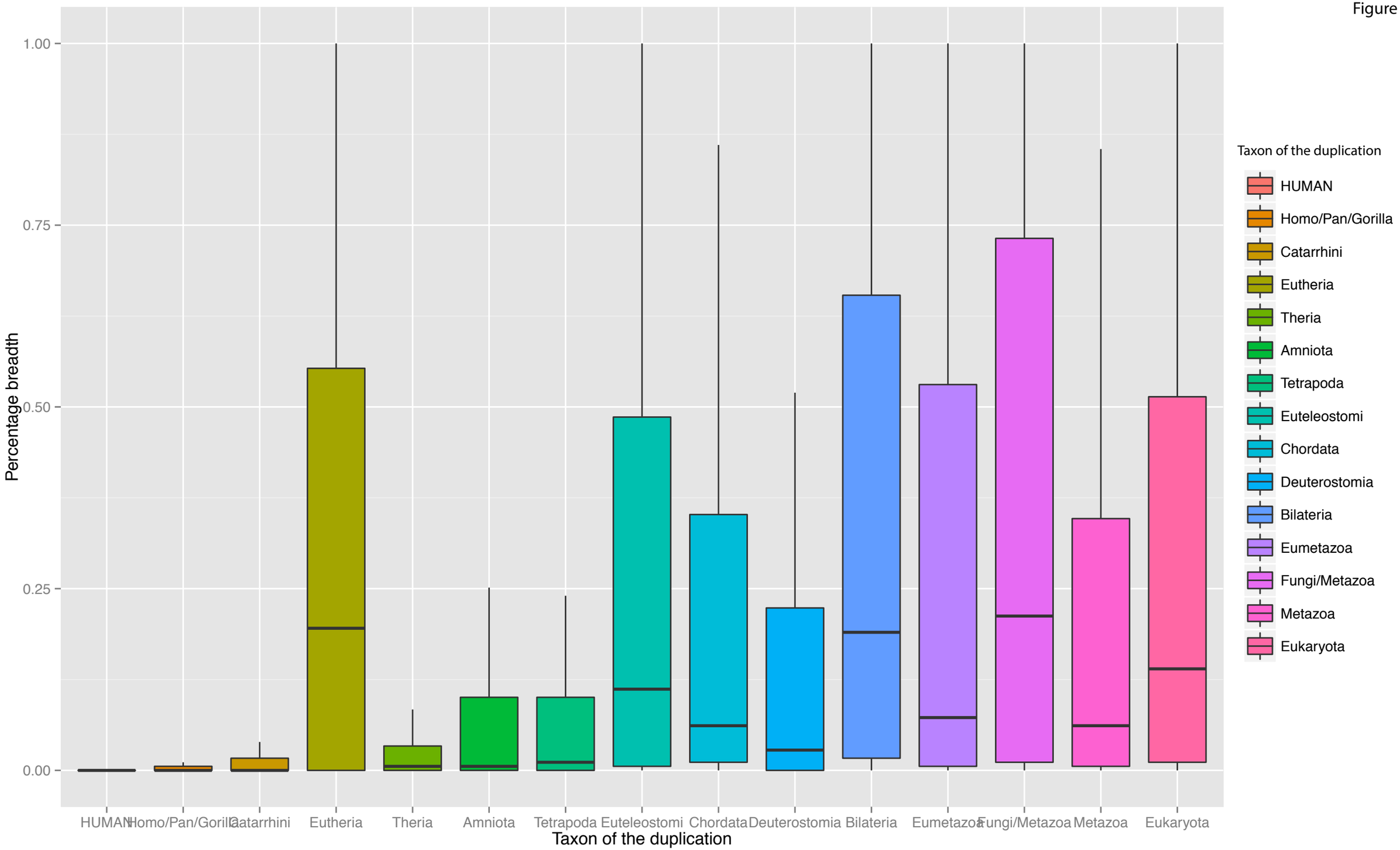

1.00
0.75
0.50
0.25
0.00
Percentage breadth
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
Taxon of the duplication
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 8, a)

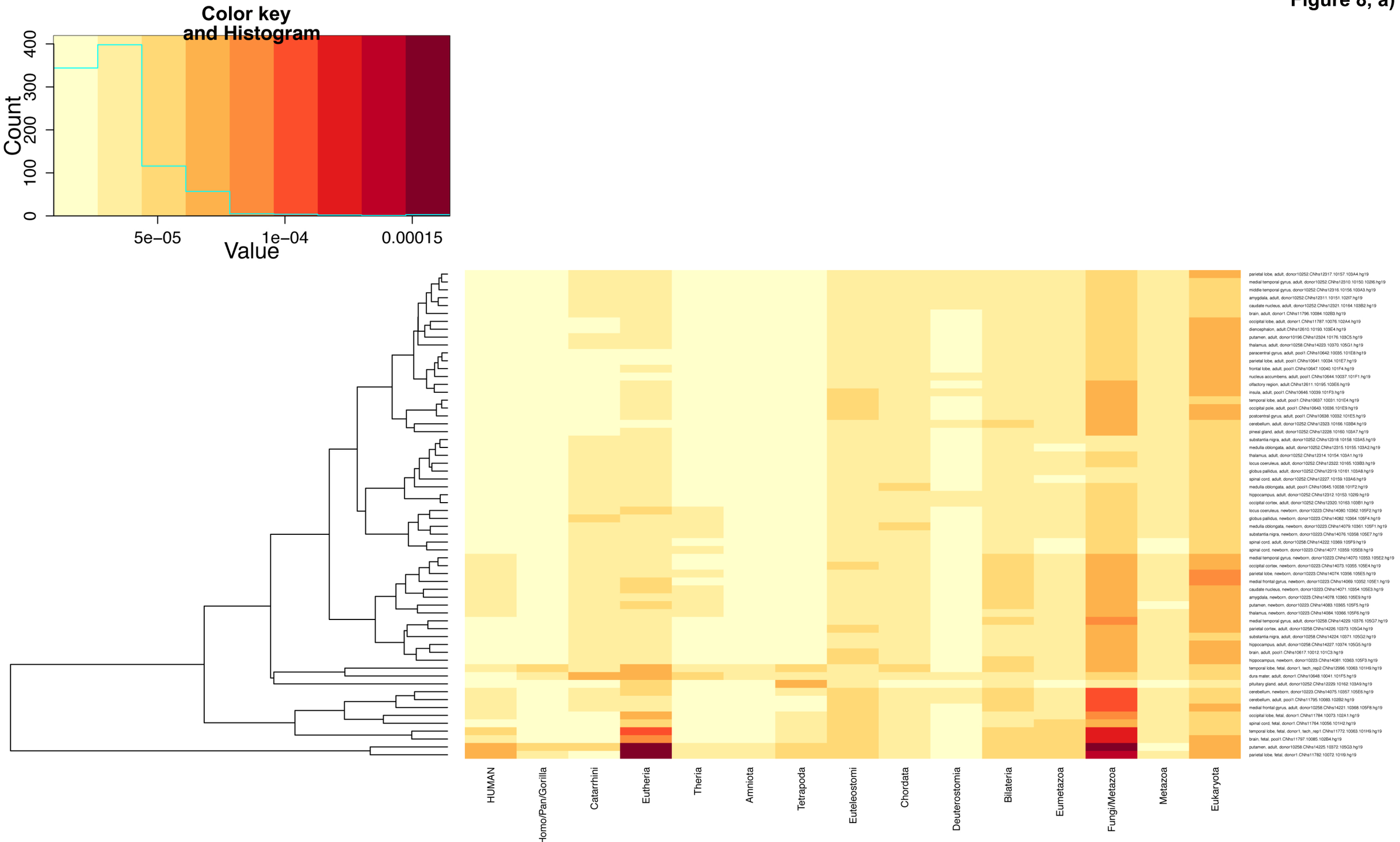

Color key
and Histogram
Count
0
100
200
300
400
5e−05
1e−04
0.00015
Value
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
medial temporal gyrus, adult, donor10252.CNhs12310.10150.102I6.hg19
middle temporal gyrus, donor10252.CNhs12316.10156.103A3.hg19
amygdala, adult, donor10252.CNhs12311.10151.102I7.hg19
caudate nucleus, adult, donor10252.CNhs12321.10164.103B2.hg19
brain, adult, donor1.CNhs11796.10084.102B3.hg19
occipital lobe, adult, donor1.CNhs11787.10076.102A4.hg19
diencephalon, adult.CNhs12610.10193.103E4.hg19
putamen, adult, donor10196.CNhs12324.10176.103C5.hg19
thalamus, adult, donor10258.CNhs14223.10370.105G1.hg19
paracentral gyrus, adult, pool1.CNhs10642.10035.101E8.hg19
parietal lobe, adult, pool1.CNhs10641.10034.101E7.hg19
frontal lobe, adult, pool1.CNhs10647.10040.101F4.hg19
nucleus accumbens, adult, pool1.CNhs10644.10037.101F1.hg19
olfactory region, adult.CNhs12611.10195.103E6.hg19
insula, adult, pool1.CNhs10646.10039.101F3.hg19
temporal lobe, adult, pool1.CNhs10637.10031.101E4.hg19
occipital pole, adult, pool1.CNhs10643.10036.101E9.hg19
postcentral gyrus, adult, pool1.CNhs10638.10032.101E5.hg19
cerebellum, adult, donor10252.CNhs12323.10166.103B4.hg19
pineal gland, adult, donor10252.CNhs12228.10160.103A7.hg19
substantia nigra, adult, donor10252.CNhs12318.10158.103A5.hg19
medulla oblongata, adult, donor10252.CNhs12315.10155.103A2.hg19
thalamus, adult, donor10252.CNhs12314.10154.103A1.hg19
locus coeruleus, adult, donor10252.CNhs12322.10165.103B3.hg19
globus pallidus, adult, donor10252.CNhs12319.10161.103A8.hg19
spinal cord, adult, donor10252.CNhs12227.10159.103A6.hg19
medulla oblongata, adult, pool1.CNhs10645.10038.101F2.hg19
hippocampus, adult, donor10252.CNhs12312.10153.102I9.hg19
occipital cortex, adult, donor10252.CNhs12320.10163.103B1.hg19
locus coeruleus, newborn, donor10223.CNhs14080.10362.105F2.hg19
globus pallidus, newborn, donor10223.CNhs14082.10364.105F4.hg19
medulla oblongata, newborn, donor10223.CNhs14079.10361.105F1.hg19
substantia nigra, newborn, donor10223.CNhs14076.10358.105E7.hg19
spinal cord, adult, donor10258.CNhs14222.10369.105F9.hg19
spinal cord, newborn, donor10223.CNhs14077.10359.105E8.hg19
medial temporal gyrus, newborn, donor10223.CNhs14070.10353.105E2.hg19
occipital cortex, newborn, donor10223.CNhs14073.10355.105E4.hg19
parietal lobe, newborn, donor10223.CNhs14074.10356.105E5.hg19
medial frontal gyrus, newborn, donor10223.CNhs14069.10352.105E1.hg19
caudate nucleus, newborn, donor10223.CNhs14071.10354.105E3.hg19
amygdala, newborn, donor10223.CNhs14078.10360.105E9.hg19
putamen, newborn, donor10223.CNhs14083.10365.105F5.hg19
thalamus, newborn, donor10223.CNhs14084.10366.105F6.hg19
medial temporal gyrus, adult, donor10258.CNhs14229.10376.105G7.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
brain, adult, pool1.CNhs10617.10012.101C3.hg19
hippocampus, newborn, donor10223.CNhs14081.10363.105F3.hg19
temporal lobe, fetal, donor1, tech_rep2.CNhs12996.10063.101H9.hg19
dura mater, adult, donor1.CNhs10648.10041.101F5.hg19
pituitary gland, adult, donor10252.CNhs12229.10162.103A9.hg19
cerebellum, newborn, donor10223.CNhs14075.10357.105E6.hg19
cerebellum, adult, pool1.CNhs11795.10083.102B2.hg19
medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
occipital lobe, fetal, donor1.CNhs11784.10073.102A1.hg19
spinal cord, fetal, donor1.CNhs11764.10056.101H2.hg19
temporal lobe, fetal, donor1, tech_rep1.CNhs11772.10063.101H9.hg19
brain, fetal, pool1.CNhs11797.10085.102B4.hg19
putamen, adult, donor10258.CNhs14225.10372.105G3.hg19
parietal lobe, fetal, donor1.CNhs11782.10072.101I9.hg19

Figure 8, b)

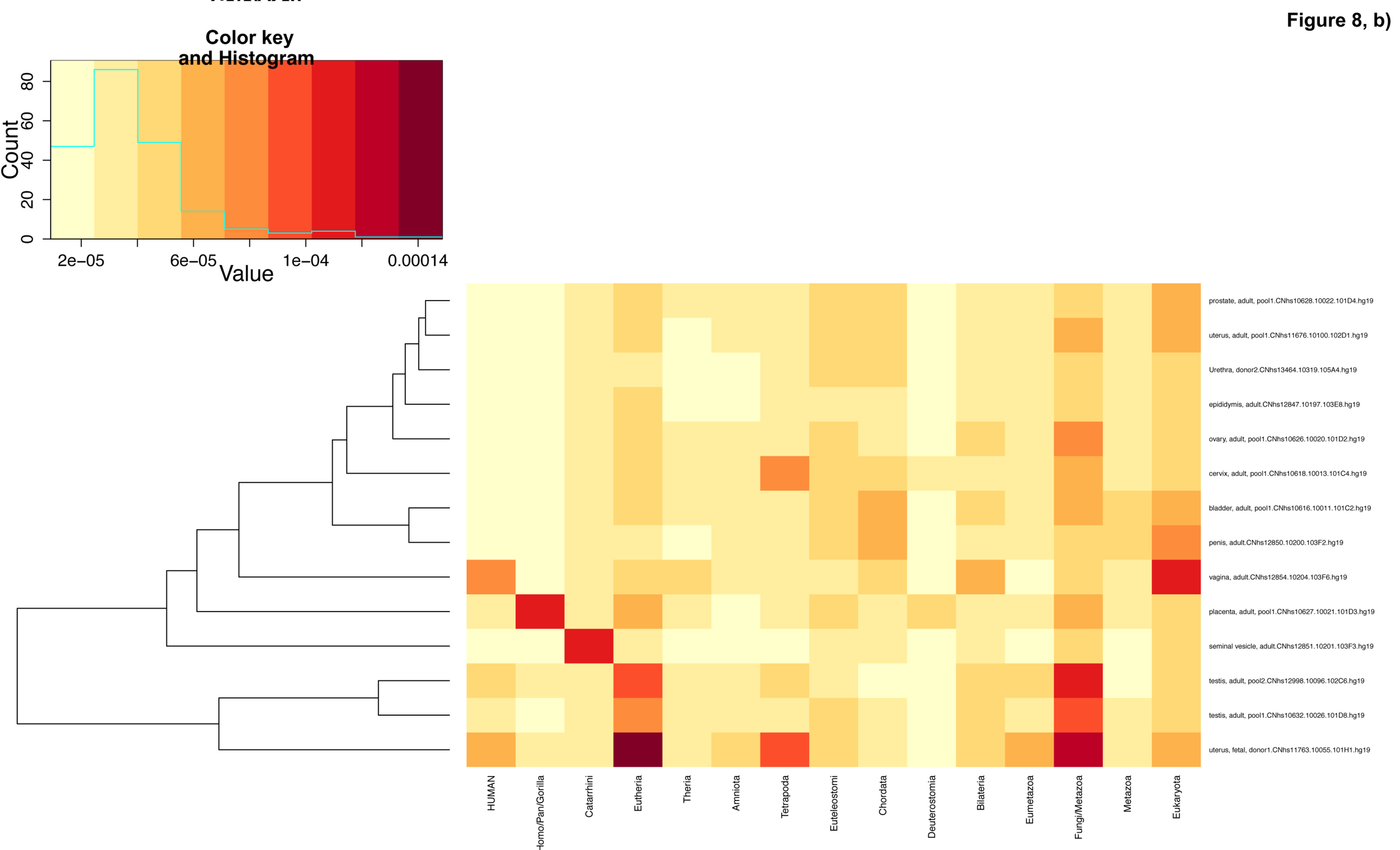

Color key
and Histogram
Count
0
20
40
60
80
2e−05
6e−05
1e−04
0.00014
Value
prostate, adult, pool1.CNhs10628.10022.101D4.hg19
uterus, adult, pool1.CNhs11676.10100.102D1.hg19
Urethra, donor2.CNhs13464.10319.105A4.hg19
epididymis, adult.CNhs12847.10197.103E8.hg19
ovary, adult, pool1.CNhs10626.10020.101D2.hg19
cervix, adult, pool1.CNhs10618.10013.101C4.hg19
bladder, adult, pool1.CNhs10616.10011.101C2.hg19
penis, adult.CNhs12850.10200.103F2.hg19
vagina, adult.CNhs12854.10204.103F6.hg19
placenta, adult, pool1.CNhs10627.10021.101D3.hg19
seminal vesicle, adult.CNhs12851.10201.103F3.hg19
testis, adult, pool2.CNhs12998.10096.102C6.hg19
testis, adult, pool1.CNhs10632.10026.101D8.hg19
uterus, fetal, donor1.CNhs11763.10055.101H1.hg19
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 10

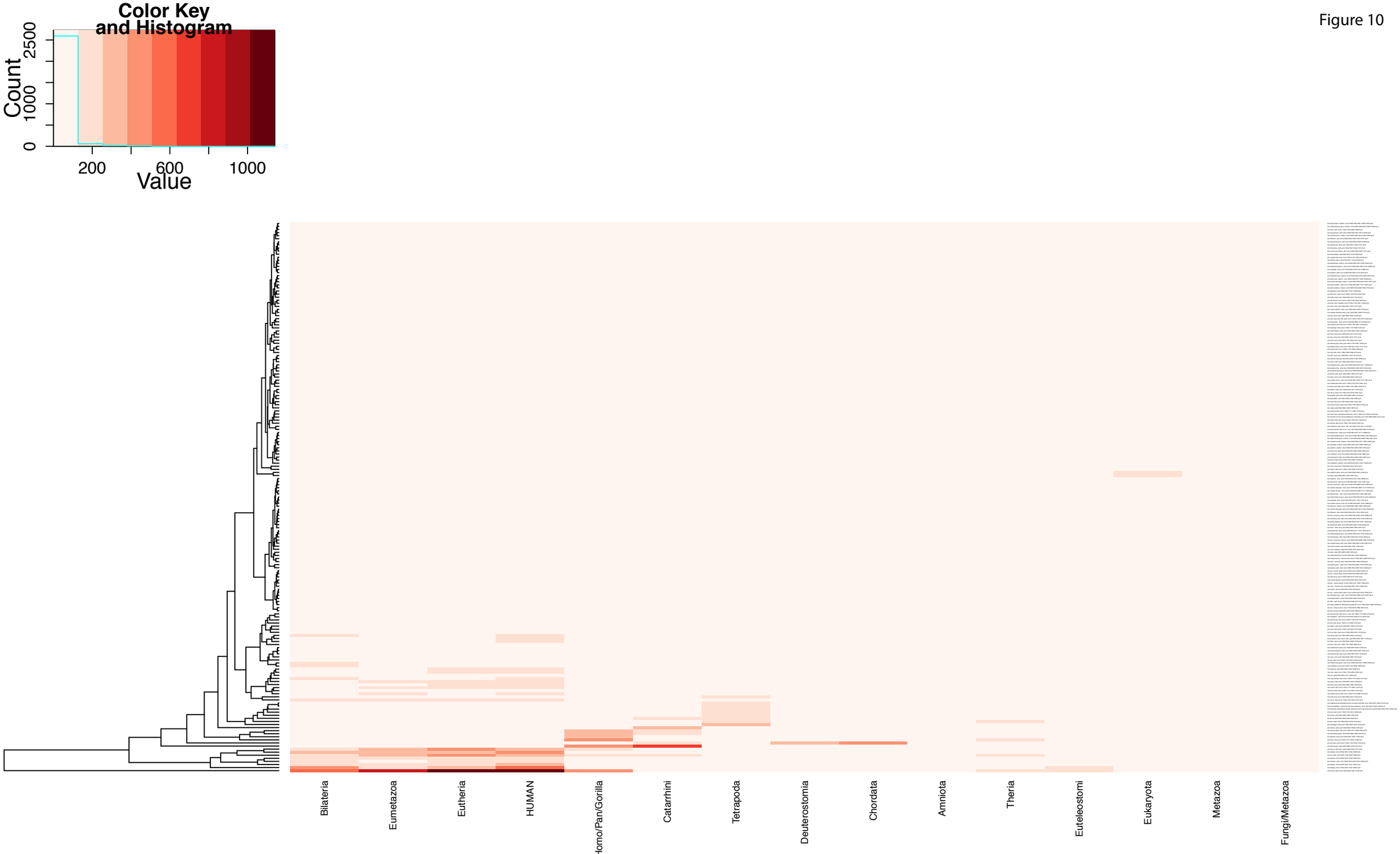

Color Key
and Histogram
Count
0
1000
2500
200
600
1000
Value
Bilateria
Eumetazoa
Eutheria
HUMAN
Homo/Pan/Gorilla
Catarrhini
Tetrapoda
Deuterostomia
Chordata
Amniota
Theria
Euteleostomi
Eukaryota
Metazoa
Fungi/Metazoa

Figure S1

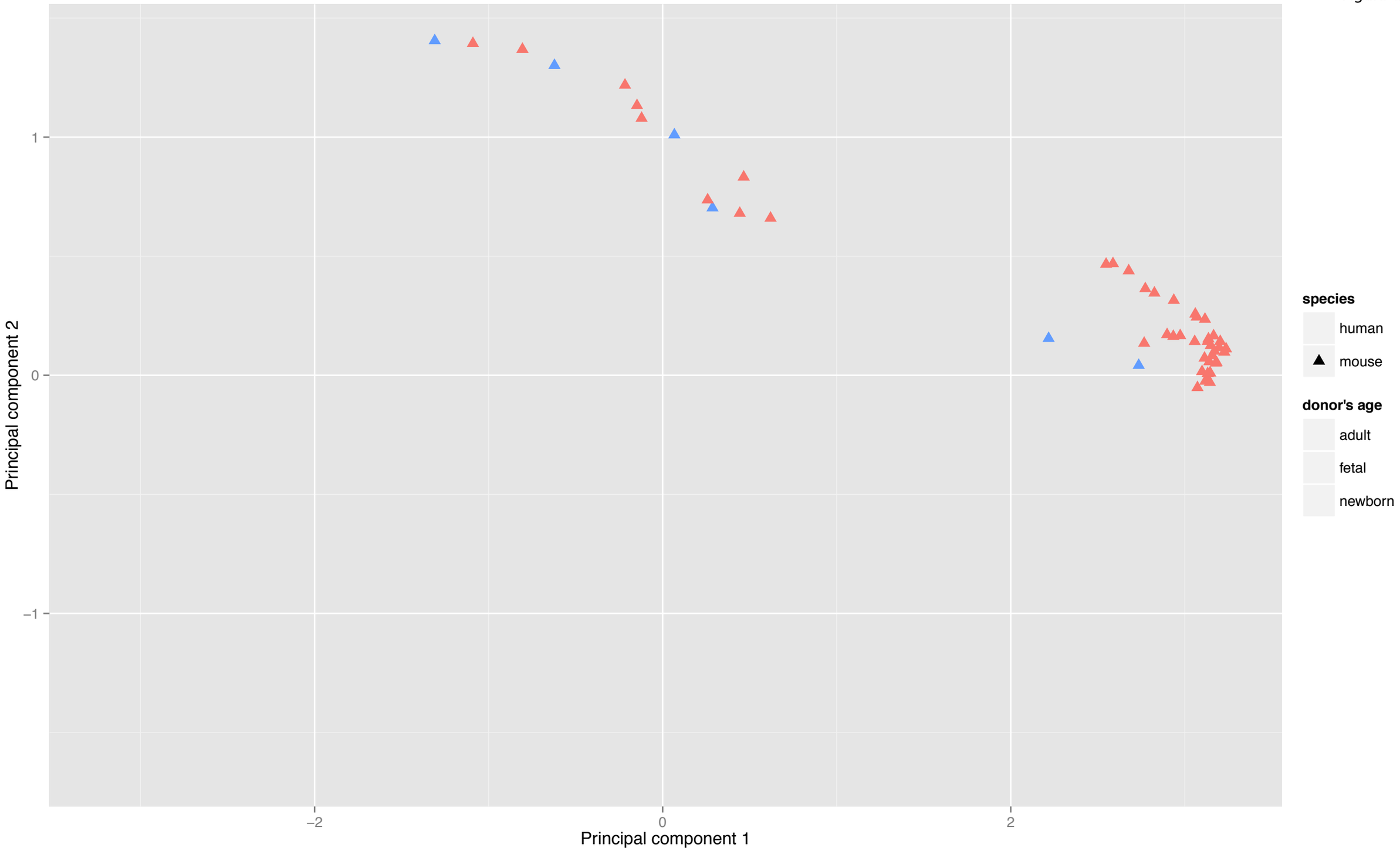
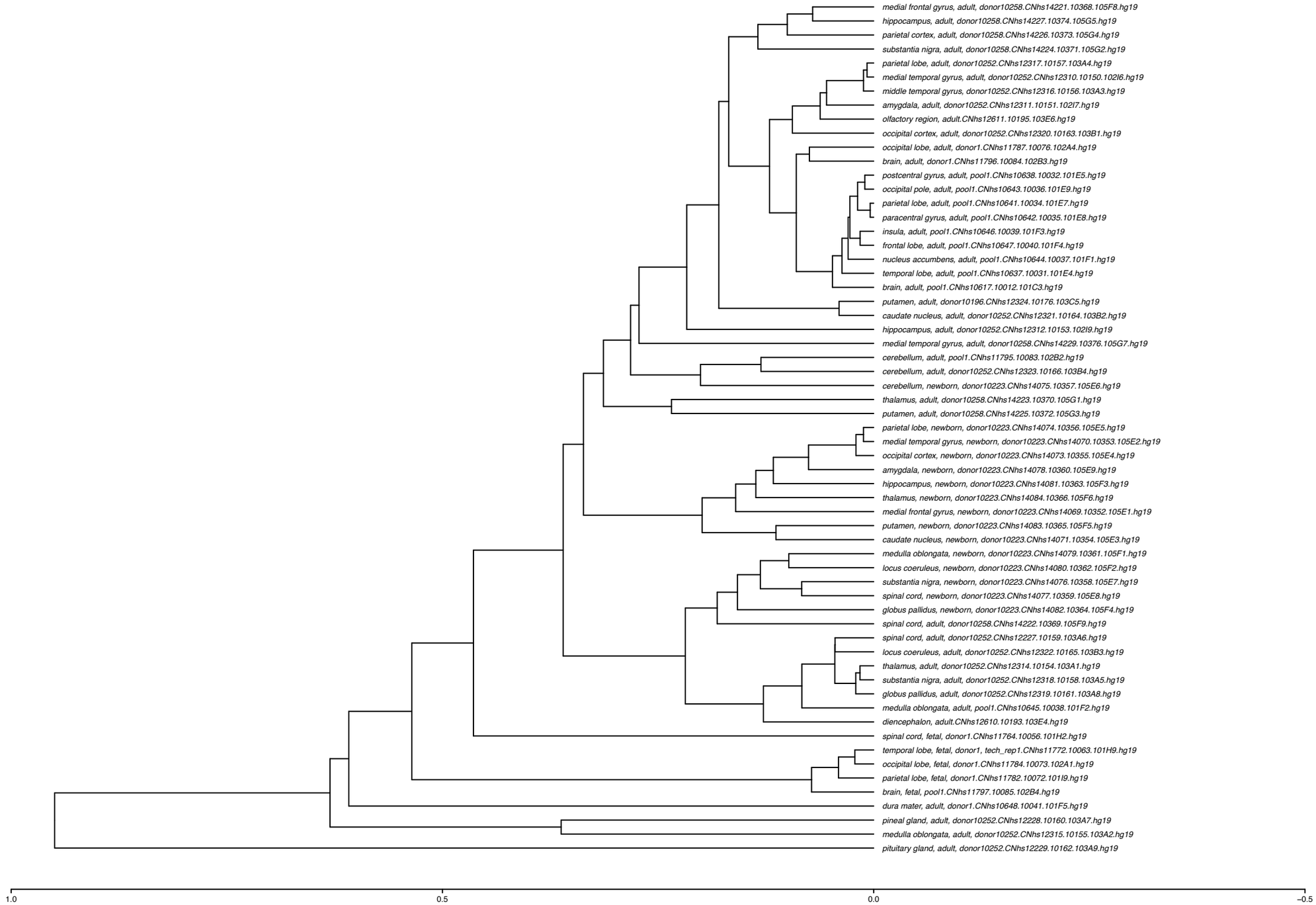

1
0
−1
Principal component 2
−2
0
2
Principal component 1
species
human
mouse
donor's age
adult
fetal
newborn

Figure S2


medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
medial temporal gyrus, adult, donor10252.CNhs12310.10150.102I6.hg19
middle temporal gyrus, donor10252.CNhs12316.10156.103A3.hg19
amygdala, adult, donor10252.CNhs12311.10151.102I7.hg19
olfactory region, adult.CNhs12611.10195.103E6.hg19
occipital cortex, adult, donor10252.CNhs12320.10163.103B1.hg19
occipital lobe, adult, donor1.CNhs11787.10076.102A4.hg19
brain, adult, donor1.CNhs11796.10084.102B3.hg19
postcentral gyrus, adult, pool1.CNhs10638.10032.101E5.hg19
occipital pole, adult, pool1.CNhs10643.10036.101E9.hg19
parietal lobe, adult, pool1.CNhs10641.10034.101E7.hg19
paracentral gyrus, adult, pool1.CNhs10642.10035.101E8.hg19
insula, adult, pool1.CNhs10646.10039.101F3.hg19
frontal lobe, adult, pool1.CNhs10647.10040.101F4.hg19
nucleus accumbens, adult, pool1.CNhs10644.10037.101F1.hg19
temporal lobe, adult, pool1.CNhs10637.10031.101E4.hg19
brain, adult, pool1.CNhs10617.10012.101C3.hg19
putamen, adult, donor10196.CNhs12324.10176.103C5.hg19
caudate nucleus, adult, donor10252.CNhs12321.10164.103B2.hg19
hippocampus, adult, donor10252.CNhs12312.10153.102I9.hg19
medial temporal gyrus, adult, donor10258.CNhs14229.10376.105G7.hg19
cerebellum, adult, pool1.CNhs11795.10083.102B2.hg19
cerebellum, adult, donor10252.CNhs12323.10166.103B4.hg19
cerebellum, newborn, donor10223.CNhs14075.10357.105E6.hg19
thalamus, adult, donor10258.CNhs14223.10370.105G1.hg19
putamen, adult, donor10258.CNhs14225.10372.105G3.hg19
parietal lobe, newborn, donor10223.CNhs14074.10356.105E5.hg19
medial temporal gyrus, newborn, donor10223.CNhs14070.10353.105E2.hg19
occipital cortex, newborn, donor10223.CNhs14073.10355.105E4.hg19
amygdala, newborn, donor10223.CNhs14078.10360.105E9.hg19
hippocampus, newborn, donor10223.CNhs14081.10363.105F3.hg19
thalamus, newborn, donor10223.CNhs14084.10366.105F6.hg19
medial frontal gyrus, newborn, donor10223.CNhs14069.10352.105E1.hg19
putamen, newborn, donor10223.CNhs14083.10365.105F5.hg19
caudate nucleus, newborn, donor10223.CNhs14071.10354.105E3.hg19
medulla oblongata, newborn, donor10223.CNhs14079.10361.105F1.hg19
locus coeruleus, newborn, donor10223.CNhs14080.10362.105F2.hg19
substantia nigra, newborn, donor10223.CNhs14076.10358.105E7.hg19
spinal cord, newborn, donor10223.CNhs14077.10359.105E8.hg19
globus pallidus, newborn, donor10223.CNhs14082.10364.105F4.hg19
spinal cord, adult, donor10258.CNhs14222.10369.105F9.hg19
spinal cord, adult, donor10252.CNhs12227.10159.103A6.hg19
locus coeruleus, adult, donor10252.CNhs12322.10165.103B3.hg19
thalamus, adult, donor10252.CNhs12314.10154.103A1.hg19
substantia nigra, adult, donor10252.CNhs12318.10158.103A5.hg19
globus pallidus, adult, donor10252.CNhs12319.10161.103A8.hg19
medulla oblongata, adult, pool1.CNhs10645.10038.101F2.hg19
diencephalon, adult.CNhs12610.10193.103E4.hg19
spinal cord, fetal, donor1.CNhs11764.10056.101H2.hg19
temporal lobe, fetal, donor1, tech_rep1.CNhs11772.10063.101H9.hg19
occipital lobe, fetal, donor1.CNhs11784.10073.102A1.hg19
parietal lobe, fetal, donor1.CNhs11782.10072.101I9.hg19
brain, fetal, pool1.CNhs11797.10085.102B4.hg19
dura mater, adult, donor1.CNhs10648.10041.101F5.hg19
pineal gland, adult, donor10252.CNhs12228.10160.103A7.hg19
medulla oblongata, adult, donor10252.CNhs12315.10155.103A2.hg19
pituitary gland, adult, donor10252.CNhs12229.10162.103A9.hg19
1.0
0.5
0.0
−0.5