Version: Monday, March 25, 2013

Title page

Oxana Sachenkova (1, 2), Alistair R. R. Forrest (3), Carsten Daub (3), Lukasz Huminiecki (1, 2, 4, 5) and the FANTOM consortium

1. Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
2. Science for Life Laboratory, SciLifeLab, Sweden
3. RIKEN Omics Centre, Yokohama, Japan
4. Department of Cell and Molecular Biology, Karolinska Institutet, Sweden
5. BILS bioinformatics infrastructure for life sciences

Corresponding author: Lukasz.Huminiecki@scilifelab.se

Keywords: transcription start sites, cap analysis of gene expression, evolution, expression pattern, gene family, gene duplication, cancer

Evolution of expression patterns in human gene families illustrated by the *FANTOM5* encyclopedia of transcription start sites

Abstract:

Background

Human gene families emerged through consecutive rounds of gene duplication. Here we apply the cutting-edge *FANTOM5* single-nucleotide resolution atlas of transcription start sites from 1348 human and mouse libraries, to elucidate expression pattern evolution in animal gene families, with stress on comparison between human and mouse, and normal versus cancer cells.

Results

Broad over-view of *FANTOM5* was obtained with intra-species and inter-species hierarchical clustering of human and mouse samples. In the follow-up, we dated gene duplications by phylogenetic timing, and investigated the rate of expression pattern divergence between duplicates, as well as the tissue-specificity of their expression. Finally, we defined the concept of phylo-expression signatures as strong associations between duplications of certain ages and expression samples in the *FANTOM5* atlas. We show how phylo-expression signatures can be used to generate novel hypotheses on the nature of animal evolution, and discuss central nervous system and reproductive tract as two focused examples.

Conclusions

A striking trend for young genes to be narrowly expressed was revealed. Several lines of evidence suggested that emergence of placental mammals was a unique period in the evolution of animal gene families and duplicates dating to that period have broader and more conserved expression patterns, with genes involved in chromatin assembly and epigenetic control driving the trend. A major strength of the *FANTOM5* atlas is that it profiles normal tissues, primary cells, and cancer cell lines, and as expected, clustering of expression profiles showed a major divide between leukemias and solid tumors. Where the evolutionary link became apparent was that in cancer cell lines, unlike in tissues and primary cells, recent paralogs lacked the peak of highly correlated pairs. This novel finding suggests that global devolution and loss-of-evolutionary constraints on expression patterns accompany malignant transformation, and provides additional evidence in the debate on use of cancer cell lines as research models.

Abbreviations

2ROs 2R-ohnologs (paralogs derived from the 2R-WGD event)

2R-WGD two rounds of whole genome duplication

CAGE cap analysis of gene expression

DNASEI Deoxyribonuclease I

ENCODE The Encyclopedia of DNA Elements

F5 *FANTOM5*

FANTOM Functional Annotation of the Mammalian Genome (http://fantom.gsc.riken.jp/)

FANTOM5 The fifth phase of the FANTOM project seeks to generate transcriptional regulatory models to define all human cell types, using Helicos Single Molecule Sequencing and CAGE

GO gene ontology

Hs-Mm human-mouse

ISHC inter-species hierarchical clustering (human-mouse tissue clusters matched by linking ortholog gene expression)

JIB *Jaccard index*

NM-clusters human-mouse tissue clusters derived through name matching

NFKB nuclear factor kappa B

PC *Pearson correlation*

PFAM protein domains from the PFAM database

RAC 2 rho family, small GTP binding protein RAC 2

RAC 3 rho family, small GTP binding protein RAC 3

RAC subfamily of Rho GTPases (TreeFam family id TF101109)

RHOG Ras homology Growth-related

TreeFam database of animal gene families

TF6 TreeFam release 6

TF8 TreeFam release 8

TSS transcription start site

Tfbs transcription factor binding sites

TfbsClusteredV2 multi cell-line, clustered ENCODE Tfbs data (ENCODE_UCSC_Tfbs_V2); this is the subset of ENCODE Tfbs data displayed in the UCSC genome browser

TF transcription factor

WGD whole genome duplication

ZENBU *FANTOM5*'s own genome browser with single-base resolution CAGE datamining

ZBTB7A zinc finger and BTB domain containing 7A

Introduction

Gene duplication is of primary importance to understanding animal gene family evolution, since horizontal gene transfer does not play an important role. Animal gene families emerge through consecutive rounds of gene duplication, and functional divergence of gene duplicates continues to be a topic of interest (Huminiecki 2005; Huminiecki, Goldovsky et al. 2009; Huminiecki and Heldin 2010). Since animals are primarily characterized by multicellularity and the existence of distinct tissue- and cell-types, exploration of expression pattern evolution following gene duplication is of fundamental interest for understanding human evolution. To examine expression patterns across a broad collection of tissues, primary cells, and cancer cell lines, we used the single molecule Cap Analysis of Gene Expression - CAGE (Kanamori-Katayama, Itoh et al. 2011) data generated for the *FANTOM5* promoter level expression atlas (Forrest et al. supporting manuscript 1).

The *FANTOM5* promoter level expression atlas is arguably the most complete functional genomics dataset generated to date (Forrest et al. supporting manuscript 1). The *FANTOM5* data include 952 human and 396 mouse tissues, primary cells and cancer cell-lines. The major strengths of the *FANTOM5* dataset are comprehensiveness and technological uniformity, unmatched by any other expression datasource available today. Moreover, CAGE explores the entire genome space, unbiased by any prior assumptions or gene models, and allows for systematic comparison between normal tissue samples, primary cells in culture, and cancer cell-lines. **Table** 1 briefly summarizes the first release of the *FANTOM5* resource.

FANTOM5 should settle many heated debates on expression pattern evolution in animal gene families and following gene duplication. In comparison to ESTs and microarrays, *FANTOM5* data are less susceptible to cross-hybridization between paralogs, as CAGE tags target fast diverging promoter regions, instead of conserved protein coding regions (Masatoshi Nei 2000). CAGE also enables investigation of the entire genome in a unbiased fashion, not being limited to a set of pre-selected genes chosen by the chip maker. Furthermore, expression datasets available to date, such as the Gene Expression Atlas, were somewhat limited in scope to a narrow set of somatic tissues (Su, Cooke et al. 2002) while meta-analysis of expression datasets was associated with enormous problems due to technological differences between platforms and analysis pipelines (Huminiecki, Lloyd et al. 2003). Due to these limitations, many controversies with regards to animal expression pattern evolution remained unresolved.

In summary, herein, we combine new extensive, technologically uniform, and reproducible single-molecule sequencing-based *FANTOM5* encyclopedia defining both human and mouse transcription start sites and expression patterns, with data on animal phylogenomics to investigate spatial expression pattern evolution following gene duplication and in the context of gene family evolution.

This work is part of the *FANTOM5* project. Data downloads, genomic tools and co-published manuscripts are summarized here http://fantom.gsc.riken.jp/5/.

Results

Initial over-view of expression patterns in the *FANTOM5* atlas.

Broad over-view of *FANTOM5* was obtained with hierarchical clustering of human and mouse tissue samples. *FANTOM5* samples clustered primarily depending on cell and tissue type of origin, not donor or developmental stage. Clustering was performed for three different subgroups of samples in *FANTOM5*, tissues, primary cell lines, and cancer cell lines. Please, see **Figure 1** for heatmap of human tissues and **Figure 2** for heatmap of cancer cell lines. Other sample subgroups, namely human primary cells and mouse tissues are illustrated in supplementary figures (**Figure S1**a and **Figure S1**b, respectively). Multiple donors were available for a high proportion of samples, but neither hierarchical clustering nor principal component analysis (data not shown) grouped samples by donors. Clearly, tissue-of-origin differences are more important than individual variability between donors. Interestingly, for both human and mouse, samples derived from brain sub-locations and anatomical structures tended to cluster together (**Figure 1**, **Figure S1**b and **Figure S1**c).

First evolutionary comparison

Assignment of ortholog tissues between human and mouse.

In analogy to ortholog genes, ortholog tissues can be defined as homolog tissues derived from a common ancestral tissue trough the process of speciation. Therefore, assignment of ortholog tissues has been our first evolutionary comparison. We first assigned human and mouse ortholog tissues by simply comparing sample names (**Table 2**). This is a simple name matching procedure, resulting in name matching or NM-clusters (for example, human liver matches mouse liver). Secondly, we developed ortholog-based inter-species hierarchical clustering (shortened as the ISHC), where human and mouse samples are combined into one common expression matrix which is then clustered (**Figure 3**).

Could name clusters be recovered by the ISHC? **Table 2** summarizes the comparison between clusters identified through name matching, with those inferred using the ISHC. The full dataset is provided in supplementary **Table S1**. Eight NM-clusters were recovered by the ISHC, but the majority were not. The NM-clusters which were recovered included: skin, liver, tongue, heart, pancreas, pituitary gland, thymus and “total RNA control”. However, twice as many NM-clusters (namely 19) were not recovered. This difficulty seemed robust to alterations of the ISHC procedure. For example, we have experimented with other expression distance measures, as well as ISHC-variant based on whole family-averaging rather than ortholog genes, but higher rate of recovery could not be achieved (data not shown).

Figure 3b displays stack histogram with distance distributions, obtained during the course of the ISHC procedure, for the two intra-species comparisons (human, and mouse), and the inter-species comparison (human-mouse). Mean inter-species distances (*Hs*-*Mm*) were somewhat higher than intra-species distances (*Hs* and *Mm*): 0.78, 0.68 and 0.72, respectively. This suggests that expression pattern evolution rate is rather rapid, and even on moderate evolutionary distances, such as the human-mouse comparison, inter-species expression pattern differences dominate over inter-organ and inter-tissue differences.

Second evolutionary comparison

Rates of expression pattern evolution vary widely between tissues.

The overall rate of expression pattern evolution and whether it is subject to purifying selection has been hotly debated (Huminiecki and Wolfe 2004; Khaitovich, Weiss et al.

2004; Jordan, Marino-Ramirez et al. 2005). *FANTOM5* tissues could be subdivided into three groups of widely differing expression pattern evolution rates, as calculated by Euclidean distance between paralog pairs (**Figure 4**). The three groups were: (**a**) dynamic, (**b**) intermediate and (**c**) static. Dynamic tissues included thymus, adipose, liver, pancreas and blood. Brain samples were split between intermediate and static, suggesting that different brain regions and anatomical structures evolve at different rates. This high variability in expression evolutionary rates depending on tissue, suggested that conservation of expression profiles is a tissue-specific phenomenon, likely to depend on transcription factors used by a cell and cell's physiological function. In contrast, protein sequence evolutionary rates are not known to vary widely depending on tissue of expression, although a trend for tissue-specific genes to evolve faster was reported using several expression datasets, including the Gene Expression Atlas (Huminiecki and Wolfe 2004).

Phylogenetic timing of gene duplications using TreeFam8 database.

In our previous work, we used TreeFam6 database (Li, Coghlan et al. 2006) to phylogenetically time gene duplication events (Huminiecki and Heldin 2010). The current study used an updated and expanded release of the database, TreeFam8, which showed the same overall distribution of duplication events, linking the emergence of vertebrates and bilaterian animals with the two most abundant waves of gene duplications in the history of animal kingdom.

Recently evolved duplicates tend to be tissue-specific with the exception of histones.

We observed two broad evolutionary trends related to animal expression patterns, a novel trend for recently evolved genes to be more tissue-specific in their expression domain shown in **Figure 5**, and a previously described trend for gradual paralog expression pattern divergence illustrated in **Figure 6** (Huminiecki and Wolfe 2004; Huminiecki and Heldin 2010).

When the dynamics of these two trends are compared, paralog expression divergence appears faster (**Figure 6**), reaching plateau at the timescales similar to divergence between mammals and reptiles (taxon Amniota, around 300 million years). In contrast, the trend for older genes becoming increasingly housekeeping, did not reach a plateau until the base of the animal tree of life (taxa *Eumetazoa-Metazoa*, **Figure 5**).

However, placental mammals (taxon *Eutheria*) were an outlier to both of the above trends, with an enrichment in housekeeping genes and co-expressed paralogs. Several lines of evidence suggested that genes involved in epigenetic regulation, in particular histone families, lie at the root of unusual characteristics of duplications associated with placental mammals. Here, three large families of histones, *H2B*, *H2A* and *H3* (TreeFam accessions TF300212, TF300137 and TF314241, respectively) contribute heavily to the set of highly correlated paralogs (*Pearson correlation* higher than 0.9), with thymus and testis as the main tissues where paralog histones were highly co-expressed (data not shown). It seems logical that fast proliferating tissues, such as thymus and testis, express abundantly all *H2B*, *H2A* and *H3* variants. At the same time, these histones are expressed at low to average level in almost all tissues, resulting in the signature of broad expression. The fact that mammalian duplications in histone families gave rise to housekeeping genes, rather than tissue-specific genes, which seems a rule for other functional classes of genes, suggests regulatory rather than structural novelty. Taken together these data imply that diversification of mammals was accompanied by the expansion of the part set of the epigenetic regulatory toolkit.

We then applied gene ontology and domain enrichment for functional characterisation of genes lying at the extremes of distribution of these two broad trends. Supplementary tables **Table S3a** and **Table S3b** summarize the results for fast expression divergence rate, while **Table S3b** relates to high breadth of expression. In both cases, the cut-off at the top *0.75* quantile of the distribution was used, with several observations emerging as being of particular interest.

(a) For **placental mammals (**taxon *Eutheria***)**, highly co-expressed genes were associated with chromatin assembly or disassembly, nucleosome assembly, and DNA packaging (GO: 0006333, p = 4.65e-06; GO:0006334, p = 4.65e-06; GO:0006323, p = 8.38e-06, respectively);

(b) Association of **histone domain** (PF00125; p = 1.13e-10) with **human-specific** highly housekeeping genes;

(c) Associations of extracellular region, and extracellular space (GO:0005576; p = 1.8e-07, GO:0005615; p = 5.79e-08, respectively), with broadly expressed gene duplicates which emerged during **diversification of primates** (taxa *Homo/Pan/Gorilla* and *Catarrhini*);

(d) Under-representation of terms: intracellular, cell, cytoplasm, organelle (GO:0005622, p = 4.21e-07; GO:0005623, p = 2.54e-06; GO:0005737, p = 7.84e-06; GO:0043226, p = 1.05e-05) among broadly expressed duplicates mapping to taxon *Catarrhini*.

Differences between normal tissues and cancer cell lines, loss of evolutionary constraints on gene expression accompanies malignant transformation.

When paralog analysis was extended into cancer cell lines, no signature for recently evolved duplicates to be co-expressed could be seen (**Figure 7**). This suggested important global differences in regulation of gene expression between normal tissues and primary cells, versus malignantly transformed cancer cell lines. As co-expression signature could be seen in primary cells, cancerous transformation not cell culture conditions lie at the root of the effect.

Phylo-expression signatures

Phylo-expression signatures were defined as a strong associations between individual *FANTOM5* samples, and gene duplicates derived from a given taxon. **Table S4** lists top three phylo-expression signatures for each bilaterian taxon. For example, human specific duplications tended to be expressed in thymus, liver and adipose tissue, suggesting immune and metabolic innovation. Interestingly, parotid gland, one of salivary glands, is associated with strong expression of gene duplicates dating to diversification of primates (taxa *Homo/Pan/Gorilla* and *Catarrhini*). Finally, genes expressed in thymus, both adult and fetal, were included in phylo-expression signatures of many taxa, suggesting constant immune system innovation was the persistent leading theme throughout animal evolution.

We then focused on the organ systems of unique interest to bilateral animal evolution, reproductive system and central nervous system (**Figure 8** and **Figures 9**, respectively). **Figures 8** and **9** can be analyzed from the point of view of evolutionary history of a given tissue, for example placenta is associated with gene duplications during diversification of primates and emergence of placental mammals, while the vaginal sample has strongest association with human-specific duplications. Another way of looking at **Figure 8** is to ask questions about expression signature of a given evolutionary period. For example, emergence of placental mammals was associated with genes highly expressed in uterus, placenta, and testis.

Illustrating FANTOM5 through specific family example

We illustrate the power of *FANTOM5* for elucidating evolutionary histories of gene families by focusing on the Rac family example. The RAC subfamily of Rho GTPases (TreeFam accession TF101109) features an interesting phenomenon of mutually exclusive expression pattern of three paralogs, namely RAC 2, RAC 3, and RHOG (**Figure 10**). RAC 3 was highly expressed in five fetal tissues from which RHOG, and RAC 2 were excluded. The five tissues in question were of fetal origin, parietal lobe, temporal lobe, duodenum, occipital lobe, and brain pool. RHOG, on the other hand, was highly expressed in adult corpus callosum, where the other two genes were not detectable (**Figure 10**). Finally, a cluster of tissues associated with the immune and circulatory systems (namely adult thymus, blood, tonsil, appendix, vein, spleen, and lymph node; as well as fetal thymus and spleen), expressed RHOG and RAC 2 but not RAC 3. These dramatic expression pattern shifts are ellucidated with transcription factor binding site data in discussion.

Discussion:

Clustering of human and mouse samples.

Hierarchical clustering of samples according to their protein-coding gene expression profiles provided the first overview of the evolutionary expression space defined by *FANTOM5*. **Figure 1** demonstrates distinct clusters formed by brain samples, with one brain subcluster formed by retina, eye, optic nerve, brain glands and brain stem, and another formed by cortex, cerebellum, midbrain, and the limbic system. **Figure S1**b shows that mouse brain samples also clustered, with subclusters formed by visual cortex and cerebellum. **Figure 2**, **Figure S1**a, and **Figure S1**b show heatmaps for human cancer cell lines, human primary cells, and mouse tissues respectively. Finally, human cancer cell lines showed a major divide between leukemias and solid tumors, suggesting that these two major clinical subclasses of human malignancies are readily distinguished by their CAGE expression profiles (**Figure 2**).

Assignment of human-mouse ortholog tissues.

Previous studies comparing evolutionary profiles between species accepted a rather simple method for matching tissues, automatically assuming that samples with matching names in different species are ortholog tissues. Herein, we show that this may be an over-simplification of a complex problem. When we clustered human and mouse tissues together, most samples grouped by species not tissue type (**Figure 3**a). For example, human and mouse brain samples formed two adjacent but distinct clusters. However, some tissue-type clusters could be seen, for example human and mouse testis, heart, liver, kidney, skin, pancreas, and intestine grouped together.

How to interpret these discrepancies? Firstly, as demonstrated by expression divergence between paralogs (**Figure 6**), expression pattern evolution is rapid. Lineage-specific expression pattern shifts and tissue-specific evolutionary novelties put into question the very assumption of the existence of ortholog tissues. For example, conceptually it may be wrong to assume that human brain sublocations correspond directly to those in mouse brain, as behavioral, reproductive, and ecological differences between these two species are profound. Secondly, tissues are complex mixtures of cell types, and proportions of different cell types building a tissue may differ between species. The latter hypothesis may be investigated in *FANTOM5* data using tissue-specific molecules as markers, for example endothelial-specific molecules to estimate the degree of vascularization in different organs and species (Huminiecki and Bicknell 2000; Huminiecki, Gorn et al. 2002).

Recently duplicated genes are tissue-specific in their expression domain, with the exception of histone families.

The emergence of placental mammals (taxon *Eutheria*) was associated with a burst of duplications or unusual characteristics in terms their expression patterns, relatively wide and co-expressed (**Figure 5** and **Figure 6**) with enrichment in gene ontology terms suggesting functions in chromatin assembly and epigenetic control. Three families of histones, *H2A*, *H2B*, and *H3*, contributed strongest to these trends, as these genes are both widely expressed in most tissues, and very significantly up-regulated in proliferating tissues such as thymus and testis. Expansion of histone families hints towards chromatin structure regulation and epigenetic regulatory mechanism. Combined with other published studies, these results raise an intriguing possibility that major animal evolutionary transitions were accompanied by regulatory innovation involving different functional types of molecules, signal transduction pathways at the emergence of vertebrates (Huminiecki and Heldin 2010) and histones at the emergence of placental mammals as suggested herein.

Comparison of normal versus cancer cells.

When three broad human *FANTOM5* sample subtypes were compared (tissues, primary cells, and cancer cell-lines), we found that the peak of highly co-expressed paralogs (top quartile of *Pearson correlation* distribution) was missing in cancer cell-lines (**Figure 7**). In other words, while recent closely related paralogs tended to be expressed in the same tissues or primary cells, they were not co-expressed across cancer cell-lines. The effect cannot be attributed to cell culture conditions alone, as the co-expression signature was seen in primary cell samples. We propose a novel term, carcinogenic devolution of expression patterns, to suggest that normal evolutionary constraints on expression patterns do not exist in cancer. Carcinogenic expression pattern devolution is indicative of global distortion of cancer expression space, perhaps through accumulation of gross genomic abnormalities, such as those just reported for the HeLa genome.

To investigate differences between normal tissues and cancer samples further, we looked at gene families with differential average expression in tissues, versus primary cells and cancer cell-lines (**Table S2**). It makes sense that most of the top cancer cell-line over-expressed families were associated with cell cycle and proliferation, for example cyclins A and B, kinesins, cyclin-dependent kinases, aurora kinase, F-box only protein 5 (also known as early mitotic inhibitor 1), and cell division cycle associated 7. These differences could be interpreted as simple consequence of malignant transformation and uncontrolled proliferation. In contrast, gene families over-expressed in tissues were mostly involved in cell adhesion, for example PRKA3 and 4 described in the context of sperm-oocyte adhesion, patched which functions as a negative regulator of Hedgehog signaling pathway, and a tumor suppressor involved in cell adhesion and attachment Adenomatous Polyposis Coli. Taken together, these results illustrate the power of *FANTOM5* to differentiate between alterations introduced by cell culture conditions alone, from those which are the true consequence of malignant transformation.

Specific family example, the RAC family and mutually exclusive expression of paralogs.

RAC's family varied expression patterns were visualized in **Figure 10**. Can dramatic expression domain shifts and mutually exclusive expression pattern seen for RAC 2, RAC 3 and RHOG be correlated with transcription factor binding sites for these genes? Promoter regions of these three genes were examined with the F5-ZENBU and UCSC genome browsers (**Figure S2**), and ENCODE transcription factor binding sites (**Table 3**). Narrowly expressed but with a broad transcription start site, RAC 3 features only one strong transcription factor binding site for ZBTB7A. In contrast, broadly expressed RAC 2 and RHOG have many transcription factor binding sites with strong NFKB signature and top sites of expression in immune cells. Taken together, these results suggested a scenario where ancestral NFKB binding site was replaced with ZBTB7A, resulting in a shift from a broad expression pattern associated with immune cells, to a narrow expression domain associated with fetal brain. Indeed, in literature NFKB was reported to be widely expressed in animal cells and play a role in immune and stress responses (Karin 2006), while ZBTB7A was shown to be an oncogenic transcription factor implicated in glioma (Rovin and Winn 2005).

Intriguingly, dramatic expression shifts in the RAC family were not reflected directly in its phylogenetic history. The RAC tree (TreeFam accession TF101109) showed that RAC 2 and RAC 3 were 2R-ohnologs deriving from 2R whole genome duplication (Huminiecki and Heldin 2010). In contrast, divergence between RAC 2/3 ancestor and RHOG predated the origin of animals. Thus, RAC 2 and RAC 3 are more closely related in evolutionary terms but starkly different in their expression profiles.

Phylo-expression signatures and novel evolutionary hypotheses.

Examination of genes behind top phylo-expression associations provide rich material for formulation of hypotheses on animal evolution (**Table S4** and **S5**). In the first example, salivary cystatins (Baron, DeCarlo et al. 1999) and proline-rich salivary proteins (Amado, Lobo et al. 2010) are part of the parotid gland/*Catarrhini* phylogenetic signature, suggesting adaptation to novel food sources in this division of higher primates. In the second example, histone proteins are involved in multiple top mammalian phylo-expression signatures. Histones *H3* and *H4* are linked with human adipose, parietal_lobe, and putamen, as well as testis/*Eutheria* phylogenetic signatures, linking this analysis with trends discussed previously for histones *H2A* and *H2B* in the context of diversification of placental mammals. In the third example, semenogelin 1 and 2, predominant proteins in semen (Lilja, Abrahamsson et al. 1989), derive from a gene duplication associated with the seminal vesicle/*Catarrhini* phylo-expression signature, and have been previously implicated in evolution of sociosexual behavior in hominoid primates (Jensen-Seaman and Li 2003). Finally, in most fish esophagus like the pharynx is extremely short, but was strongly extended in tetrapods. Accordingly, keratins, making up basal layers for epithelia in esophagus, are the top proteins of the phylo-expression signature for esophagus and Tetrapoda.

Conclusions.

To recapitulate, using *FANTOM5* and phylogenomics, we demonstrate a trend for young genes to be specific in their expression domain, but mammalian histone duplicates defy the trend being both widely expressed in normal tissues and highly up-regulated and co-expressed in proliferating tissues, such as thymus and testis. Comparison of normal versus cancer samples suggested that global devolution of expression patterns accompanied malignant transformation. Finally, we illustrate *FANTOM5* explanatory potential on the individual family level using RAC family example, and introduce the concept of phylo-expression signatures with potential for generation of novel hypotheses on animal evolution.

Data access:

Access to *FANTOM5* is provided at the *FANTOM5* public website, including the UCSC genome browser mirror, and *FANTOM5*'s own CAGE-focused ZENBU genome browser.

Methods:

TreeFam8 database

The version eight of the TreeFam database (released on 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.

FANTOM5 **dataset and gene expression tables**

FANTOM5 is arguably the most complete functional genomics dataset generated to date, including 952 human and 396 mouse tissues, primary cells and cancer cell-lines.

To produce gene expression tables, RefSeq transcripts were linked with all CAGE tags +/- 500 bps from the RefSeq's TSS. Expression values were normalized to tags per million (TPM) values. TPM of 10 was accepted as threshold for a gene to be "on" in a given tissue.

Linking TreeFam8 with ***FANTOM5***

TF8 trees were linked with *FANTOM5* to produce a unified database of phylogenetics and gene expression information. In the first step, ENSEMBL gene ids used in TreeFam were linked to EntrezIDs and those were then linked to RefSeq ids.

Later analysis stages were performed in R/Bioconductor (v2.11) using among others BioC packages: Biodist, gplot, ggplot, rtracklayer, TxDb.Hsapiens.UCSC.hg19.knownGene, and GOstats.

Expression distances and clustering

Several different expression distance measures were used, as there is not one single measure which works equally well for different applications. *Spearman distances* were used for tissues, primary cells, and cancer cell-line heatmaps (**Figure 1** and **2**, **Figure S1 a**,**b**), as this type of correlation is rank-based and less affected by strong outliers. Hierarchical clustering was used to generate **Figures 1**, **2** and **3**. Between-paralog *Euclidean distances* were used for **Figure 4**, where within-tissue expression divergence rates are measured. *Pearson correlation* was used for **Figure 6** and **7** as it works well for tissue-specific genes (Huminiecki and Wolfe 2004). ISHC expression distances were calculated using Bioconductor package bioDist Release (2.11) for all pairwise comparisons using the function cor.dist and the 1-ICORI formula.

Heatmaps were prepared with R package heatmap.2 which automatically generates color-coded histogram legends.

Preferential Expression Measure (PEM)

PEM is a ratio of maximal expression in any of the samples over average expression in all samples, and is high for tissue-specific genes and low for housekeeping genes.

Ortholog tissue assignment

a) Inter-species hierarchical clustering of samples (ISHC).

In inter-species hierarchical clustering of samples (ISHC) human and mouse samples are clustered according to expression profiles of orthologs. *Pearson correlation* was used as the expression distance measure.

b) Name clusters (NM-clusters) were identified using a custom Python script using standard text analysis approaches.

Expression in tissues and primary cells, versus cancer cell lines

Across family members, average expression values were calculated and compared for three major *FANTOM5* subclasses (supplementary **Table S4** is a summary table, while the full dataset is shown in **Table S5**). Data in **Table S5** can be uploaded to a relational database or statistical environment and queried in multiple additional ways, depending on the biological question of interest to the reader. For example, one can identify (**a**) all families preferentially expressed in tissues and primary cells versus cancer cell lines; (**b**) preferentially expressed in tissues versus primary cells and cancer cell lines; (**c**) narrowly expressed in tissues but not in cancer cell lines, etc.

Jaccard index **calculations**

Jaccard index, that is the ratio of intersection over the union, is used here as a measure of similarity between promoters of the RAC family in terms of their transcription factor binding profiles. In our hands, *Jaccard index* works well as a distance metric for measuring the rate of promoter divergence, and correlates with divergence in expression profiles between pairs of duplicates (manuscript in preparation).

ENCODE data

Transcription factor binding sites reported here derive from ENCODE data. Multi cell-line clustered ENCODE dataset published by the ENCODE consortium in 2012 (ENCODE) was used to analyze promoter regions of genes of interest within the 500 bps window (-/+ 250 bps from the TSS). The ENCODE transcription factor binding site dataset (weblink ENCODE_UCSC_Tfbs_V2) is contained in the file TfbsClusteredV2 and includes data for 148 transcription factors, including 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. Strong transcription factor binding sites had a score higher than 750, overall score varies between 0-1000.

Promoter regions of several genes were also inspected manually using the UCSC genome browser on the GRCh37/hg19 assembly, which includes tracks with the same subset of ENCODE data .

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).

Acknowledgments

Lukasz Huminiecki is supported by Bioinformatics Infrastructure for Life Sciences (BILS). BILS is a distributed national research infrastructure supported by the Swedish Research Council (Vetenskapsrådet) providing bioinformatics support to life science researchers in Sweden. BILS is also the Swedish contact point to the European infrastructure for biological information ELIXIR. We would like to thank Prof. Bengt Persson and staff at BILS for creating friendly and intellectually stimulating environmnet for sustained development of bioinformatics in Sweden.

FANTOM5 was made possible by a Research Grant for RIKEN Omics Science Center from MEXT to Yoshihide Hayashizaki and a Grant of the Innovative Cell Biology by Innovative Technology (Cell Innovation Program) from the MEXT, Japan to YH. We would like to thank all members of the *FANTOM5* consortium for contributing to generation of samples and analysis of the data-set and thank GeNAS for data production.

OS is funded by VR project research grant 2011-5893.

Author contributions:

Lukasz Huminiecki designed the study, performed the analyses, and wrote the manuscript.

OS provided technical assistance for some of the analyses.

A.R.R.F. and C.O.D were involved in the *FANTOM5* concepts and management.

Disclosure declaration:

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Clustering of human tissues.

Central nervous system samples clustered separately from all other tissues, with two distinct sub-clusters: (**b**) brain stem - medulla oblongata, spinal cord, brain glands, eye; (**c**) midbrain and higher brain. Several additional clusters were clearly visible, and were annotated (**d**-**i**).

Figure 2. Clustering of human cancer cell-line samples.

Clustering of cancer cell-line samples showed a major divide between leukemias and solid tumors, annotated as clusters (**a**) and (**b**).

Figure 3. Inter-species hierarchical clustering of samples (ISHC).

Panel (**a**) shows the tree for inter-species hierarchical clustering of samples (ISHC), where human and mouse samples were clustered according to expression profiles of orthologs. *Pearson correlation* was used as the expression distance measure. Panel (**b**) shows a stacked histogram with distance distributions, with all-against-all sample comparisons, for *Hs* - intra-species human, *Mm* - intra-species mouse, and *Hs-Mm* inter-species human-mouse.

Figure 4. Expression pattern evolution rates varied widely between tissues.

Tissues could be subdivided into three groups of differing expression pattern evolution rates: (**a**) dynamic (for example, thymus, adipose, liver, pancreas and blood), (**b**) intermediate and (**c**) static. Brain samples were split between intermediate and static clusters.

Figure 5. Recently evolved genes were tissue-specific.

In the evolutionary lineage leading to humans, as organisms grew in complexity and additional tissues formed, new genes became more tissue-specific in their expression domain. Placental mammals (taxon *Eutheria*) were an outlier to the trend.

Figure 6. **Expression pattern divergence between paralogs.**

Paralogs diverge in expression profiles over evolutionary time, but placental mammals (taxon *Eutheria*) are an outlier to the trend.

Figure 7. Paralog expression divergence in human tissues versus cancer cell-lines.

Distribution of paralog expression distances (*Pearson correlation*) offered unexpected evidence for global transcriptional deregulation in human cancer cell-lines. Panel (**a**) human tissues, panel (**b**) human cancer cell-lines. No peak for paralogs with high *Pearson correlatio*n could be seen in human cancer cell-lines.

Contributions of pairs from different taxa were visualized using colors as described in the legend key histogram. Interestingly, paralogs from the top quartile of *Pearson correlation* distribution tended to be recent gene duplicates, the great majority not older than taxon *Eutheria*.

Figure 8. Evolutionary history of mammalian tissues, reproductive system.

Placenta is associated with gene duplications dating from diversification of primates and emergence of placental mammals. The vaginal sample has strongest association with human-specific duplications. Emergence of placental mammals was associated with genes highly expressed in uterus, placenta, and testis.

Dark color bands on the heatmap correspond to phylo-expression signatures, that is strong taxon/tissue associations. Each band corresponds to average TPM expression values for all duplicates associated with a given taxon. Key to the heatmap shows color-codes and associated histogram of frequencies.

Figure 9. Evolutionary history of mammalian tissues, brain samples.

Dark color bands on the heatmap correspond to phylo-expression signatures, that is strong taxon/tissue associations. Each band corresponds to average TPM expression values for all duplicates associated with a given taxon. Key to the heatmap shows color-codes and histogram of associated frequencies.

Emergence of placental mammals appears associated with a wave of duplications in parietal lobe, temporal lobe, fetal brain, putamen, and dura mater.

Figure 10. Evolution of expression patterns in the RAC family.

Panel (**a**) shows heatmap of expression patterns for the RAC family in human tissues, while panel (**b**) shows heatmap of the values for *Jaccard index* for ENCODE transcription factor binding sites in pairwise comparisons. **Table 4** shows actual values of the *Jaccard index*, which had bimodal distribution with peaks in two intervals: 0-0.2 and 0.2-0.5 (**Figure 10b** and **Table 4**).

RAC 2 and 3 are 2R-ohnologs. RHOJ and RHOQ were very divergent family members, with weak expression in human tissues. RAC 1 and cdc42 were also highly diverged, but similar in their expression pattern. Neither the expression distance nor the *Jaccard index*-based dendrogram directly reflected the phylogenetic history of the family.

Supplementary figures:

Figure S1a. Clustering of human primary cell samples.

Hierarchical clustering and *Spearman correlation* were used.

Figure S1b. Clustering of mouse tissues.

Hierarchical clustering and *Spearman correlation* were used.

Figure S2. Promoter regions of RAC 2, RAC 3 and RHOG were examined with the F5-ZENBU and UCSC genome browsers.

RAC3 is narrowly expressed but has a broad transcription start site driven by ZBTB7A. In contrast, broadly expressed RAC 2 and RHOG have many transcription factor binding sites, but narrow transcription start sites.

TABLES

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

Numbers of human and mouse samples in different subsets are listed (tissues, primary cells, and cancer cell lines).

Table 2. Two strategies for ortholog tissue assignment.

Comparison of two different approaches for ortholog tissue identification: “name matching” (NM) and inter-species hierarchical clustering (ISHC). Eight name clusters were also ISHC clusters, signified by “YES” in the last column. However, many name clusters were not recovered as inter-species hierarchical clusters (signified by “NO” in the last column), while two were split and difficult to classify (signified by “Not certain” in the last column).

Table 3. Transcription factor binding sites in the RAC family.

The table lists all transcription factors linked to the promoters of genes in the RAC family (“All Tfbs”), and a subset of the the dataset with the strongest signal (“Strong Tfbs”).

Table 4. *Jaccard index* for pairwise comparisons between RAC family members.

High *Jaccard index* indicates similar transcription factor binding profile between two gene promoters (*Jaccard index* between identical promoters is 1). Values higher than 0.2 are shown in bold and using larger font. Please, see **Figure 10**b for clustering and heatmap representation of the table. At least in this family, there is no direct correlation between *Jaccard index* and co-expression. For example, RAC 1 and CDC42 have high *Jaccard index* (0.34) and cluster together in expression profiles. However, RAC 2 and CDC42 have higher *Jaccard index* (0.43) and substantially different expression patterns.

Supplementary tables

Tables S2 and **S4** are included in this document. For other supplementary tables see attached files with corresponding file names (for example, OS_LH_FANTOM5_TableS1).

Table S1. The full *FANTOM5* ortholog tissue clustering dataset is shown here:
(a) NM-dataset; (b) ISHC clusters.

Table S2. Families with differential average expression in tissues versus primary cells and cancer cell lines.

Table S3a. GO term and PFAM domain enrichment, for all taxa, in genes with unusually fast expression divergence rate.

Table S3b. GO term and PFAM domain enrichment, for all taxa, in genes with unusually high breadth of expression.

Table S4. Strongest associations between timing of gene duplication and expression site in *FANTOM5*.

Table S5. Example genes behind strongest associations between timing of gene duplication and expression site in *FANTOM5*.

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

Numbers of human and mouse samples in different subsets are listed (tissues, primary cells, and cancer cell lines).

***FANTOM5* samples**	**human**	**mouse**
total	952	396
tissues	179	280
primary cells	513	116
cancer cell lines	260	-
brain tissues	60	51
reproductive tissues	14	21

Table 2.

Name matching (NM) cluster	No. of human samples in the NM cluster	No. of mouse samples in the name 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 3.

Symbol	Name, Promoter location	All Tfbs	Strong Tfbs
RHOQ	ras homolog family member Q *chr2: 46769617-46770116*	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	HA-E2F1, ZNF263
RHOG	ras homolog family member G *chr11: 3861964-3862463*	HA-E2F1, CCNT2, Pol2, HEY1, ELF1_(SC-631), Pol2-4H8, GABP, PU.1, Egr-1, HA-E2F1, **NFKB**	
RHOJ	ras homolog family member J *chr14: 63670852-63671351*	KAP1, c-Jun, c-Fos, JunD, GATA-2, CTCF, HDAC2_(SC-6296), Rad21, p300, ELF1_(SC-631), Pol2(b), SRF, Pol2	c-Jun, Pol2(b)
RAC 1	ras-related C3 botulinum toxin substrate 1 *chr7: 6413876-6414375*	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)	TBP, HA-E2F1, YY1, YY1_(C-20), E2F1, GABP
RAC 2	ras-related C3 botulinum toxin substrate 2 *chr22: 37640056-37640555*	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	**NFKB**, Pol2
RAC 3	ras-related C3 botulinum toxin substrate 3 *chr17: 79989282-79989781*	ETS1, Sin3Ak-20, **ZBTB7A**_(SC-34508), Egr-1, SRF	**ZBTB7A**_(SC-34508)
CDC 42	cell division cycle 42 *chr1: 22378870-22379369*	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	Nrf1, Pol2, ELF1_(SC-631), PU.1, HMGN3, eGFP-JunD, Pol2-4H8, TAF1, HEY1

	RHOQ	RHOG	RHOJ	RAC 1	RAC 2	RAC 3	CDC42
RHOQ	***1***						
RHOG	**0.24**	***1***					
RHOJ	0.07	0.1	***1***				
RAC 1	**0.28**	**0.35**	0.1	***1***			
RAC 2	**0.25**	**0.31**	0.09	**0.25**	***1***		
RAC 3	0.17	0.07	0.06	0.04	0.16	***1***	
CDC42	**0.23**	**0.21**	0.12	**0.34**	**0.43**	0.07	***1***

Table 4.

TableS2. 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.

expression high in human cancer cell lines, low in human tissues		**fold difference**	**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
expression high in tissues, low in primary cells and cancer cell lines		**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

expression high in primary cells, and cancer cell lines, low in human tissues		**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
expression high in tissues, low in cancer cell lines		**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

TableS4. *FANTOM5* phylo-expression signatures

Strongest associations between timing of gene duplication and expression site in *FANTOM5*. Average expression values for each association in TPM are given.

Thymus contributes to almost all top phylo-expression signatures.

Taxon	Tissue name	Expression
Homo sapiens	thymus adult liver fetal adipose	245 113 108
Homo/Pan/Gorilla	parotid gland adult salivary gland adult blood adult	334 324 320
Catarrhini	parotid gland adult liver fetal trachea adult	243 124 70
Eutheria	thymus adult liver fetal thymus fetal	566 225 208
Theria	blood adult liver fetal thymus adult	139 134 104
Amniota	thymus adult thymus fetal breast adult	116 50 45
Tetrapoda	thymus adult esophagus adult pancreas adult	185 149 139
Euteleostomi	thymus adult adipose thymus fetal	78 48 42
Chordata	pancreas adult skeletal muscle adult artery adult	114 74 58
Deuterostomia	pancreas adult placenta adult dura mater adult	50 31 28
Bilateria	thymus adult adipose pancreas adult	108 73 56

Table S5. Individual genes behind several *FANTOM5* phylo-expression signatures. Phylo-expression signatures are strong associations between timing of gene duplication and expression sites in *FANTOM5*. Here individual genes involved in several phylo-expression signatures are listed.

esophagus, adult, ***Tetrapoda***		
gene	average expression	name
NM_002274	13082.2	keratin 13
NM_153490	13082.2	keratin 13
NM_002272	11030.4	keratin 4
NM_000424	6259.86	keratin 5
NM_005554	4055.68	keratin 6A
NM_000700	3702.54	annexin A1
NM_001199893	2896.67	actin, gamma 2, smooth muscle, enteric
NM_001615	2896.67	actin, gamma 2, smooth muscle, enteric
NM_001613	1913.05	actin, alpha 2, smooth muscle, aorta
NM_002275	1681.03	keratin 15
adipose, ***Homo sapiens***		
NM_003542	3140.93	histone cluster 1, H4c
NM_003545	2515.18	histone cluster 1, H4e
NM_003537	1922.87	histone cluster 1, H3b
NM_003539	1719.11	histone cluster 1, H4d
NM_003544	1652.23	histone cluster 1, H4b
NM_003533	1608.02	histone cluster 1, H3i
NM_003546	1493.81	histone cluster 1, H4l
NM_003540	1022.51	histone cluster 1, H4f
NM_021018	704.817	histone cluster 1, H3f
NM_003532	704.251	histone cluster 1, H3e
parotid gland, adult, ***Homo/Pan/Gorilla***		

NM_002723	136988	proline-rich protein BstNI subfamily 4
NM_000200	126071	histatin 3
NM_006249	47975.3	proline-rich protein BstNI subfamily 3
NM_006685	38292.5	submaxillary gland androgen regulated protein 3B
NM_002159	25146.6	histatin 1
NM_002568	1033.51	poly(A) binding protein, cytoplasmic 1
NM_003299	206.46	heat shock protein 90kDa beta (Grp94), member 1
NM_012423	149.031	ribosomal protein L13a
NM_001014	134.32	ribosomal protein S10
NM_001203245	134.32	ribosomal protein S10
salivary gland, adult, ***Homo/Pan/Gorilla***		
NM_006685	295980	submaxillary gland androgen regulated protein 3B
NM_000200	25866.1	histatin 3
NM_002159	24816.8	histatin 1
NM_001322	16457	cystatin SA
NM_002568	693.334	poly(A) binding protein, cytoplasmic 1
NM_001898	575.298	cystatin SN
NM_012390	229.766	submaxillary gland androgen regulated protein 3A
NM_001203245	108.578	ribosomal protein S10
NM_001014	108.452	ribosomal protein S10
NM_001204091	108.452	ribosomal protein S10
parotid gland, adult, ***Catarrhini***		
NM_002723	136988	proline-rich protein BstNI subfamily 4
NM_006250	57348.8	proline-rich protein HaeIII subfamily 1

NM_006249	47975.3	proline-rich protein BstNI subfamily 3
NM_001110213	31692.8	proline-rich protein HaeIII subfamily 2
NM_005042	31692.8	proline-rich protein HaeIII subfamily 2
NM_001900	10157.6	cystatin D
NM_000099	1494.35	cystatin C
NM_001004	1105.09	ribosomal protein, large, P2
NM_001098538	512.105	proline rich 4 (lacrimal)
NM_007244	512.105	proline rich 4 (lacrimal)
salivary gland, adult, ***Catarrhini***		
NM_001322	16457	cystatin SA
NM_006250	15336.8	proline-rich protein HaeIII subfamily 1
NM_001900	8931.88	cystatin D
NM_001110213	6711.33	proline-rich protein HaeIII subfamily 2
NM_005042	6711.33	proline-rich protein HaeIII subfamily 2
NM_001098538	6034.26	proline rich 4 (lacrimal)
NM_007244	6034.26	proline rich 4 (lacrimal)
NM_000099	2302.08	cystatin C
NM_001004	859.543	ribosomal protein, large, P2
NM_001898	575.298	cystatin SN
vagina, adult, ***Homo sapiens***		
NM_005345	19049.3	heat shock 70kDa protein 1A
NM_005346	11243.1	heat shock 70kDa protein 1B
NM_003542	288.354	histone cluster 1, H4c
NM_003544	145.209	histone cluster 1, H4b
NM_006164	139.704	nuclear factor (erythroid-derived 2)-like 2
NM_005526	116.993	heat shock transcription factor 1
NM_003543	114.241	histone cluster 1, H4h

NM_003545	99.7885	histone cluster 1, H4e
NM_017971	86.0246	mitochondrial ribosomal protein L20
NM_005406	80.519	Rho-associated, coiled-coil containing protein kinase 1
placenta, adult, ***Homo/Pan/Gorilla***		
NM_021016	5474.12	pregnancy specific beta-1-glycoprotein 3
NM_000518	3718.3	hemoglobin, beta
NM_002784	1765.17	pregnancy specific beta-1-glycoprotein 9
NM_001184825	1693.14	pregnancy specific beta-1-glycoprotein 1
NM_001184826	1693.14	pregnancy specific beta-1-glycoprotein 1
NM_006905	1693.14	pregnancy specific beta-1-glycoprotein 1
NM_001632	1430.24	alkaline phosphatase, placental
NM_001031850	1212.55	pregnancy specific beta-1-glycoprotein 6
NM_002782	1212.55	pregnancy specific beta-1-glycoprotein 6
NM_003299	1189.73	heat shock protein 90kDa beta (Grp94), member 1
seminal vesicle, adult, ***Catarrhini***		
NM_003007	70653.1	semenogelin I
NM_003008	21241.3	semenogelin II
NM_013230	1660.22	CD24 molecule
NM_001540	1436.72	heat shock 27kDa protein 1
NM_001004	792.435	ribosomal protein, large, P2
NM_184041	329.578	aldolase A, fructose-bisphosphate
NM_000099	328.991	cystatin C
NM_005022	197.473	profilin 1
NM_007278	161.658	GABA(A) receptor-associated protein

NM_005953	158.918	metallothionein 2A
testis, adult, ***Eutheria***		
NM_170610	7779.16	histone cluster 1, H2ba
NM_170745	4616.21	histone cluster 1, H2aa
NM_005323	1436.98	histone cluster 1, H1t
NM_000099	1093.46	cystatin C
NM_003529	1029.76	histone cluster 1, H3a
NM_001004	783.309	ribosomal protein, large, P2
NM_002568	774.761	poly(A) binding protein, cytoplasmic 1
NM_005319	718.593	histone cluster 1, H1c
NM_018955	660.592	ubiquitin B
NM_006082	656.726	tubulin, alpha 1b
uterus, fetal, ***Eutheria***		
NM_002274	4112.03	keratin 13
NM_153490	4112.03	keratin 13
NM_003295	3509.84	tumor protein, translationally-controlled 1
NM_005554	3223.21	keratin 6A
NM_001097589	2454.6	small proline-rich protein 3
NM_005416	2454.6	small proline-rich protein 3
NM_001004	2421.19	ribosomal protein, large, P2
NM_001028	1780.75	ribosomal protein S25
NM_003125	1493.34	small proline-rich protein 1B
NM_000978	1463.44	ribosomal protein L23
putamen, adult, ***Homo sapiens***		
NM_003537	1631.66	histone cluster 1, H3b
NM_003542	1507.45	histone cluster 1, H4c
NM_003535	834.907	histone cluster 1, H3j
NM_003533	464.921	histone cluster 1, H3i
NM_002295	453.387	ribosomal protein SA
NM_003539	431.206	histone cluster 1, H4d

NM_003536	365.549	histone cluster 1, H3h
NM_003545	342.48	histone cluster 1, H4e
NM_002266	316.75	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
NM_003495	238.672	histone cluster 1, H4i
parietal lobe, fetal, ***Homo sapiens***		
NM_003537	727.108	histone cluster 1, H3b
NM_003533	436.071	histone cluster 1, H3i
NM_003542	423.15	histone cluster 1, H4c
NM_003545	419.92	histone cluster 1, H4e
NM_003535	274.24	histone cluster 1, H3j
NM_003531	272.625	histone cluster 1, H3c
NM_021018	270.687	histone cluster 1, H3f
NM_003530	254.698	histone cluster 1, H3d
NM_003539	175.074	histone cluster 1, H4d
NM_003546	167.968	histone cluster 1, H4l
putamen, adult, ***Homo sapiens***		
NM_003537	1631.66	histone cluster 1, H3b
NM_003542	1507.45	histone cluster 1, H4c
NM_003535	834.907	histone cluster 1, H3j
NM_003533	464.921	histone cluster 1, H3i
NM_002295	453.387	ribosomal protein SA
NM_003539	431.206	histone cluster 1, H4d
NM_003536	365.549	histone cluster 1, H3h
NM_003545	342.48	histone cluster 1, H4e
NM_002266	316.75	karyopherin alpha 2 (RAG cohort 1, importin alpha 1)
NM_003495	238.672	histone cluster 1, H4i
parietal lobe, fetal, ***Homo sapiens***		
NM_003537	727.108	histone cluster 1, H3b
NM_003533	436.071	histone cluster 1, H3i
NM_003542	423.15	histone cluster 1, H4c

NM_003545	419.92	histone cluster 1, H4e
NM_003535	274.24	histone cluster 1, H3j
NM_003531	272.625	histone cluster 1, H3c
NM_021018	270.687	histone cluster 1, H3f
NM_003530	254.698	histone cluster 1, H3d
NM_003539	175.074	histone cluster 1, H4d
NM_003546	167.968	histone cluster 1, H4l
dura mater, adult, ***Catarrhini***		
NM_005953	7672.17	metallothionein 2A
NM_021034	2825.61	interferon induced transmembrane protein 3
NM_175617	2239.11	metallothionein 1E
NM_005514	1775.07	major histocompatibility complex, class I, B
NM_006435	1698.99	interferon induced transmembrane protein 2
NM_000099	1672.46	cystatin C
NM_001004	1605.27	ribosomal protein, large, P2
NM_005952	1461.66	metallothionein 1X
NM_033554	1436.79	major histocompatibility complex, class II, DP alpha 1
NM_021103	1241.36	thymosin beta 10

Bibliography

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.

ENCODE_UCSC_Tfbs_V2. "TfbsClusteredV2." from http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV2.

Huminiecki, L. (2005). Expression pattern divergence in the type A GPCR family. CMB, Karolinska Institutet, Pfizer, UK.

Huminiecki, L. and R. Bicknell (2000). "In silico cloning of novel endothelial-specific genes." Genome research **10**(11): 1796-1806.

Huminiecki, L., L. Goldovsky, et al. (2009). "Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom." BMC evolutionary biology **9**: 28.

Huminiecki, L., M. Gorn, et al. (2002). "Magic roundabout is a new member of the roundabout receptor family that is endothelial specific and expressed at sites of active angiogenesis." Genomics **79**(4): 547-552.

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

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 research **14**(10A): 1870-1879.

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.

Jensen-Seaman, M. I. and W. H. Li (2003). "Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen." Journal of Molecular Evolution **57**(3): 261-270.

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

Kanamori-Katayama, M., M. Itoh, et al. (2011). "Unamplified cap analysis of gene expression on a single-molecule sequencer." Genome Research **21**(7): 1150-1159.

Karin, M. (2006). "Nuclear factor-kappaB in cancer development and progression." Nature **441**(7092): 431-436.

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.

Lilja, H., P. A. Abrahamsson, et al. (1989). "Semenogelin, the predominant protein in human semen. Primary structure and identification of closely related proteins in the male accessory sex glands and on the spermatozoa." The Journal of biological chemistry **264**(3): 1894-1900.

Masatoshi Nei, S. K. (2000). Molecular Evolution and Phylogenetics, Oxford University Press, USA.

Rovin, R. A. and R. Winn (2005). "Pokemon expression in malignant glioma: an application of bioinformatics methods." Neurosurgical focus **19**(4): E8.

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.

TF101109. "Rac/cdc42 TreeFam family tree." from http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF101109.

Figure 1

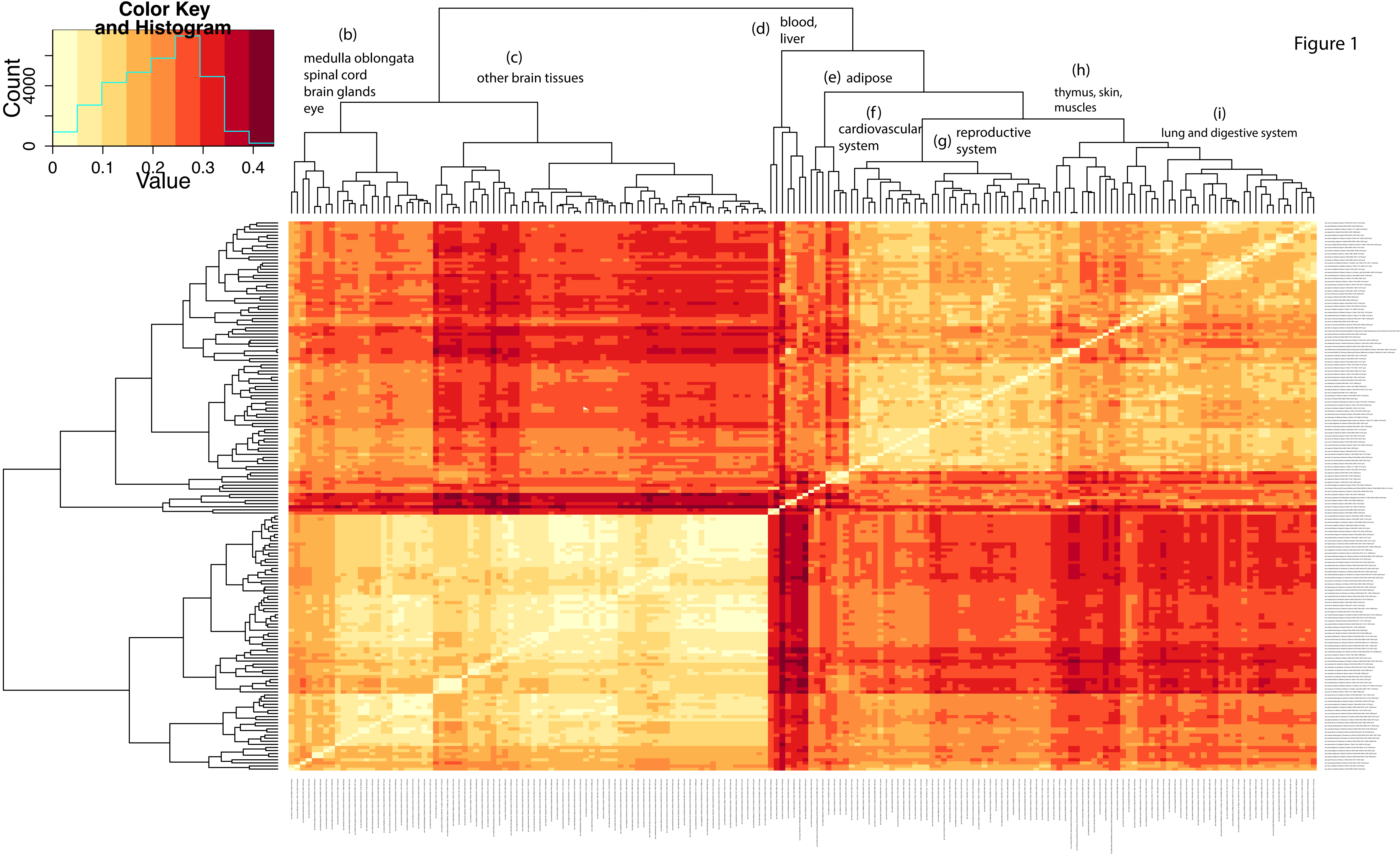
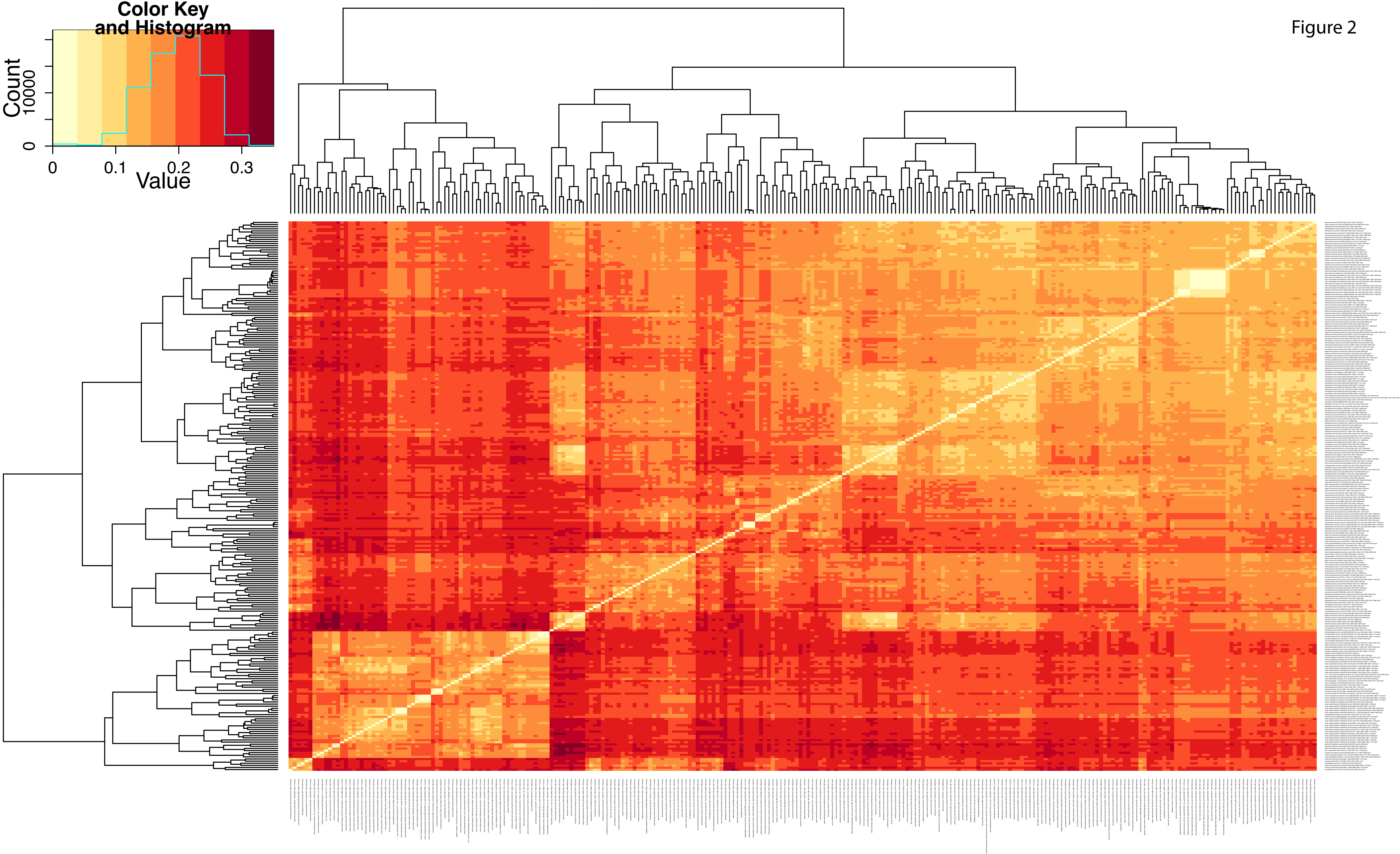

Color Key
and Histogram
Count
0
4000
0
0.1
0.2
0.3
0.4
Value
(b)
medulla oblongata
spinal cord
brain glands
eye
(c)
other brain tissues
(d)
blood,
liver
(e) adipose
(f)
cardiovascular
system
(g)
reproductive
system
(h)
thymus, skin,
muscles
(i)
lung and digestive system

Figure 2



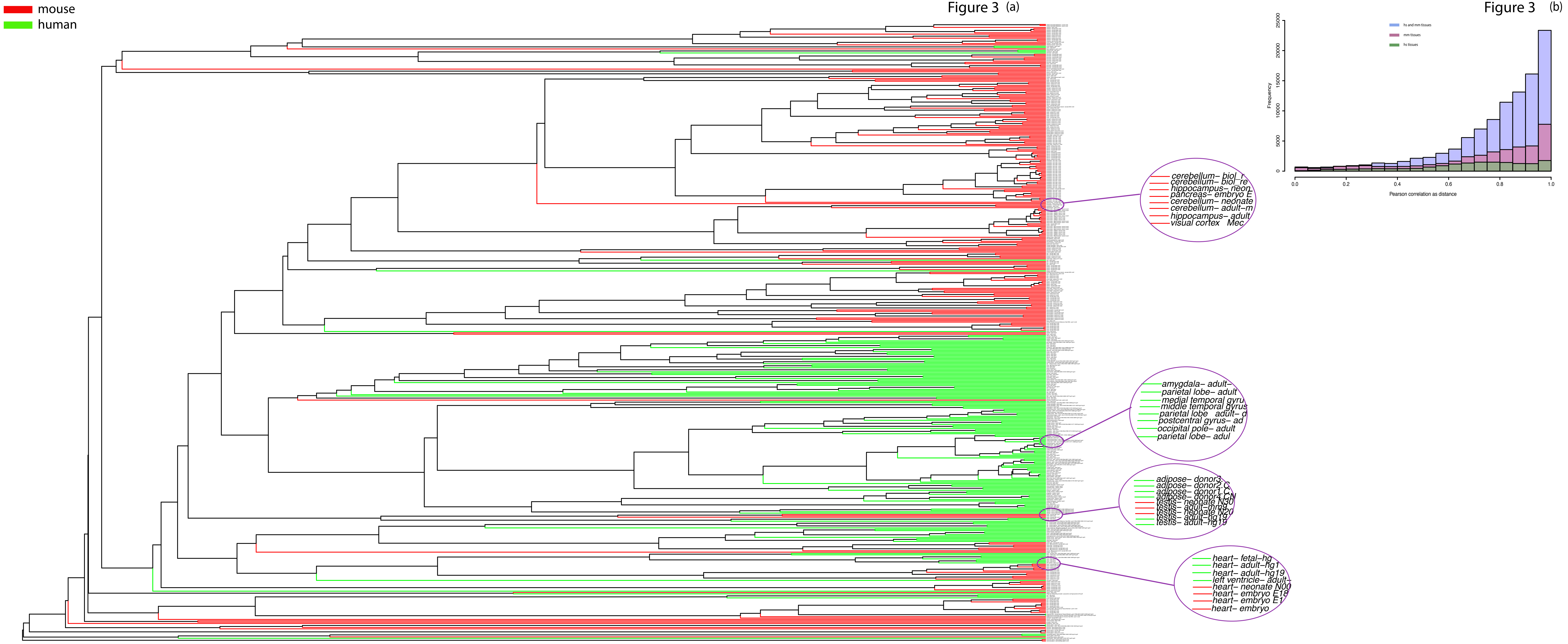


Figure 3 (a)

Figure 3 (b)

Figure 4

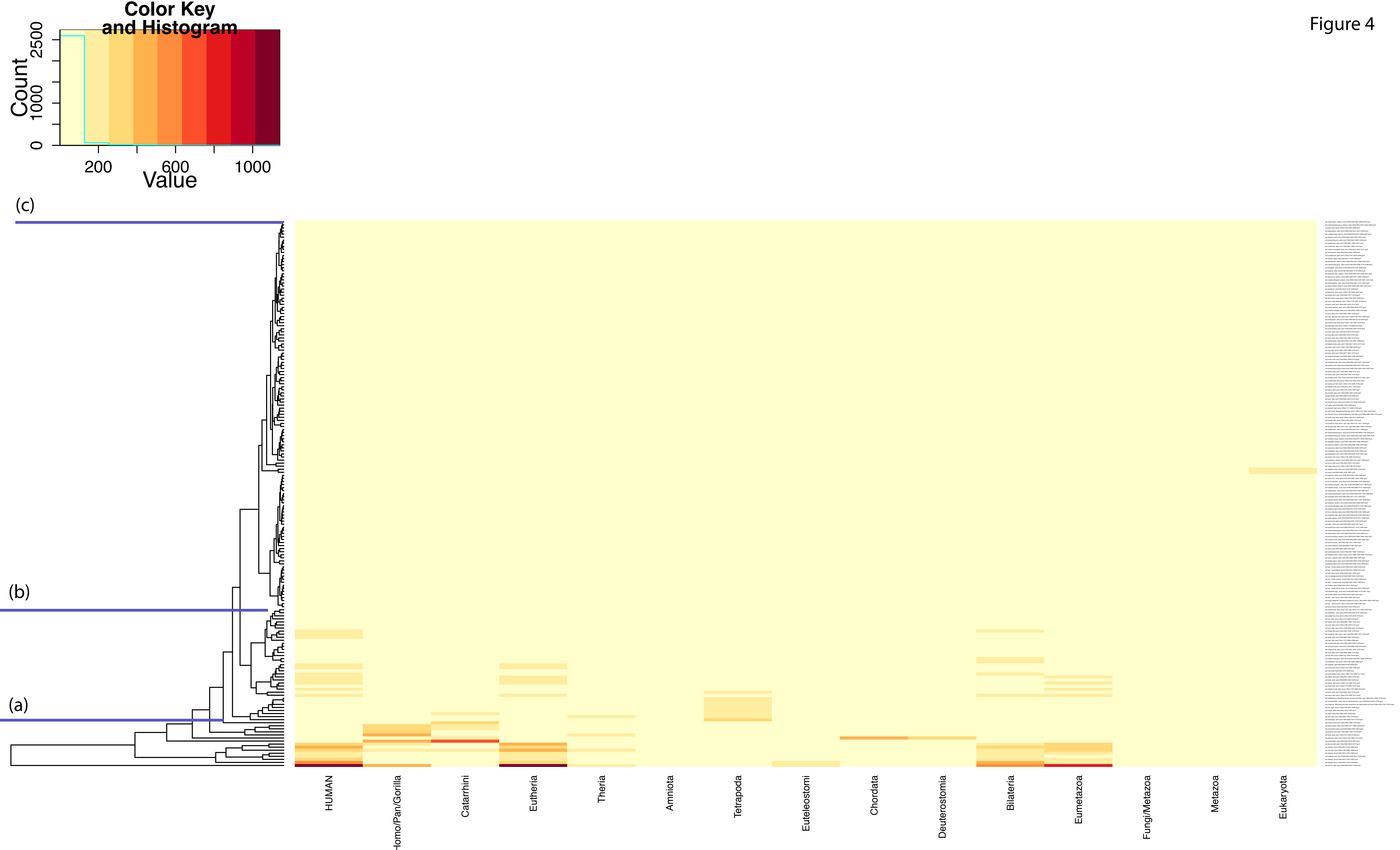
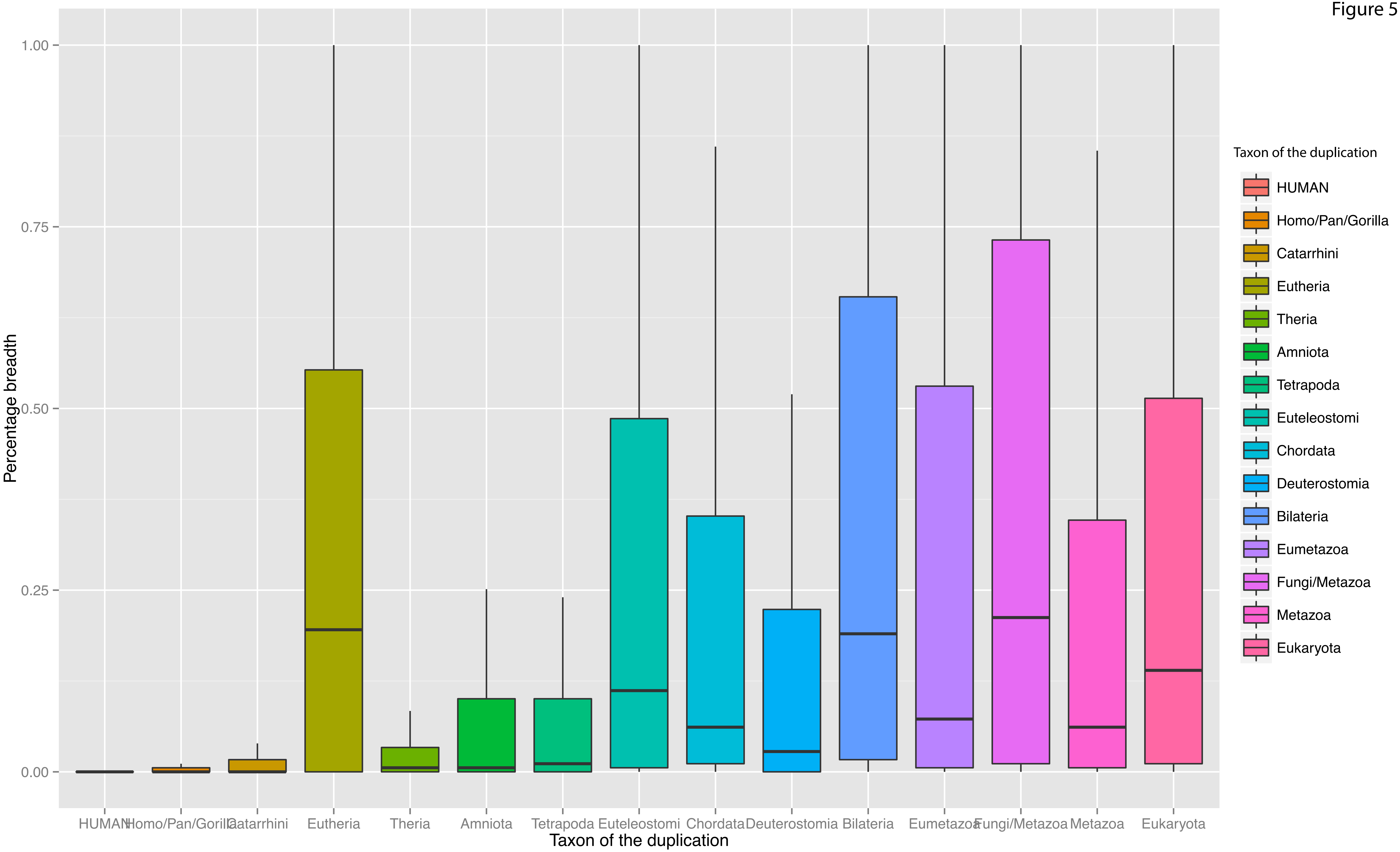
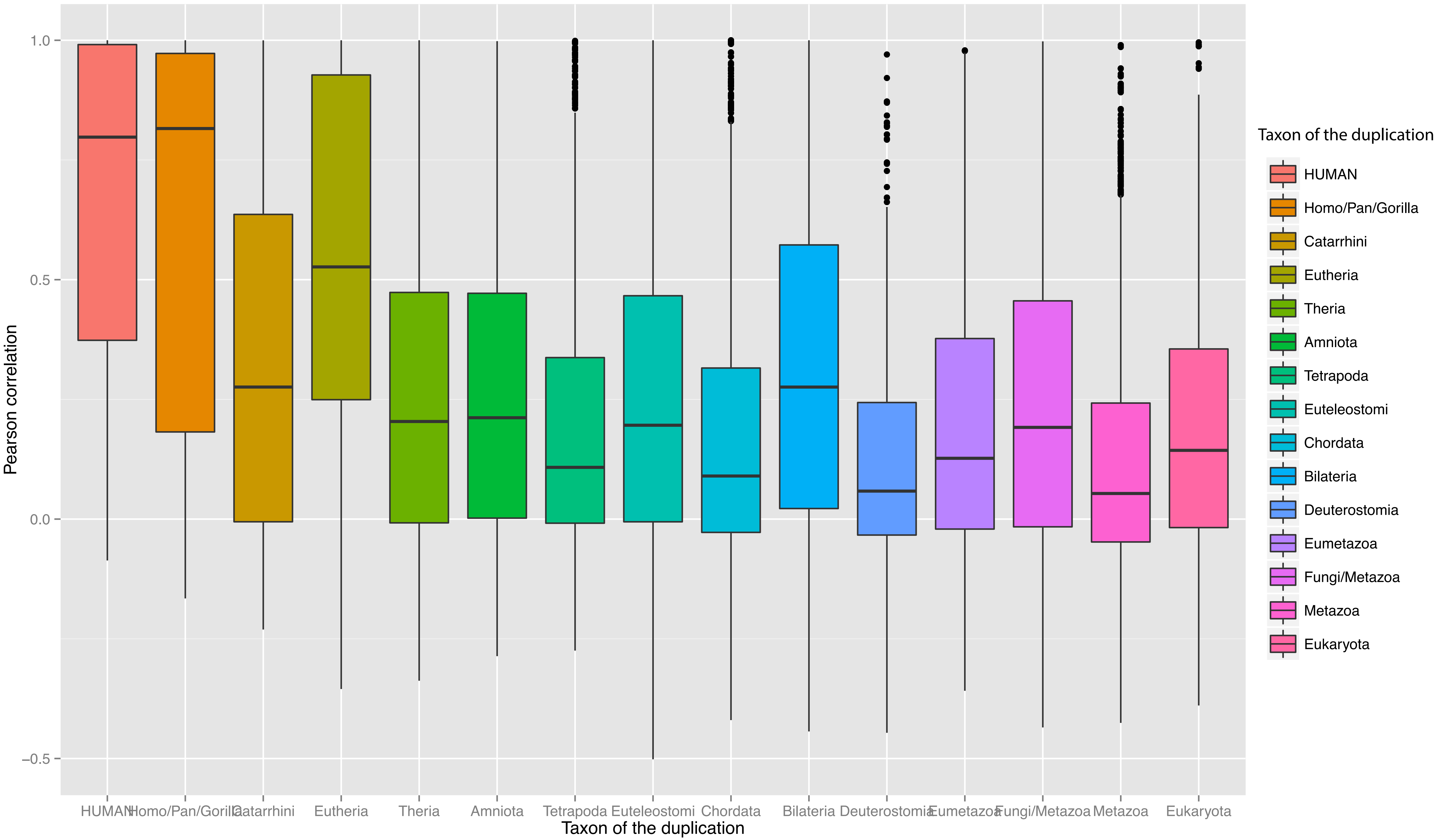

Color Key
and Histogram
Count
0
1000
2500
200
600
1000
Value
(c)
(b)
(a)
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 5



Percentage breadth
1.00
0.75
0.50
0.25
0.00
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 6



Pearson correlation
1.0
0.5
0.0
−0.5
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

(a)
Number of pairs
0
10000
20000
-0.5
0.0
0.5
1.0
Pearson correlation
(b)
Number of pairs
0
20000
40000
60000
-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 8

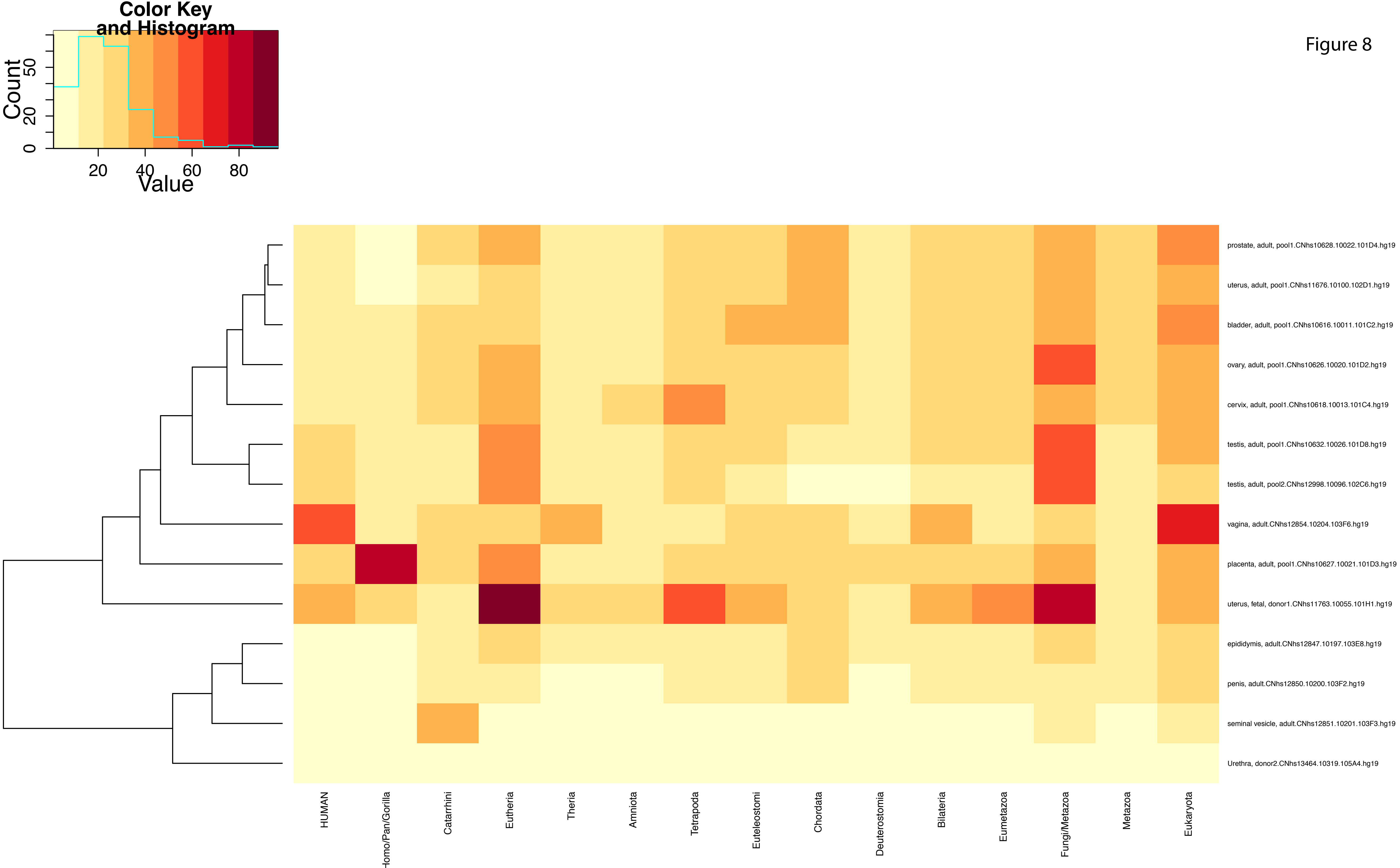

Color Key
and Histogram
Count
0
20
50
20
40
60
80
Value
prostate, adult, pool1.CNhs10628.10022.101D4.hg19
uterus, adult, pool1.CNhs11676.10100.102D1.hg19
bladder, adult, pool1.CNhs10616.10011.101C2.hg19
ovary, adult, pool1.CNhs10626.10020.101D2.hg19
cervix, adult, pool1.CNhs10618.10013.101C4.hg19
testis, adult, pool1.CNhs10632.10026.101D8.hg19
testis, adult, pool2.CNhs12998.10096.102C6.hg19
vagina, adult.CNhs12854.10204.103F6.hg19
placenta, adult, pool1.CNhs10627.10021.101D3.hg19
uterus, fetal, donor1.CNhs11763.10055.101H1.hg19
epididymis, adult.CNhs12847.10197.103E8.hg19
penis, adult.CNhs12850.10200.103F2.hg19
seminal vesicle, adult.CNhs12851.10201.103F3.hg19
Urethra, donor2.CNhs13464.10319.105A4.hg19
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 9

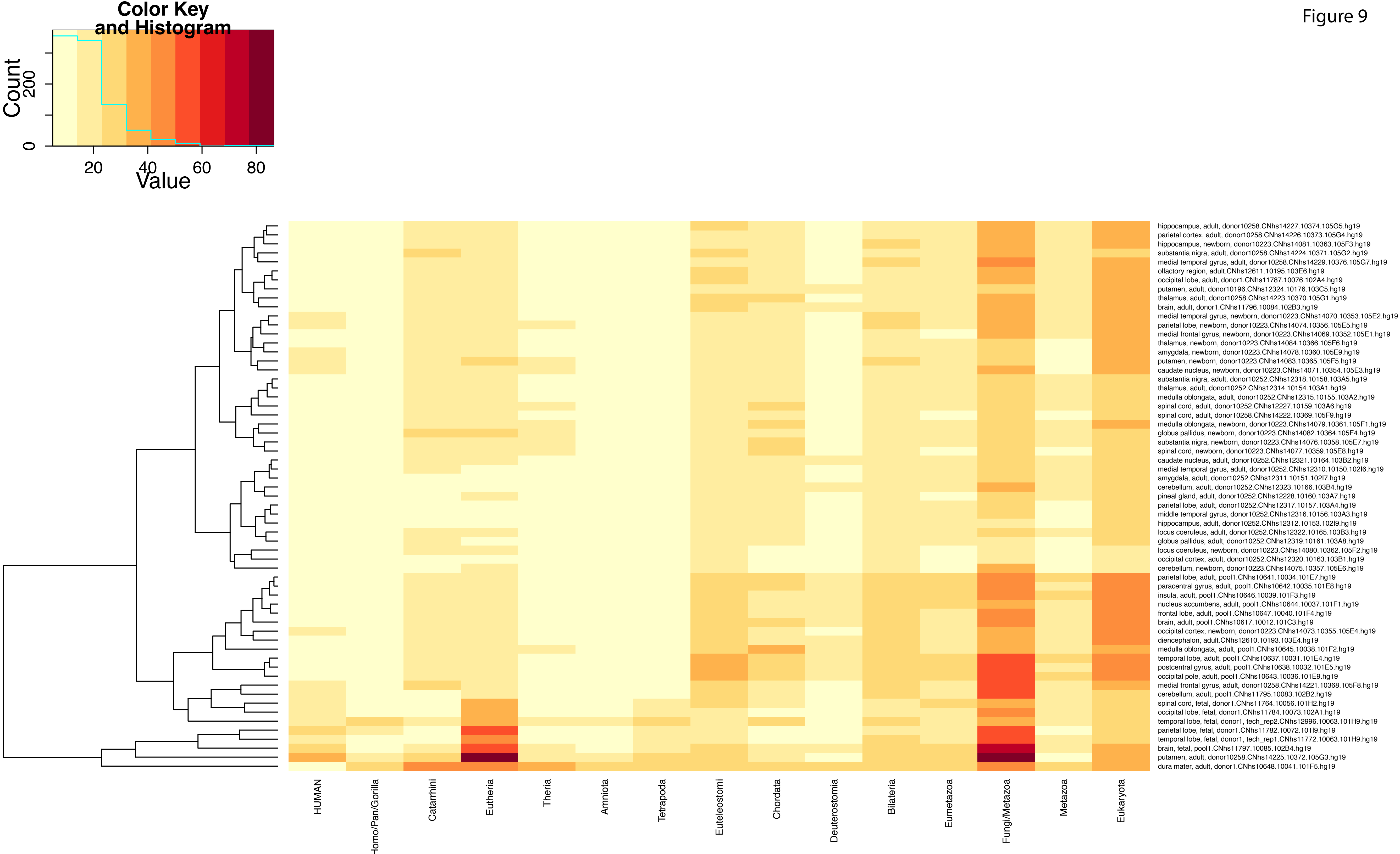

Color Key
and Histogram
Count
0
200
20
40
60
80
Value
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
hippocampus, newborn, donor10223.CNhs14081.10363.105F3.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
medial temporal gyrus, adult, donor10258.CNhs14229.10376.105G7.hg19
olfactory region, adult.CNhs12611.10195.103E6.hg19
occipital lobe, adult, donor1.CNhs11787.10076.102A4.hg19
putamen, adult, donor10196.CNhs12324.10176.103C5.hg19
thalamus, adult, donor10258.CNhs14223.10370.105G1.hg19
brain, adult, donor1.CNhs11796.10084.102B3.hg19
medial temporal gyrus, newborn, donor10223.CNhs14070.10353.105E2.hg19
parietal lobe, newborn, donor10223.CNhs14074.10356.105E5.hg19
medial frontal gyrus, newborn, donor10223.CNhs14069.10352.105E1.hg19
thalamus, newborn, donor10223.CNhs14084.10366.105F6.hg19
amygdala, newborn, donor10223.CNhs14078.10360.105E9.hg19
putamen, newborn, donor10223.CNhs14083.10365.105F5.hg19
caudate nucleus, newborn, donor10223.CNhs14071.10354.105E3.hg19
substantia nigra, adult, donor10252.CNhs12318.10158.103A5.hg19
thalamus, adult, donor10252.CNhs12314.10154.103A1.hg19
medulla oblongata, adult, donor10252.CNhs12315.10155.103A2.hg19
spinal cord, adult, donor10252.CNhs12227.10159.103A6.hg19
spinal cord, adult, donor10258.CNhs14222.10369.105F9.hg19
medulla oblongata, newborn, donor10223.CNhs14079.10361.105F1.hg19
globus pallidus, newborn, donor10223.CNhs14082.10364.105F4.hg19
substantia nigra, newborn, donor10223.CNhs14076.10358.105E7.hg19
spinal cord, newborn, donor10223.CNhs14077.10359.105E8.hg19
caudate nucleus, adult, donor10252.CNhs12321.10164.103B2.hg19
medial temporal gyrus, adult, donor10252.CNhs12310.10150.102I6.hg19
amygdala, adult, donor10252.CNhs12311.10151.102I7.hg19
cerebellum, adult, donor10252.CNhs12323.10166.103B4.hg19
pineal gland, adult, donor10252.CNhs12228.10160.103A7.hg19
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
middle temporal gyrus, donor10252.CNhs12316.10156.103A3.hg19
hippocampus, adult, donor10252.CNhs12312.10153.102I9.hg19
locus coeruleus, adult, donor10252.CNhs12322.10165.103B3.hg19
globus pallidus, adult, donor10252.CNhs12319.10161.103A8.hg19
locus coeruleus, newborn, donor10223.CNhs14080.10362.105F2.hg19
occipital cortex, adult, donor10252.CNhs12320.10163.103B1.hg19
cerebellum, newborn, donor10223.CNhs14075.10357.105E6.hg19
parietal lobe, adult, pool1.CNhs10641.10034.101E7.hg19
paracentral gyrus, adult, pool1.CNhs10642.10035.101E8.hg19
insula, adult, pool1.CNhs10646.10039.101F3.hg19
nucleus accumbens, adult, pool1.CNhs10644.10037.101F1.hg19
frontal lobe, adult, pool1.CNhs10647.10040.101F4.hg19
brain, adult, pool1.CNhs10617.10012.101C3.hg19
occipital cortex, newborn, donor10223.CNhs14073.10355.105E4.hg19
diencephalon, adult.CNhs12610.10193.103E4.hg19
medulla oblongata, adult, pool1.CNhs10645.10038.101F2.hg19
temporal lobe, adult, pool1.CNhs10637.10031.101E4.hg19
postcentral gyrus, adult, pool1.CNhs10638.10032.101E5.hg19
occipital pole, adult, pool1.CNhs10643.10036.101E9.hg19
medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
cerebellum, adult, pool1.CNhs11795.10083.102B2.hg19
spinal cord, fetal, donor1.CNhs11764.10056.101H2.hg19
occipital lobe, fetal, donor1.CNhs11784.10073.102A1.hg19
temporal lobe, fetal, donor1, tech_rep2.CNhs12996.10063.101H9.hg19
parietal lobe, fetal, donor1.CNhs11782.10072.101I9.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
dura mater, adult, donor1.CNhs10648.10041.101F5.hg19
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 10 a

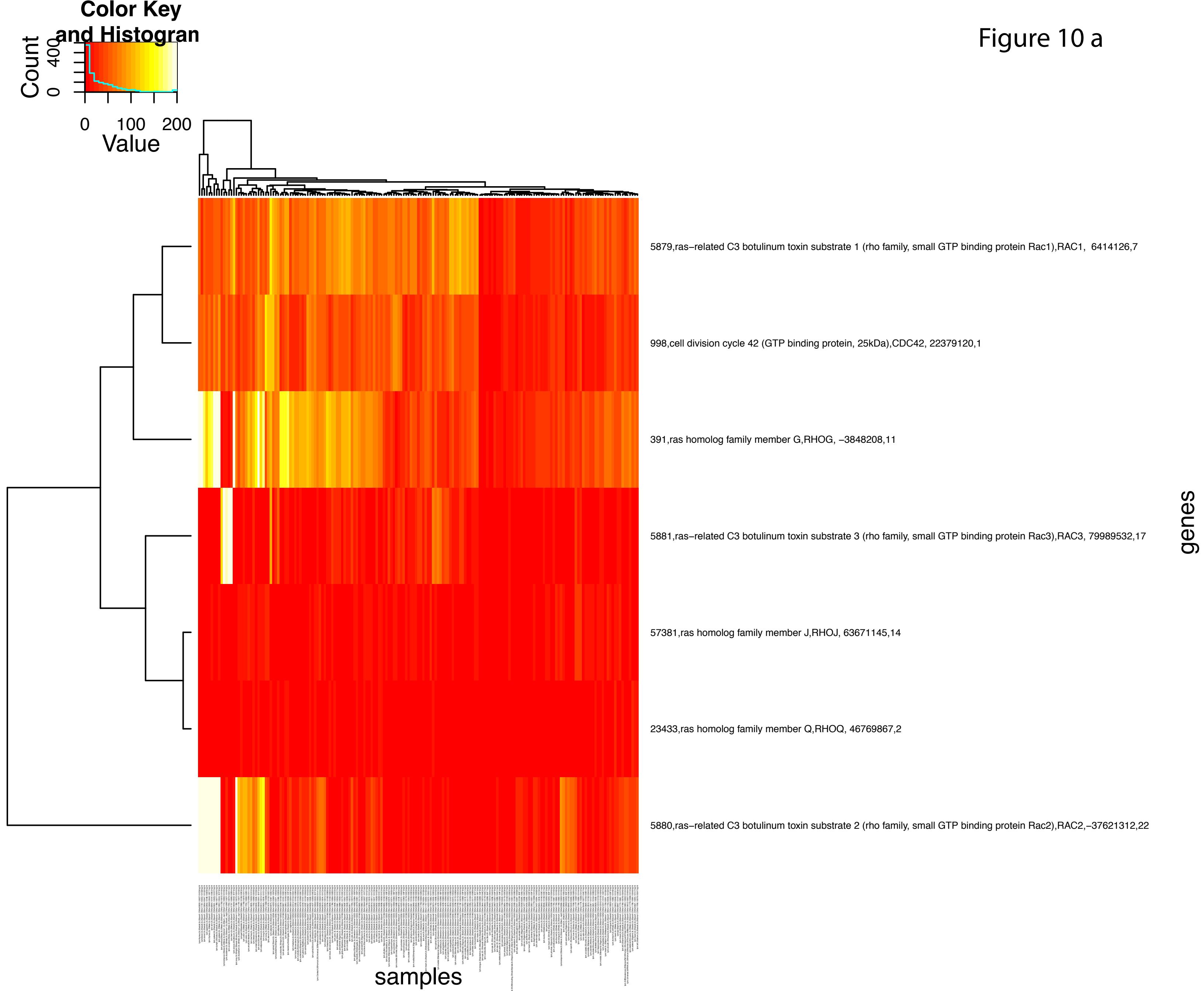


Figure 10b

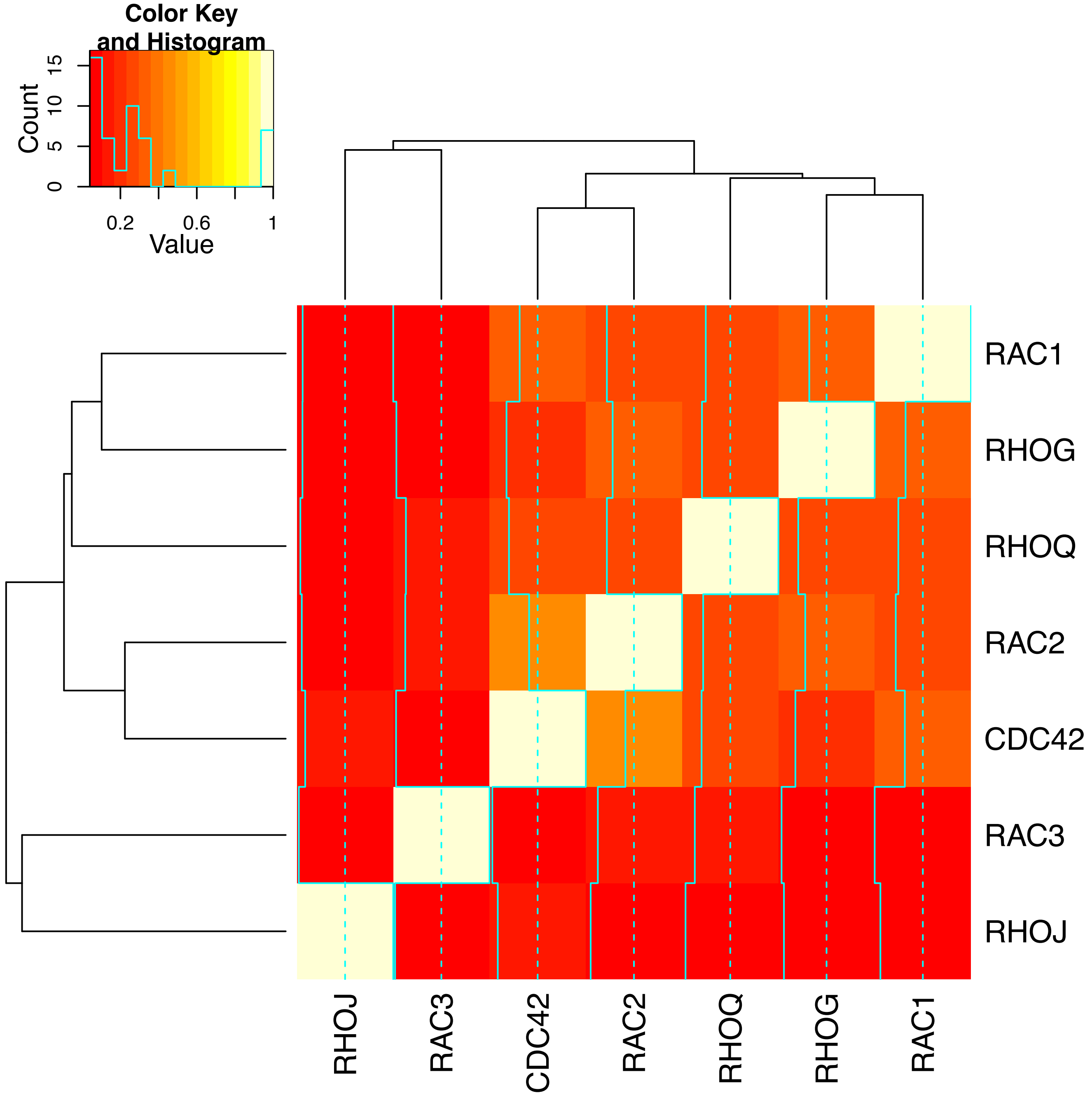


Figure S1 (a)

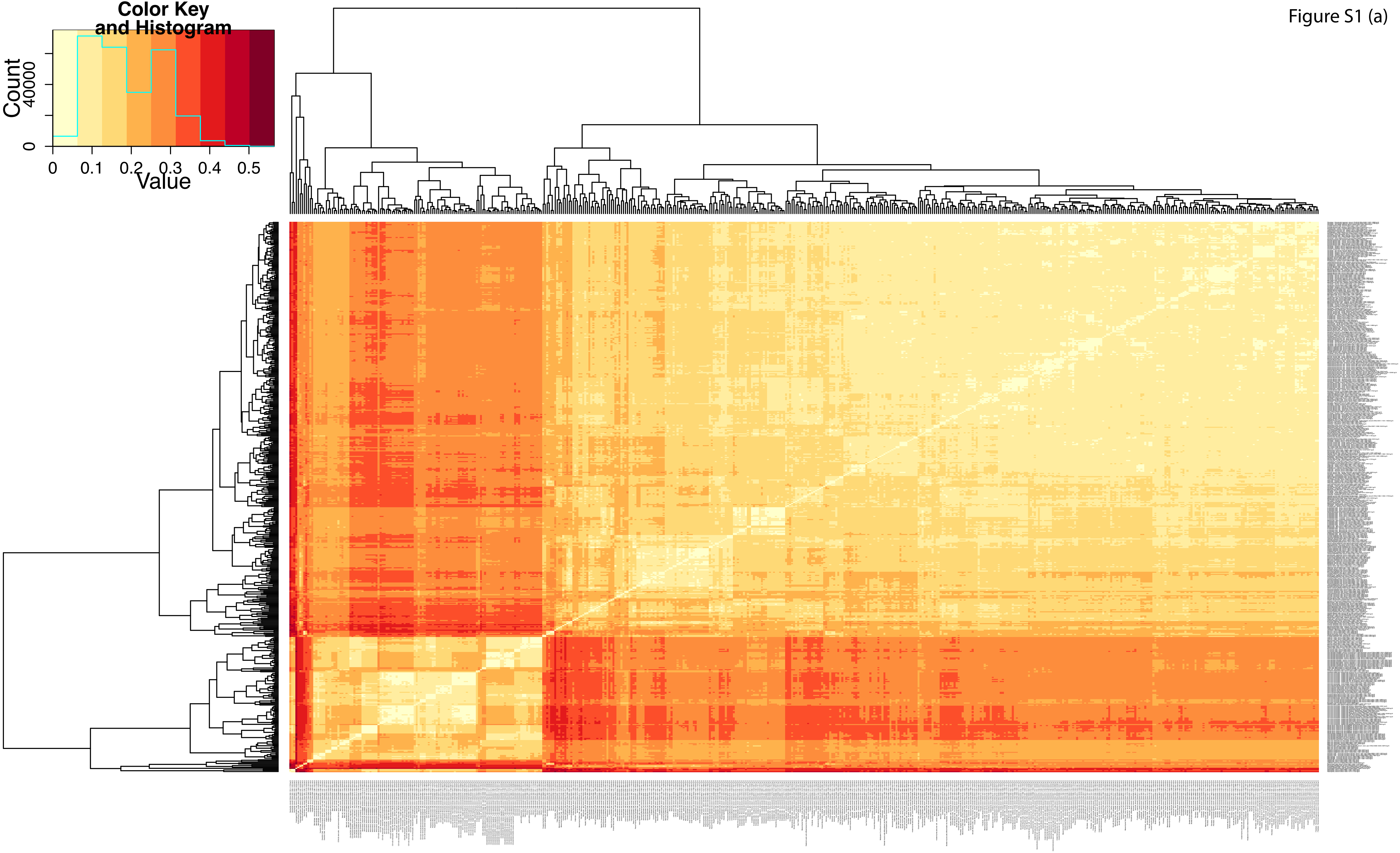


Figure S1 (b)

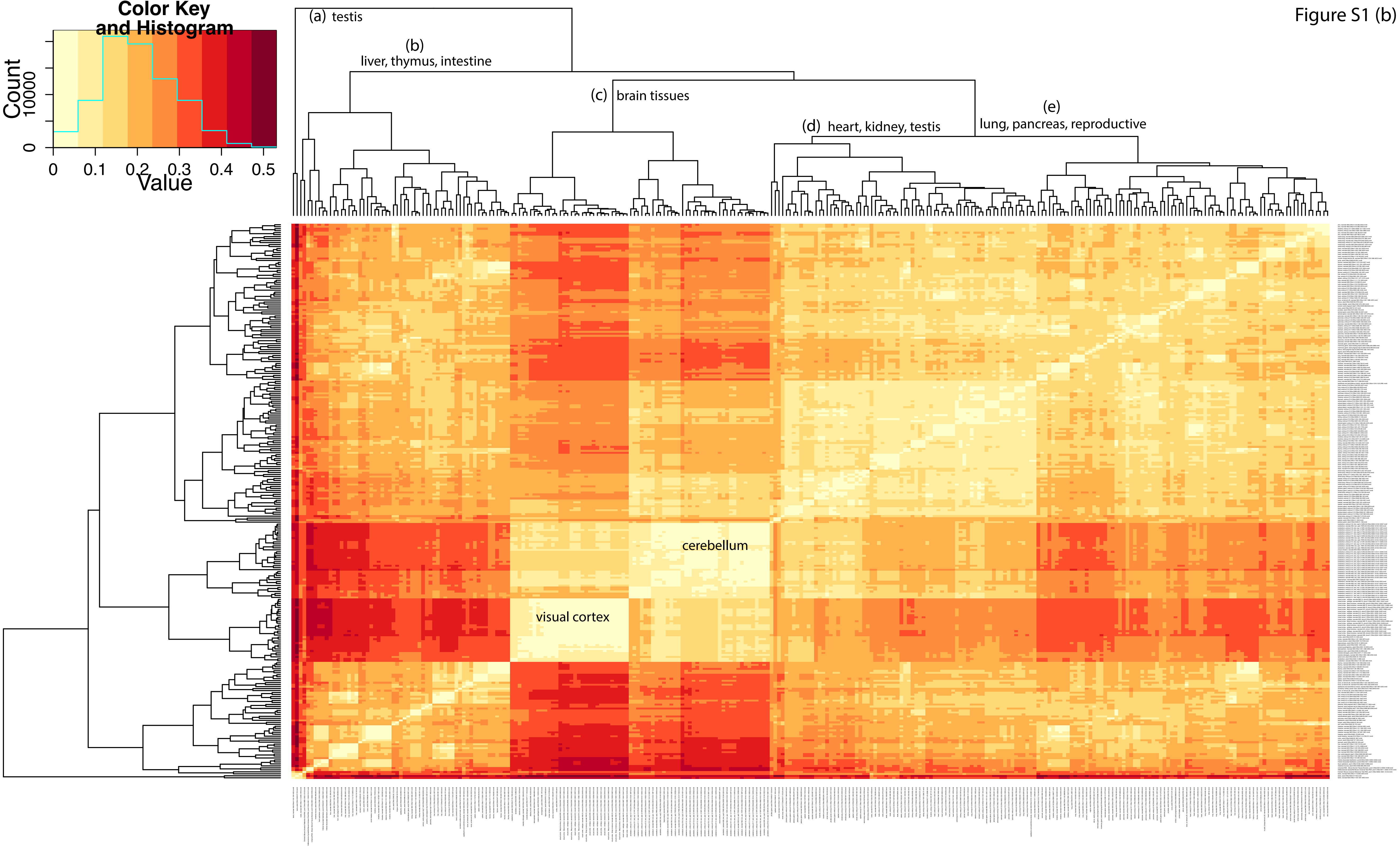

Color Key
and Histogram
Count
0
10000
0
0.1
0.2
0.3
0.4
0.5
Value
(a) testis
(b)
liver, thymus, intestine
(c) brain tissues
(d) heart, kidney, testis
(e)
lung, pancreas, reproductive
cerebellum
visual cortex

Figure S1 (c)

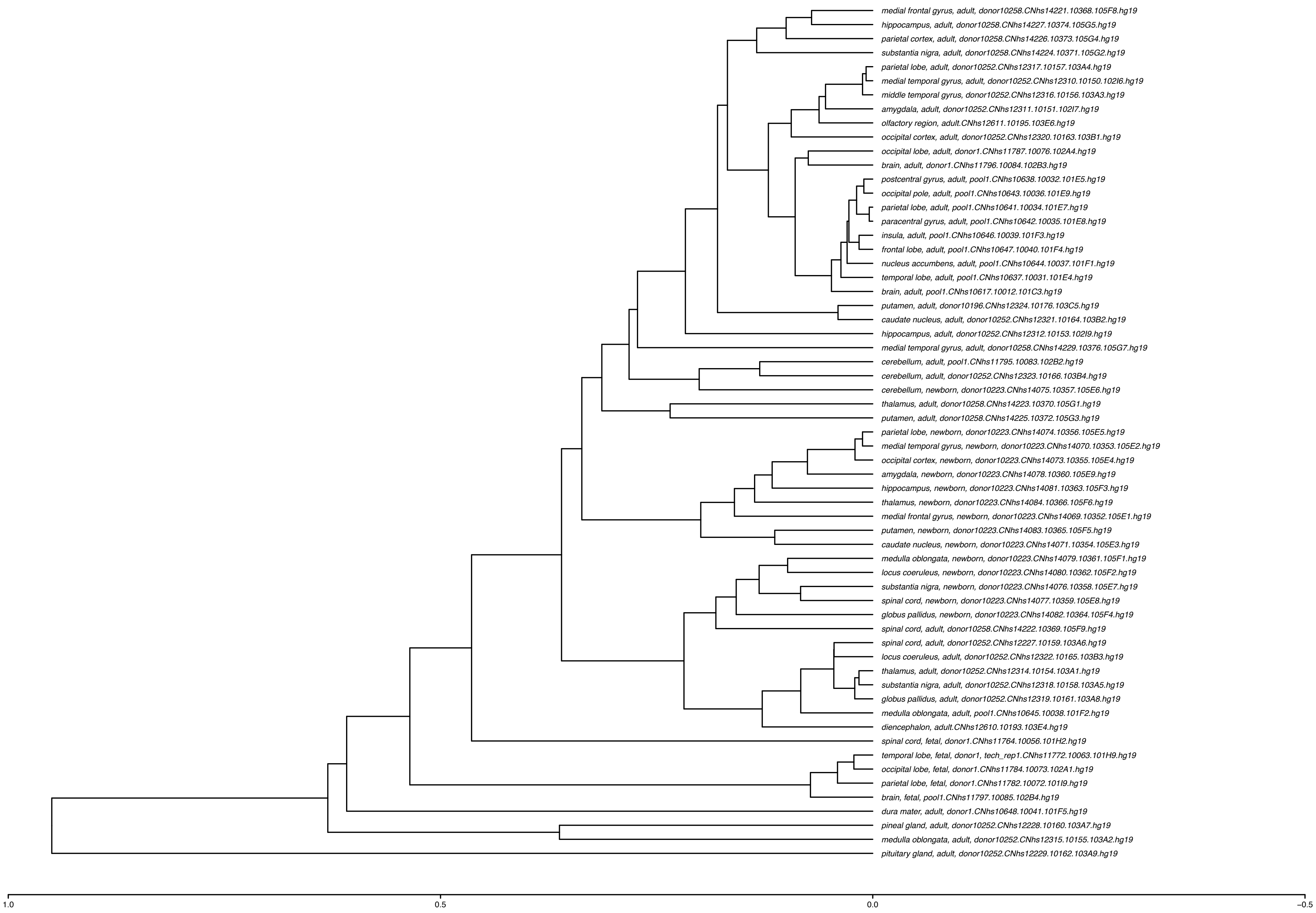
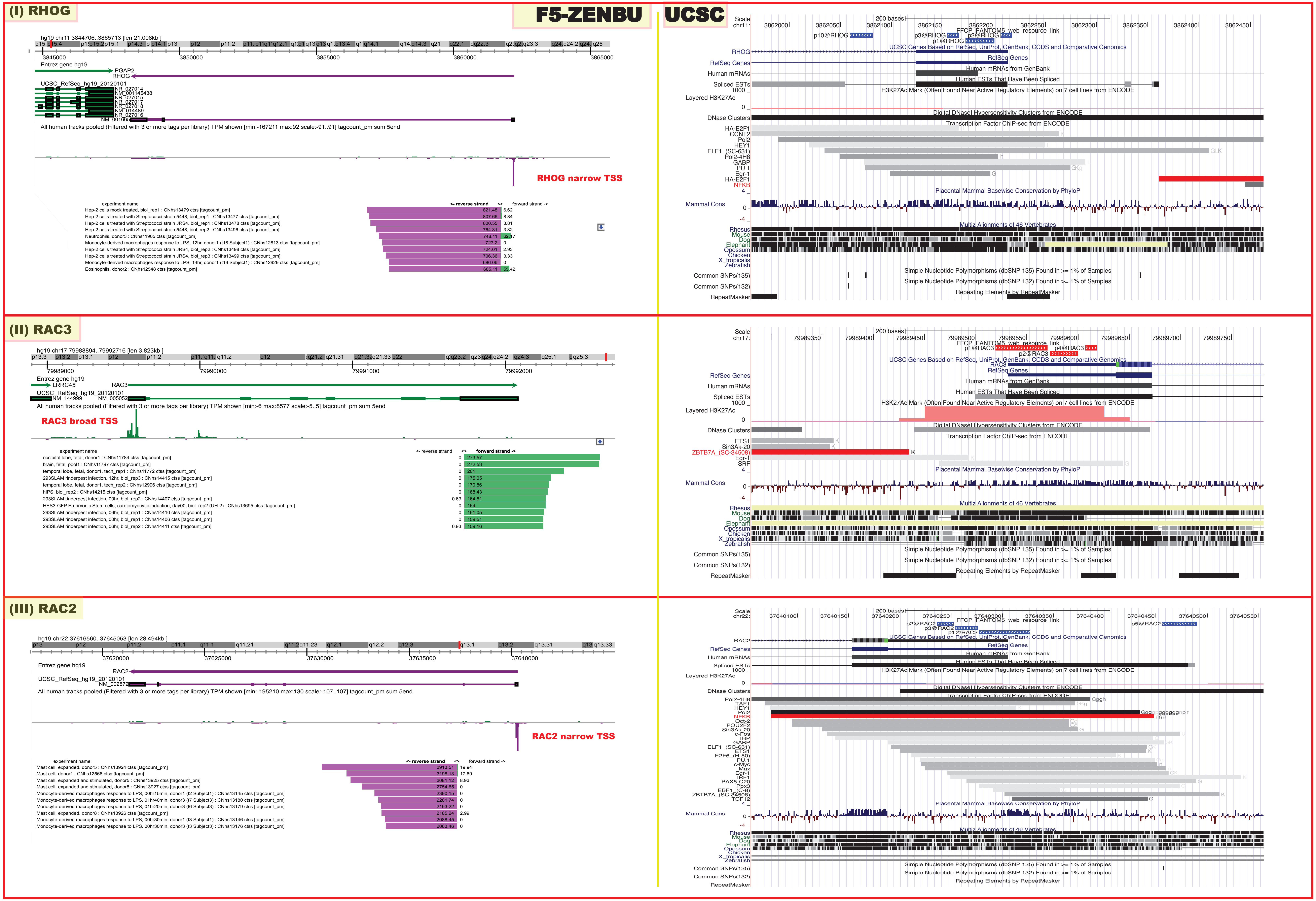

medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
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

Figure S2


(I) RHOG
F5-ZENBU
UCSC
hg19 chr11 3844706..3865713 [len 21.008kb]
3845000
3850000
3855000
3860000
3865000
Entrez gene hg19
PGAP2
RHOG
UCSC_RefSeq_hg19_20120101
NR_027014
NM_001145438
NR_027015
NR_027017
NR_027018
NM_014489
NR_027016
NM_001665
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [min:-167211 max:92 scale:-91..91] tagcount_pm sum 5end
RHOG narrow TSS
Scale
chr11:
200 bases
FFCP_FANTOM5_web_resource_link
p10@RHOG
p3@RHOG
p2@RHOG
p1@RHOG
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
HA-E2F1
CCNT2
Pol2
HEY1
ELF1_(SC-631)
Pol2-4H8
GABP
PU.1
Egr-1
HA-E2F1
NFKB
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples
Common SNPs(135)
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Common SNPs(132)
Repeating Elements by RepeatMasker
RepeatMasker
(II) RAC3
hg19 chr17 79988894..79992716 [len 3.823kb]
79989000
79990000
79991000
79992000
Entrez gene hg19
LRRC45
RAC3
UCSC_RefSeq_hg19_20120101
NM_144999
NM_005052
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [min:-6 max:8577 scale:-5..5] tagcount_pm sum 5end
RAC3 broad TSS
chr17:
p1@RAC3
p4@RAC3
p2@RAC3
RAC3
ETS1
Sin3Ak-20
ZBTB7A_(SC-34508)
Egr-1
SRF
(III) RAC2
hg19 chr22 37616560..37645053 [len 28.494kb]
37620000
37625000
37630000
37635000
37640000
RAC2
NM_002872
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [min:-195210 max:130 scale:-107..107] tagcount_pm sum 5end
RAC2 narrow TSS
chr22:
p2@RAC2
p3@RAC2
p1@RAC2
p5@RAC2
Pol2-4H8
TAF1
HEY1
Pol2
NFKB
Oct-2
POU2F2
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

(I) RHOG

experiment name	<- reverse strand	<> forward strand ->
Hep-2 cells mock treated, biol_rep1 : CNhs13479 ctss [tagcount_pm]	821.48	6.62
Hep-2 cells treated with Streptococci strain 5448, biol_rep1 : CNhs13477 ctss [tagcount_pm]	807.66	8.84
Hep-2 cells treated with Streptococci strain JRS4, biol_rep1 : CNhs13478 ctss [tagcount_pm]	800.55	3.81
Hep-2 cells treated with Streptococci strain 5448, biol_rep2 : CNhs13496 ctss [tagcount_pm]	764.31	3.32
Neutrophils, donor3 : CNhs11905 ctss [tagcount_pm]	748.11	62.17
Monocyte-derived macrophages response to LPS, 12hr, donor1 (t18 Subject1) : CNhs12813 ctss [tagcount_pm]	727.2	0
Hep-2 cells treated with Streptococci strain JRS4, biol_rep2 : CNhs13498 ctss [tagcount_pm]	724.01	2.93
Hep-2 cells treated with Streptococci strain JRS4, biol_rep3 : CNhs13499 ctss [tagcount_pm]	706.36	3.33
Monocyte-derived macrophages response to LPS, 14hr, donor1 (t19 Subject1) : CNhs12929 ctss [tagcount_pm]	686.06	0
Eosinophils, donor2 : CNhs12548 ctss [tagcount_pm]	685.11	55.42

(II) RAC3

experiment name	<- reverse strand	<> forward strand ->
occipital lobe, fetal, donor1 : CNhs11784 ctss [tagcount_pm]	0	273.57
brain, fetal, pool1 : CNhs11797 ctss [tagcount_pm]	0	272.53
temporal lobe, fetal, donor1, tech_rep1 : CNhs11772 ctss [tagcount_pm]	0	201
293SLAM rinderpest infection, 12hr, biol_rep3 : CNhs14415 ctss [tagcount_pm]	0	175.05
temporal lobe, fetal, donor1, tech_rep2 : CNhs12996 ctss [tagcount_pm]	0	170.86
hIPS, biol_rep2 : CNhs14215 ctss [tagcount_pm]	0	168.43
293SLAM rinderpest infection, 00hr, biol_rep2 : CNhs14407 ctss [tagcount_pm]	0.63	164.51
HES3-GFP Embryonic Stem cells, cardiomyocytic induction, day00, biol_rep2 (UH-2) : CNhs13695 ctss [tagcount_pm]	0	164
293SLAM rinderpest infection, 06hr, biol_rep1 : CNhs14410 ctss [tagcount_pm]	0	161.05
293SLAM rinderpest infection, 00hr, biol_rep1 : CNhs14406 ctss [tagcount_pm]	0	159.51
293SLAM rinderpest infection, 06hr, biol_rep2 : CNhs14411 ctss [tagcount_pm]	0.93	159.16

(III) RAC2

experiment name	<- reverse strand	<> forward strand ->
Mast cell, expanded, donor5 : CNhs13924 ctss [tagcount_pm]	3913.51	19.94
Mast cell, donor1 : CNhs12566 ctss [tagcount_pm]	3198.13	17.69
Mast cell, expanded and stimulated, donor5 : CNhs13925 ctss [tagcount_pm]	3081.12	8.93
Mast cell, expanded and stimulated, donor8 : CNhs13927 ctss [tagcount_pm]	2754.65	0
Monocyte-derived macrophages response to LPS, 00hr15min, donor1 (t2 Subject1) : CNhs13145 ctss [tagcount_pm]	2390.15	0
Monocyte-derived macrophages response to LPS, 01hr40min, donor3 (t7 Subject3) : CNhs13180 ctss [tagcount_pm]	2281.74	0
Monocyte-derived macrophages response to LPS, 01hr20min, donor3 (t6 Subject3) : CNhs13179 ctss [tagcount_pm]	2193.22	0
Mast cell, expanded, donor8 : CNhs13926 ctss [tagcount_pm]	2185.24	2.99
Monocyte-derived macrophages response to LPS, 00hr30min, donor1 (t3 Subject1) : CNhs13146 ctss [tagcount_pm]	2088.45	0
Monocyte-derived macrophages response to LPS, 00hr30min, donor3 (t3 Subject3) : CNhs13176 ctss [tagcount_pm]	2063.46	0