Title page

Oxana Sachenkova(1, 2), Alistair Forrest (3), Carsten Daub (3) and Lukasz Huminiecki (1, 2, 4)

1. Department of Biochemistry and Biophysics, Stockholmockholm 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

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-CAGE encyclopedia of transcription start sites

Abstract:

New genes in animal genomes originate through duplication of genes already present, and consecutive rounds of gene duplication in turn give rise to gene families. Since animals are primarily characterized by multicellularity and existence of complex organs and tissues composed of a mixture of cell types, understanding the evolution of gene expression in human gene families is of primary importance to understanding human evolution. To this purpose, we use single-nucleotide resolution atlas of human transcription start sites, FANTOM5-CAGE, arguably the most complete and uniform functional genomics dataset ever generated.

First, we examined the overall expression divergence rate in different tissues, and found that FANTOM5 samples grouped into three distinct categories with respect to global expression evolutionary rate (dynamic, intermediate and static). Next, we dated gene duplications by phylogenetic timing, investigated paralog expression divergence, and tissue-specificity of expression depending on age of duplication. There was a strong trend for recently evolved genes to be tissue-specific, with the exception of duplications mapping to placental mammals. In fact, several lines of evidence suggest 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 FANTOM5-CAGE lies in facilitating comparison of normal tissues with cancer samples. Interestingly, when clustering algorithm was applied, a major divide between leukemias and solid tumors was seen in cancer cell lines, suggesting that these two major clinical subgroups of carcinomas, are readily distinguished by their CAGE expression signatures. 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, suggesting that global devolution and loss-of-normal evolutionary constraints on expression patterns accompany malignant transformation.

Finally, we defined the concept of phyloexpression signatures as strong associations between duplications of certain ages and expression localities in the FANTOM5 atlas. We show how phyloexpression signatures can be used to generate novel hypotheses on the nature of animal evolution: central nervous system and reproductive tract are discussed as two focused examples.

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
CCL	cancer cell-lines
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
hs-T	human tissues
hs-CCL	human cancer cell lines
hs-PC	human primary cells
ISHC	inter-species hierarchical clustering (human-mouse tissue clusters matched by linking ortholog gene expression)
JI	Jaccard index
MEEP	mutually exclusive expression of paralogs
NM-clusters	human-mouse tissue clusters derived though name matching
NFKB	nuclear factor kappa B
PC	primary cells
PC	Person correlation
PC1	first principal component
PC2	second principal component
PCA	principal component analysis
PFAM	protein domains from the PFAM database
PhyloSigs	phyloexpression signatures
RAC2	rho family, small GTP binding protein Rac2
RAC3	rho family, small GTP binding protein Rac3
Rac/cdc42	subfamily of Rho GTPases (TreeFam family id TF101109)
RHOG	Ras homology Growth-related
T	tissues
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 version 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 of great interest. 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 this end, we apply the powerful and novel single-base resolution encyclopedia of vertebrate expression patterns and transcription start sites FANTOM5 cap analysis of gene expression (FANTOM5-CAGE).

FANTOM5-CAGE is arguably the most complete functional genomics dataset generated to date (FANTOM5 2013). FANTOM5-CAGE includes 952 human and 396 mouse tissues (T), primary cells (PC) and cancer cell-lines (CCL). The major strengths of FANTOM5-CAGE are comprehensiveness and technological uniformity. Moreover, FANTOM5-CAGE explores the entire genome space, and allows for comparison between tissue samples, primary cells in culture, and cancer cell-lines. Table 1 briefly summarizes the first release of the FANTOM5-CAGE resource.

FANTOM5-CAGE should settle many heated debates on expression pattern evolution in animal gene families and following gene duplication. In comparison to ESTs and microarrays, FANTOM5-CAGE are less susceptible to cross-hybridization between paralogs, as CAGE tags target fast diverging promoters, instead of conserved open reading frames. 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 were limited in scope to a narrow and biased 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. Due to these limitations, 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).

However, early in this debate, we suggested that Person R works best for tissue-specific genes, and that conservation of expression profiles is a complex phenomenon, most likely dependent on tissue-type and functional class of genes of interest (Huminiecki and Wolfe 2004). Data presented here strengthen those early conclusions.

In summary, herein, we combine powerful new resource of single-nucleotide resolution FANTOM5-CAGE encyclopedia defining both human and mouse transcription start sites and expression patterns, with the TreeFam database of animal gene families. FANTOM and TreeFam taken together, allow us to investigate spatial expression pattern evolution following gene duplication and in context of gene family evolution. The rationale for focusing on gene families, instead of individual genes, was to allow for epistatic and pleiotropic functions common in paralogs.

For brevity, henceforth we refer to subsets of FANTOM5-CAGE samples (FANTOM5 2013) as normal tissues: T, primary cells in culture: PC, and cancer cell lines: CCL. For example, hs-CCL refers to all human cancer cell line samples in FANTOM5-CAGE while mm-T refers to all murine tissue samples.

Results:

Initial over-view of expression patterns in the FANTOM5-CAGE atlas

Broad over-view of FANTOM5-CAGE was obtained with hierarchical clustering of human and mouse tissue samples, and principal component analysis (PCA). FANTOM5-CAGE samples clustered primarily depending on cell/tissue type of origin, not donor or developmental stage. Hierarchical clustering was performed for all different subgroups of samples in the massive FANTOM-CAGE resource (tissues, primary cell lines, and cancer cell lines), see Figure 1 for hs-T and Figure 2 for hs-CCL; as well as supplementary figures FigS1a for hs-PC and FigS1b for mm-T. Multiple donors were available for a high proportion of samples, but neither hierarchical clustering nor principal component analysis (data not shown) grouped samples by donor. Clearly, tissue-of-origin differences are more important than individual variability.

Interestingly, for both hs-T (Figure 1) and mm-T (FigS1b) brain tissues tended to cluster together, and hs-CCL shows a major divide between leukemias and solid tumors (Figure 2).

Principal component analysis revealed no global clustering pattern for either PC1/PC2, or PC2/PC3 comparisons (data not shown). Significantly, no clustering into ectoderm, mesoderm and endoderm could be observed.

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. Assignment of ortholog tissues is a pre-requisite step for the analysis of expression pattern evolution between two species. Here, we compared ortholog clusters derived through (a) the simple name matching procedure (NM-clusters), and the ortholog-based inter-species hierarchical clustering (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 (Figure 3), where ortholog genes are used to link human and mouse tissues, is an unsupervised learning approach.

We were interested whether name clusters could be recovered by unsupervised learning. Table 2 summarizes the comparison of ortholog human-mouse tissue clusters obtained by simple name matching, with those inferred using the ISHC. The full set of NM-

and ISHC-clusters is provided in TableS1. The ISHC clusters tissues according to ortholog gene expression patterns (Figure 3a). 8 NM-clusters were recovered by the ISHC, but the majority (19 clusters) were not (Table 2). The NM-clusters which were recovered through the ISHC 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 effect seems 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 rates 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 (hs, mm), and the inter-species comparison (hs-mm). These distances were calculated using bioDist package for all pairwise comparisons using the function cor.dist and the 1-|COR| formula . Inter-species distances (hs-mm) are somewhat higher than intra-species distances (hs and mm): 0.78, 0.68 and 0.72 were the respective means, suggesting that expression evolution is rapid. Even on moderate evolutionary distances, such as the human-mouse comparison, between species expression pattern differences dominated differences between organs and tissues.

Second evolutionary comparison: rates of expression pattern evolution vary widely between tissues

Tissues could be subdivided into three groups of widely differing expression pattern evolution rates, as calculated by Eucledian 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 clusters (b) and (c). This high variability in expression evolutionary rates for different tissues suggested that conservation of expression profiles is a tissue-specific phenomenon, likely to depend on the transcription factor-profile and cell-type function.

Phylogenetic timing of gene duplications using TreeFam8 database

Previously we used TreeFam6 (TF6) database (Li, Coghlan et al. 2006) to phylogenetically time gene duplication events (Huminiecki and Heldin 2010). An updated and expanded release of the database, TreeFam8 (TF8), was used here. TF8 showed the same overall distribution of duplication events as TF6 (Huminiecki and Heldin 2010), linking taxa Vertebrata and Bilateria with the two highest waves of animal gene duplications.

Recently evolved duplicates tend to be tissue-specific

Using hs-T data, 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 (Figure 5), and a previously described trend for gradual paralog expression pattern divergence (Figure 6). However, placental mammals (taxon Eutheria) were an outlier to both these trends, with an enrichment in widely housekeeping genes, and paralogs conserved in their expression profiles. Overall, paralog expression divergence in animals was fast (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).

We then asked what were the functional characteristics of genes lying at the extremes of distribution of these two broad trends. Supplementary tables TableS3a and TableS3b summarize the results of GO and PFAM enrichment for all taxa. TableS3a focuses on the fast expression divergence rate, while TableS3b relates to high breadth of expression. In both cases, the top 0.75 quantile of the distribution was used. Several observations were particularly interesting:

(a) placental mammals (taxon Eutheria**)**, highly co-expressed genes were associated with cellular macromolecular complex assembly (GO:0034622; p=3.43e-06), chromatin assembly or disassembly (GO:0006333; p=4.65e-06), nucleosome assembly (GO: 0006334; p=4.65e-06), and DNA packaging (GO:0006323; p=8.38e-06);

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

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

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

Taken together, these results suggest that histones and secreted proteins are outliers to the general trend for recently emerging genes to be narrowly expressed. Furthermore, the emergence of placental mammals (taxon Eutheria) was associated with a burst of duplications or unusual characteristics in terms their expression patterns: relatively wide (Figure 5) and conserved expression pattern (Figure 6), and enrichment in genes involved in chromatin assembly and epigenetic control. The latter hints towards the origins of rapid expression pattern evolution during evolutionary diversification of mammals, as having its source at the level of chromatin regulation and epigenetic regulatory mechanism, most likely involving histone modifications.

Differences between normal tissues and cancer cell lines: loss of normal 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 (0.8 < PC) could be seen in hs-CCL (Figure 7). This suggested important global differences in regulation of gene expression between normal tissues and cell-lines, versus malignantly transformed cancer cell lines. Significantly, the signature for recently evolved duplicates to be co-expressed could be observed in hs-PC, proving that malignant transformation not cell culture conditions lie at the roots of the effect.

We followed on these findings with more detailed examination of gene family expression in T, PC, and CCL. Across family members, average expression was calculated (supplementary Table S4 is a summary table for the full dataset show in Table S5). Data in TableS5 can be queried in multiple additional ways, depending on the biological question. For example, one can identify all families preferentially expressed in both T, and PC, versus CCL; preferentially expressed in T versus PC and CCL; narrowly expressed in T but not in CCL, etc.

Phyloexpression signatures (PhyloSigs)

A phyloexpression signatures (PhyloSig) was defined as a strong expression association between individual FANTOM5-CAGE samples, and gene duplicates derived from a given taxon. TableS4 lists top three PhyloSigs for each taxon. For example, the first row in TableS4 with taxon label “HUMAN”, shows that genes derived from human specific duplications, tend to be expressed in thymus, liver and adipose tissue. Interestingly, parotid gland (one of salivary glands) is associated with strong expression of gene duplicates dating to taxa Homo/Pan/Gorilla and Catarrhini. Finally, genes expressed in thymus, both adult and fetal, were included in PhyloSigs of many taxa, suggesting constant immune system innovation as the persistent leading theme throughout animal evolution.

We then focus on the organ systems of unique interest for vertebrate evolution: reproductive system and central nervous system. Figure 8 focuses on PhyloSigs associated with the reproductive system, while Figures 9 focuses on those for brain tissues. TableS5 lists individual genes “behind” several PhyloSigs from TableS4 and Figure 8, for several chosen sample+taxon PhyloSigs: seminal vesicle+Catarrhini, testis +Eutheria, uterus+Eutheria, adipose+human, vagina+human, salivary glad+Catarrhini.

Specific family example

Rac/cdc42 - gene family with complex and dynamic expression pattern evolution

The Rac/cdc42 subfamily of Rho GTPases (TF101109) illustrates mutually exclusive expression of paralogs - MEEP (Figure 10). What is being meant by the MEEP? Let us examine expression of RAC2, RAC3 and RHOG in detail. RAC3 was highly expressed in five fetal tissues from which RHOG, and RAC2 were excluded. The five tissues in question were: fetal parietal lobe, fetal temporal lobe, fetal duodenum, fetal occipital lobe, and fetal brain pool. On the other hand, RHOG 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: thymus adult, blood adult, tonsil adult, appendix adult, thymus fetal, vein adult, spleen adult, lymph node adult, spleen fetal), expressed RHOG and RAC2, but not RAC3. Although the MEEP may at first suggest expression subfunctionalization, the next paragraph shows that neofunctionalization was more likely in the cdc42 family.

Did these expression patterns reflect the phylogenetic history of the Rac/cdc42 family? Not directly. The Rac/cdc42 TreeFam family tree (TF101109), showed that RAC2 and RAC3 were 2R-ohnologs deriving from 2R-WGD (Huminiecki and Heldin 2010). In contrast, divergence between RAC2/3 ancestor and RHOG predated the origin of animals. What is then the most parsimonious scenario for the evolution of MEEP in the Rac/cdc42 family? We suspect expression pattern neofunctionalization following RAC2/3 duplication. In this scenario, RAC2 retained most of ancestral expression characteristics (i.e. those shared with RHOG), while RAC3 acquired new expression sites in fetal brain, loosing expression in other tissues. In discussion, we examine promoter structures of relevant genes using ENCODE data, and attempt to identify transcription factors involved.

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 the FANTOM5-CAGE single-base level encyclopedia of transcription start sites and expression patterns. Clustering and corresponding heatmap computed for human tissues, shown in Figure 1, demonstrated 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 2, FigS1a, and FigS1b show heatmaps for hs-CCL, hs-PC, and mm-T respectively. FigureS1b shows that mouse brain samples also clustered, with subclusters formed by visual cortex and cerebellum. Finally, a major divide between human blood malignancies and solid tumours can be seen in Figure 2.

In addition to hierarchical clustering, we used principal component analysis (PCA) but PCA could not significantly reduce the dimensions of the data. This underlined intrinsic variability of expression patterns, and argued against the existence of global modules of transcriptional regulation common to sets of tissues, for example for tissues derived from three distinct embryonic layers.

Assignment of human-mouse ortholog tissues

In the next step, we compared human and mouse tissues together using hierarchical clustering, calculating distances between tissues from different species by comparing expression values for ortholog genes. The goal was to match similar human and mouse samples and thus to assign ortholog tissues (Figure 3a). However, when this approach was tested most samples grouped by species not tissue type (Figure 3a). For example, most 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 (Figure 3a).

Previous studies comparing evolutionary profiles between species assumed a rather simple method for matching tissues: automatically assuming that samples with matching names in different species are ortholog tissues. This may be an over-simplification of a complex problem. Name clusters were not easily recovered by unsupervised learning approaches, such as the ISHC (compare Figure 3 with Table 2).

What were the reasons? Firstly, as demonstrated by rapid 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 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. More data is needed to investigate this problem. Future releases of FANTOM database will include tissues from additional vertebrate species (rat, dog and chicken), and inclusion of these samples into the comparison should give a fuller picture.

Comparison of normal and cancer cells

When three broad human FANTOM5-CAGE sample subtypes were compared (tissues - T, primary cells - PC, and cancer cell-lines - CCL), we found that the peak of highly co-expressed paralogs (top quartile of PC distribution) was missing in hs-CCL, in contrast to hs-T and hs-PC (Figure 7). In other words, while recent closely related paralogs tended to be expressed in the same T samples, they were not co-expressed in the same CCL samples. This suggested that normal conditions of promoter regulation did not exist in CCL. Paralogs from the top quartile of PC distribution tended to be recent gene duplicates (the majority not older than taxon “Eutheria”: Figure 4a), most likely with similar promoters and in tandem arrangement on the genome. Clearly, while these genes were co-regulated in normal tissues, they lost this regulation in cancer. The effect cannot be attributed to cell culture conditions alone, as paralog co-expression signature was seen in PC samples. We propose a novel term: devolution of expression patterns, to suggest that normal evolutionary constraints on expression patterns do not exist in cancer. Expression pattern devolution is indicative of global distortion of cancer expression space, perhaps through the loss of normal epigenetic regulatory mechanisms.

To investigate differences between normal tissues and cancer samples further, we looked at gene families with differential average expression in T, versus PC and CCL (TableS2). It makes sense that most of the top CCL 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 (CDCA7). These differences could be interpreted as simple consequence of malignant transformation and uncontrolled proliferation. In contrast, gene families highly expressed in tissue samples versus both primary and cancer cell lines, appeared to be mostly involved in cell adhesion, for example: PRKA3/4 described previously in the context of sperm-oocyte interaction, negative regulator of Hedgehog signaling pathway - patched, tumor suppressor APC (regulating cell attachment). Taken together, these results help 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/cdc42 family and mutually exclusive expression of paralogs (MEEP)

Expression patterns associated with the Rac/cdc42 family were investigated in Figure 10. Here, we ask whether dramatic spatial expression domain shifts, and mutually exclusive expression of paralogs (MEEP) seen in this family, can be correlated with promoter structure, revealed using ENCODE data (Table 3). Table 3 lists all transcription factor binding sites (Tfbs) associated with transcription start sites (TSSes) of these three genes, while Table 4 calculates JI for pairwise promoter comparisons (visualized in Figure 10b). JI has bimodal distribution with peaks in two intervals: 0-0.2 and 0.2-0.5 (Figure 10b-key and Table 4).

We have also examined promoter regions of these three genes manually with the UCSC-ENCODE and ZENBU genome browsers (FigS2). Narrowly expressed, RAC3 features one strong TF binding site: ZBTB7A, a weak and narrow DNASEI site, and a weak H3K27AC signal. In contrast, liberally expressed RAC2 and RHOG have broad and high-scoring DNASEI sites, and very strong H3k27AC signals. RAC2 and RHOG have many transcription factor binding sites (Tfbs) within narrow 500 bps window of their transcription start sites (TSSes), with strong NFKB signature for both RAC2, and RHOG. Taken together, these results suggested a scenario where ancestral NFKB binding site was replaced with ZBTB7A in RAC3.

Two lines of evidence in published literature support the evolutionary scenario correlating replacement of NFKB with ZBTB7A, with a shift from broad expression preferentially associated with immune and circulatory systems (RAC2 and RHOG), to narrow expression domain in fetal brain (RAC3):

(a) NFKB (NFKB) is widely expressed in animal cells and well-known to play a role in immune and stress responses (Karin 2006).

(b) ZBTB7A or zinc finger and BTB domain containing 7A, had been shown to be an oncogenic transcription factor (Maeda, Hobbs et al. 2005; Maeda, Hobbs et al. 2005), and implicated in glioma (Rovin and Winn 2005). Interestingly, ZBTB7A promoter region itself features three ZBTB7A binding sites, suggesting an autoregulatory positive feedback loop (UCSC genome browser on hg19: chr19:4,066,215-4,068,615).

Overall, there is substantial correlation between co-expression (Figure 10a) and JI (Figure 10b), although strong binding Tfbs may be more informative, and some TFs, such as ZBTB7A, may be more important than others in determining gene expression, and facilitating MEEP. This analysis is now being extended to global scale, with Jaccard index calculations for all TreeFam families.

A sideline to the examination of the Rac/cdc42 family, is that the phyloexpression tree (i.e. the gene tree derived from FANTOM5-CAGE expression clustering), has little similarity to the TreeFam phylogenetic tree (TF101109), suggesting that expression patterns cannot be readily used as phenotypic markers for phylogenetic inference. Examination of several other families supports the conclusion of limited similarity between expression-derived dendrograms and phylogenetic trees (data not shown). Limited phylogenetic signal contained in expression patterns is also evident in Figure 6: high PC variability for younger taxa, and plateau in expression divergence reached as soon as taxon Amniota.

Phylogenetic signatures (PhyloSigs)

Our PhyloSigs, i.e. associations of taxa of duplication with expression in certain FANTOM5-CAGE samples (TableS4 and S5), provide rich material for formulation of hypotheses on animal evolution. For example, cystatins (Baron, DeCarlo et al. 1999) and proline-rich salivary proteins (Amado, Lobo et al. 2010) known to be present in saliva, are part of the [parotid gland+Catarrhini] PhyloSig, suggesting adaptation to novel food sources.

Histone proteins are involved in several PhyloSigs (Table S5). For example, histone cluster 1 H3 and H4, are linked with [adipose+human], [parietal_lobe+human], [putamen +human], [testis+Eutheria], suggesting that modulation of epigenetic control of gene expression was frequently at the root of animal evolutionary novelties.

Semenogelin, the predominant protein in human semen (Lilja, Abrahamsson et al. 1989), isoforms I and II are two top proteins associated with the [seminal vesicle +Catarrhini] PhyloSig. This PhyloSig derives from SEMG1/SEMG2 gene duplication (TF342360). Interestingly, SEMG1/SEMG2 form a tight cluster on chr20, flanked on both sides by putative CTCF insulator sites, yet adjacent genes both down-stream (PI3) and up-stream (SLPI) are also co-expressed in semen (UCSC genome browser on hg19,chr20: 43,793,276-43,902,697).

[Homo/Pan/Gorilla+placenta] PhyloSig involves a number of pregnancy specific glycoproteins (PSGs). Interestingly, these proteins derive from the same TF8 gene family (TF336859), and form a tight genomic cluster with several overlapping CTCF sites (UCSC genome browser on hg19, chr19:43,172,185-43,493,834).

Semenogelins and PSGs are examples of two broader genomic trends: the trend for recently evolved genes to be specific in their expression domain (Figure 5), and for the potential role of CTCF in coordinating such domain-specific expression in clustered loci. Although CTCF was originally described as an insulator, recent data suggest that it may have a dual role, functioning either as an insulator, or in facilitating looping which brings promoter regions under the control of an enhancer (Phillips and Corces 2009).

Data access:

Access to FANTOM5-CAGE 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-CAGE is arguably the most complete functional genomics dataset generated to date (FANTOM5 2013). FANTOM5-CAGE includes 952 human and 396 mouse tissues (T), primary cells (PC) and cancer cell-lines (CCL).

To produce gene expression tables, RefSeq transcripts were linked with all CAGE tags +/- 500 bps from the RefSeq's TSS.

Linking TreeFam8 with FANTOM5

TF8 trees were linked with FANTOM5-CAGE 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, hierarchical clustering, and PEM

Expression distances were calculated using Bioconductor package bioDist Release (2.11). Pearson correlation (PC) was used for Figure 6 and 7 as it works well for tissue-specific genes. Between-paralog Euclidean distances were used for Figure 4, where within-tissue expression divergence rates are measured. Spearman distances were used for T, PC, and CCL heatmaps (Figure 1 and 2, FigS1a,b), as this type of correlation is rank-based and less affected by strong outliers.

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 (PC) was used as the expression distance measure.
b) Name clusters (NM-clusters) were identified using a custom Python script.
c) In our hands complex ontologies such as open biomedical ontologies (OBO) UBERON classification (Mungall, Torniai et al. 2012) have not worked well for the task of ortholog tissue assignment, because their hierarchical, denormalized, and multilayerd design is better suited for browsing than classification.

Jaccard index calculations

Jaccard index (JI), i.e. ratio of intersection over the union, is used here as a measure of similarity between promoters of the cdc42 family in terms of their transcription factor binding (Tfbs) profiles.

ENCODE data

Multi cell-line, clustered ENCODE Tfbs data (ENCODE_UCSC_Tfbs_V2), called from henceforth TfbsClusteredV2, published by the ENCODE consortium in 2012 (ENCODE), 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 the GRCh37/hg19 assembly, which includes tracks with subset of ENCODE data from TfbsClusteredV2 .

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:
We would like to acknowledge the FANTOM5 consortium.

Author contributions:

OS and LH designed the study, performed all analyses, and wrote the manuscript. 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. Hierarchical clustering of hs-T.
Central nervous system samples cluster separately from all other tissues, with two distinct sub-clusters: (b) medulla oblongata, spinal cord, brain glands, eye; (c) other brain tissues. Several additional clusters are clearly visible, and have been annotated (d-i). Spearman correlation was used as the expression distance measure.

Figure 2. Hierarchical clustering of hs-CCL samples (Spearman correlation).

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 are clustered according to expression profiles of orthologs. Pearson Correlation (PC) is used as the expression distance measure. Panel (**b**) shows a stacked histogram with PC distributions (all-against-all sample comparisons) for hs (human), mm (mouse), and hs-mm (human-mouse), visualized using distinct colors.

Figure 4. 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).

Figure 5. Recently evolved are tissue-specific, old genes are housekeeping.
An R-style box-and-whisker plot of EB for ordered taxa is shown. 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. However, there are exceptions to the trend: taxa Eutheria and Deuterostomia.

Figure 6. **Expression pattern divergence bar-plot between paralogs.**

Figure 7. Paralog expression divergence in hs-T versus hs-CCL.
Distribution of paralog expression distances (PC) offers unexpected evidence for global transcriptional disregulation in hs-CCL. Panel (a) hs-T, panel (b) hs-CCL. No peak for paralogs with high PC can be seen in hs-T.

Figure 8. Evolutionary history of mammalian tissues: reproductive system.
Key to the heatmap shows color-codes and associated frequencies. Dark color bands on the heatmap correspond to PhyloSigs, i.e. strong [taxon+tissue] associations. Each band corresponds to average TPM expression values for all duplicates associated with a given taxon in TF8.

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

Figure 10. Evolution of expression patterns in the cdc42 family.
Panel (a) shows heatmap of expression patterns for the cdc42 family in hs-T. Panel (b) shows heatmap of the values for Jaccard-index (JI) for ENCODE Tfbs in pairwise comparisons. Table 4 shows actual values of the JI.

Rac 2 and 3 are 2R-ohnologs. RHOJ and RHOQ are very divergent family members, with weak expression in human tissues. RAC1 and cdc42 are also highly diverged, but similar in their expression pattern. Neither the expression distance nor the JI-based dendrograms directly reflect the phylogenetic history of the family.

Please, see Table 4 for actual values of Jaccard index in the cdc42 family.

Supplementary figures:

FigS1a. Hierarchical clustering of hs-PC samples (Spearman correlation).

FigS1b. Hierarchical clustering of mm-T samples (Spearman correlation).

FigS1c. Dendrogram of human brain samples.

FigS2. RAC2, RAC3 and RHOG in the UCSC-ENCODE and ZENBU genome browsers.

TABLES:

Table 1. The structure of the F5 encyclopedia of expression patterns.
Numbers of samples in different sample subsets are listed (T, PC, and CCL), in human and mouse.

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. 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 3. cdc42 family ENCODE Tfbs in a 500 bp window (-/+ 250 bps from the TSS).
Table 3 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 4. Tfbs Jaccard index for the cdc42 family ENCODE Tfbs
Jaccard index (JI, intersection over the union) values for pairwise comparisons between Rac/cdc42 family members. High JI indicates similar Tfbs profile between paralog promoters. RHOJ and RAC3 have lowest sums of JI values (1.54 and 1.57, respectively), while other members of the family have JI between 2.24 and 2.49.

Supplementary tables:

TableS1. The full FANTOM5-CAGE clustering dataset is shown: (a) NM-dataset; (b) ISHC clusters.

TableS2. Families with differential average expression in T versus PC and CCL.

TableS3a. GO term and PFAM domain enrichment, for all taxa, in genes with unusually fast expression divergence rate.

TableS3b. GO term and PFAM domain enrichment, for all taxa, in genes with unusually high breadth of expression.

TableS4. Strongest associations between timing of gene duplication and expression site in FANTOM5-CAGE.

TableS5. Example genes behind strongest associations between timing of gene duplication and expression site in FANTOM5-CAGE.

TableS2 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 1. Number of samples in distinct FANTOM5-CAGE categories for human and mouse: T, PC, CCL.

	human (hs)	mouse (mm)
total	952	396
tissues (T)	179	280
primary cells (PC)	513	116
cancer cell lines (CCL)	260	-
brain tissues (subset of T)	60	51
reproductive tissues (subset of T)	14	21

Table 2. Comparison of ortholog human-mouse (hs-mm) tissue (T) 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 3. 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***						
RHOG	**0.24**	***1***					
RHOJ	0.07	0.1	***1***				
RAC1	**0.28**	**0.35**	0.1	***1***			
RAC2	**0.25**	**0.31**	0.09	**0.25**	***1***		
RAC3	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. Jaccard index (JI) for the cdc42 family
JI between identical promoters is 1. The table is symmetrical. Values higher than 0.2 are shown in bold and using larger font. Please, see Figure 10b for clustering and heatmap representation of the table.

TableS2. Families with differential average expression in T versus PC and CCL. 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 hs-CCL, low in hs-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 hs-T, low in hs-PC and hs-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 hs-PC and hs-CCL, low in hs-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 hs-T, low in hs-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

TableS4. FANTOM5-CAGE phyloexpression signatures
Strongest associations between timing of gene duplication and expression site in FANTOM5-CAGE. Average expression values, in TPM, for human genes linked with all duplication events times to a given taxon, by phylogenetic timing, are given.

Taxon	Tissue short name	Full F5 sample name (with CN-id code)	Expression
HUMAN	thymus liver adipose	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19	245.55 113.83 108.23
Homo/Pan/ Gorilla	parotid gland salivary gland blood	parotid gland, adult.CNhs12849.10199.103F1.hg19 salivary gland, adult, pool1.CNhs11677.10093.102C3.hg19 blood, adult, pool1.CNhs11761.10053.101G8.hg19	334.94 324.45 320.09
Catarrhini	parotid gland liver trachea	parotid gland, adult.CNhs12849.10199.103F1.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19 trachea, adult, pool1.CNhs10635.10029.101E2.hg19	243.46 124.30 70.86
Eutheria	thymus liver thymus	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19 thymus, fetal, pool1.CNhs10650.10043.101F7.hg19	566.32 225.78 208.24
Theria	blood liver thymus	blood, adult, pool1.CNhs11761.10053.101G8.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19 thymus, adult, pool1.CNhs10633.10027.101D9.hg19	139.02 134.02 104.06
Amniota	thymus thymus breast	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 thymus, fetal, pool1.CNhs10650.10043.101F7.hg19 breast, adult, donor1.CNhs11792.10080.102A8.hg19	116.93 50.29 45.65
Tetrapoda	thymus esophagus pancreas	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 esophagus, adult, pool1.CNhs10620.10015.101C6.hg19 pancreas, adult, donor1.CNhs11756.10049.101G4.hg19	185.17 149.77 139.67
Euteleostomi	thymus adipose thymus	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 thymus, fetal, pool1.CNhs10650.10043.101F7.hg19	78.64 48.70 42.70
Chordata	pancreas skeletal muscle artery	pancreas, adult, donor1.CNhs11756.10049.101G4.hg19 skeletal muscle, adult, pool1.CNhs10629.10023.101D5.hg19 artery, adult.CNhs12843.10190.103E1.hg19	114.33 74.92 58.70
Deuterostomia	pancreas placenta dura mater	pancreas, adult, donor1.CNhs11756.10049.101G4.hg19 placenta, adult, pool1.CNhs10627.10021.101D3.hg19 dura mater, adult, donor1.CNhs10648.10041.101F5.hg19	50.02 31.04 28.25
Bilateria	thymus adipose pancreas	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 pancreas, adult, donor1.CNhs11756.10049.101G4.hg19	108.48 73.82 56
Eumetazoa	thymus thymus liver	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 thymus, fetal, pool1.CNhs10650.10043.101F7.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19	117.86 51.50 50.88
Fungi/Metazoa	thymus adipose adipose	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 adipose, donor3.CNhs13974.10186.103D6.hg19	450.2 232.43 177.11
Metazoa	skeletal muscle artery pancreas	skeletal muscle, adult, pool1.CNhs10629.10023.101D5.hg19 artery, adult.CNhs12843.10190.103E1.hg19 pancreas, adult, donor1.CNhs11756.10049.101G4.hg19	42.76 40.91 37.04
Eukaryota	thymus skeletal muscle vagina	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 skeletal muscle, adult, pool1.CNhs10629.10023.101D5.hg19 vagina, adult.CNhs12854.10204.103F6.hg19	100.76 75.12 72.49

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.

FANTOM5 (2013). A promoter level mammalian expression atlas

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

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.

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.

NFKB. "NFKB." from http://www.factorbook.org/mediawiki/index.php/NFKB.

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.

Phillips, J. E. and V. G. Corces (2009). "CTCF: master weaver of the genome." Cell **137**(7): 1194-1211.

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.

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.

TF101109. "TF101109." from (http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF101109.

TF336859. "TF336859." from http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF336859.

TF342360. "TF342360." from http://www.treefam.org/cgi-bin/TFinfo.pl?ac=TF342360.

Figure 1

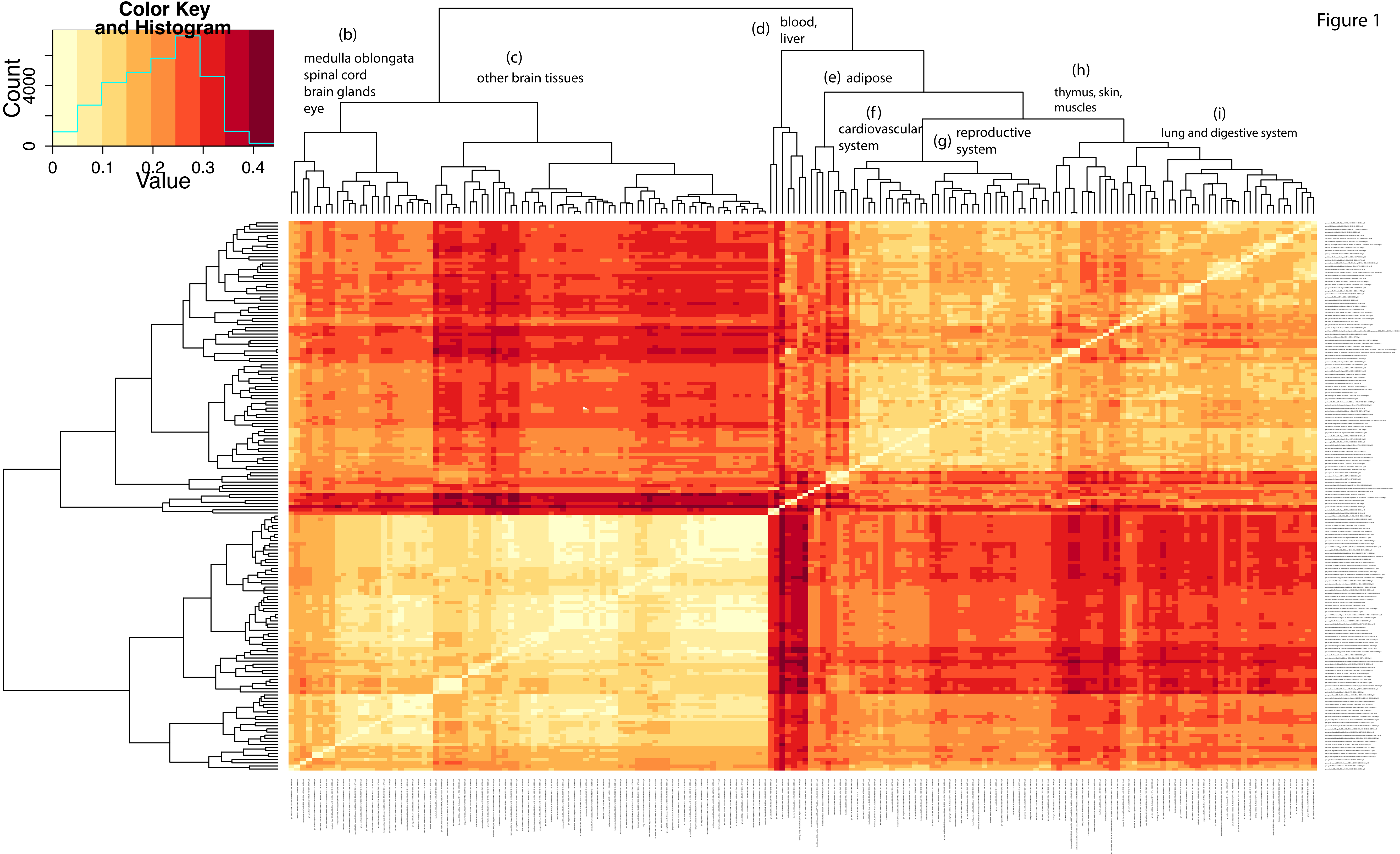
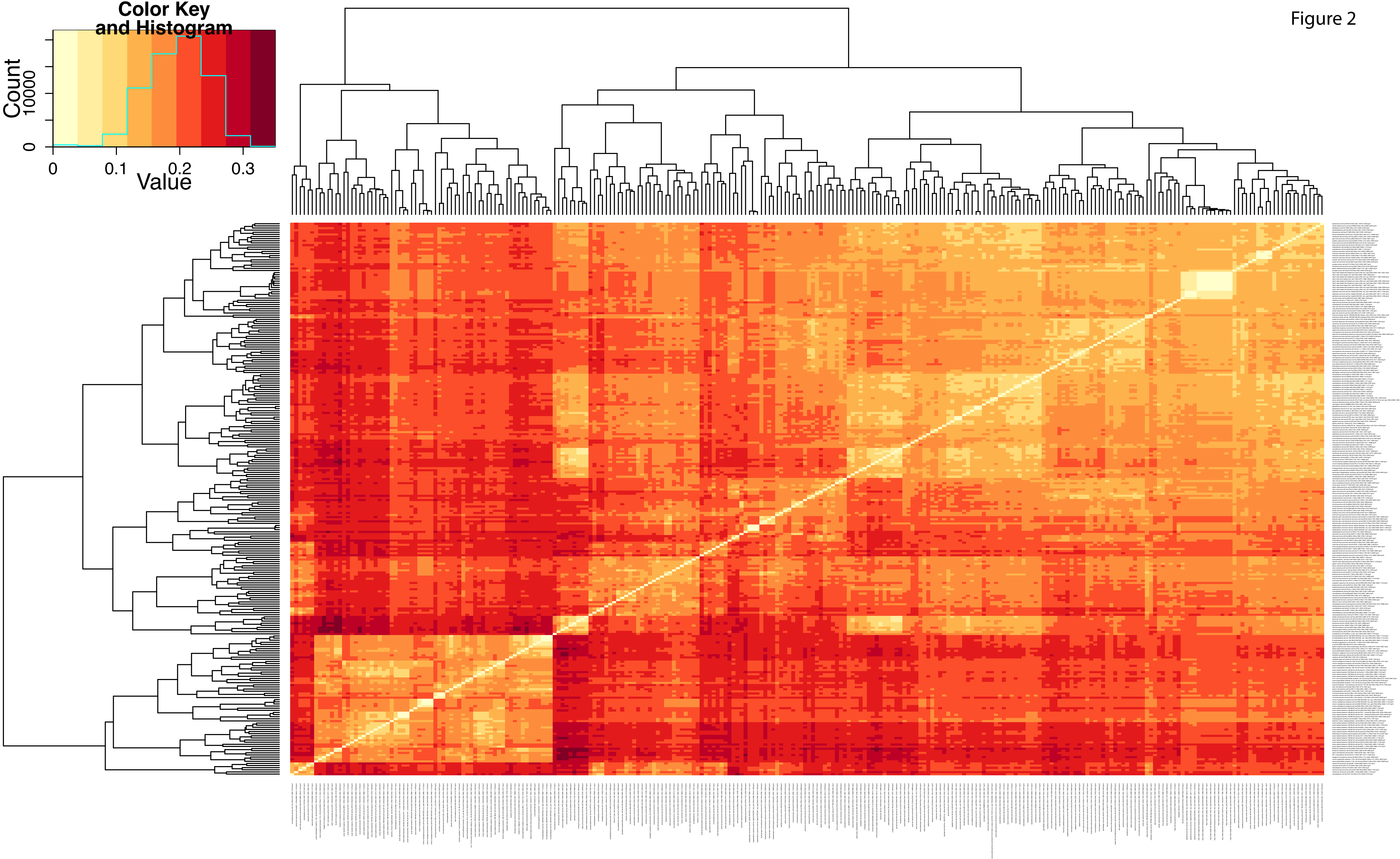

Color Key
and Histogram
Count
0
4000
Value
0
0.1
0.2
0.3
0.4
(b)
medulla oblongata
spinal cord
brain glands
eye
(c)
other brain tissues
(d)
blood,
liver
(e) adipose
(f)
cardiovascular
system
(g)
reproductive
system
(h)
thymus, skin,
muscles
(i)
lung and digestive system

Figure 2



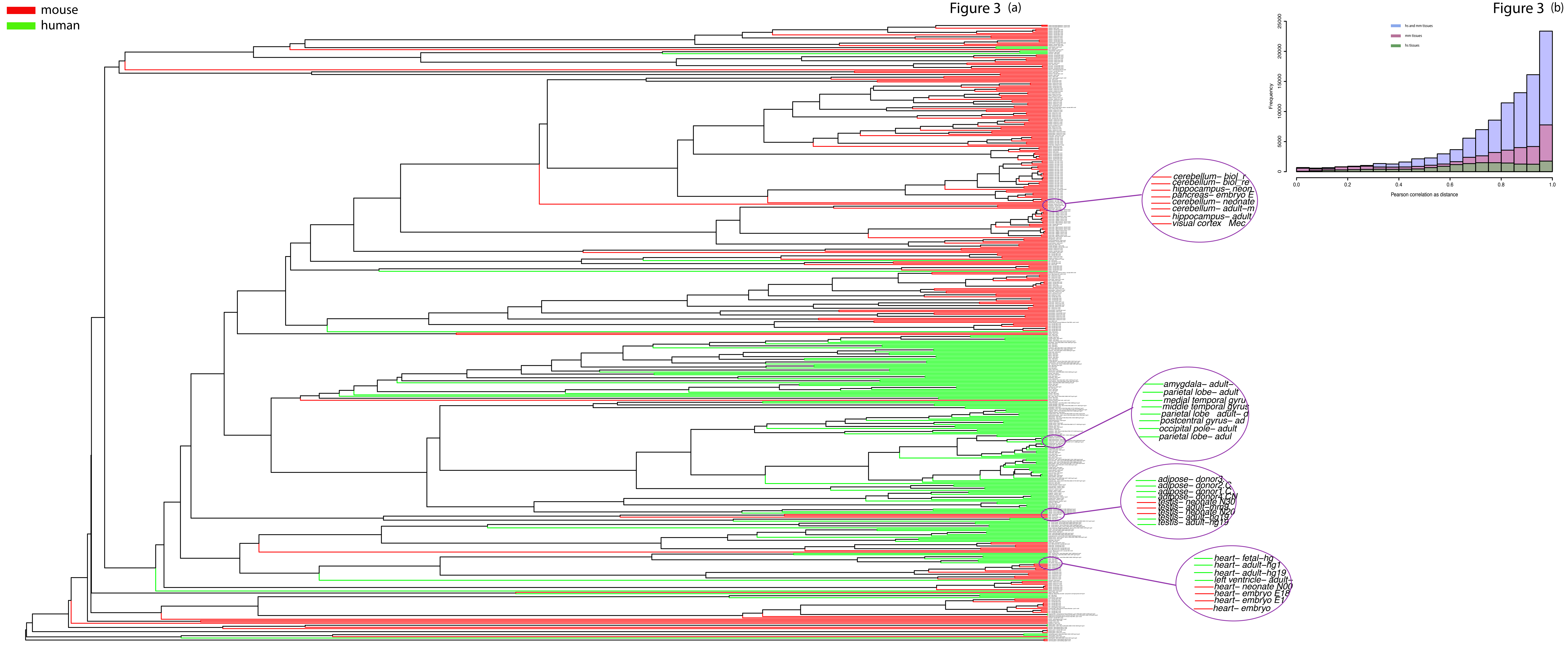


Figure 3 (a)

Figure 3 (b)

Figure 4

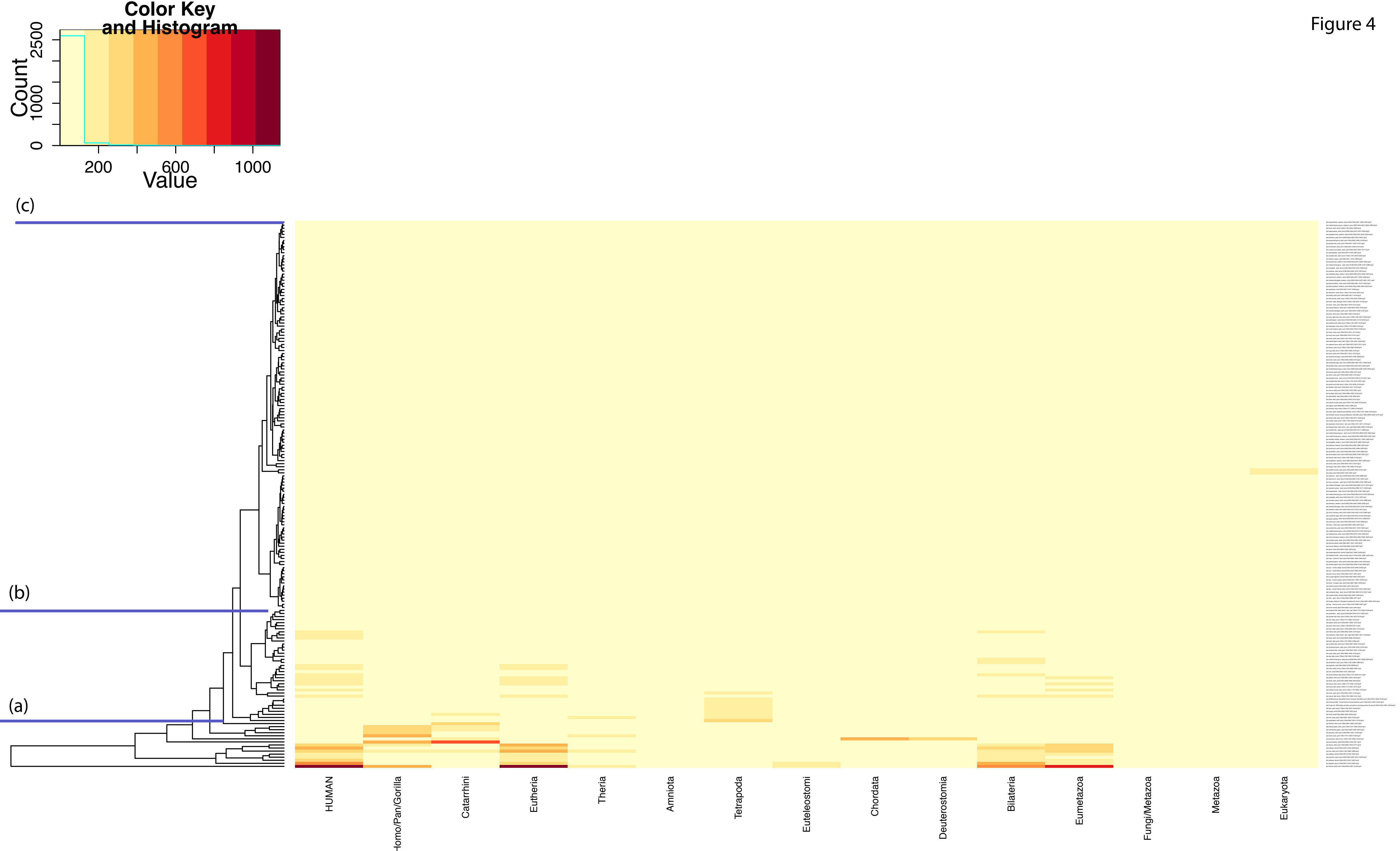

Color Key
and Histogram
Count
0
1000
2500
200
600
1000
Value
(c)
(b)
(a)
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure 5

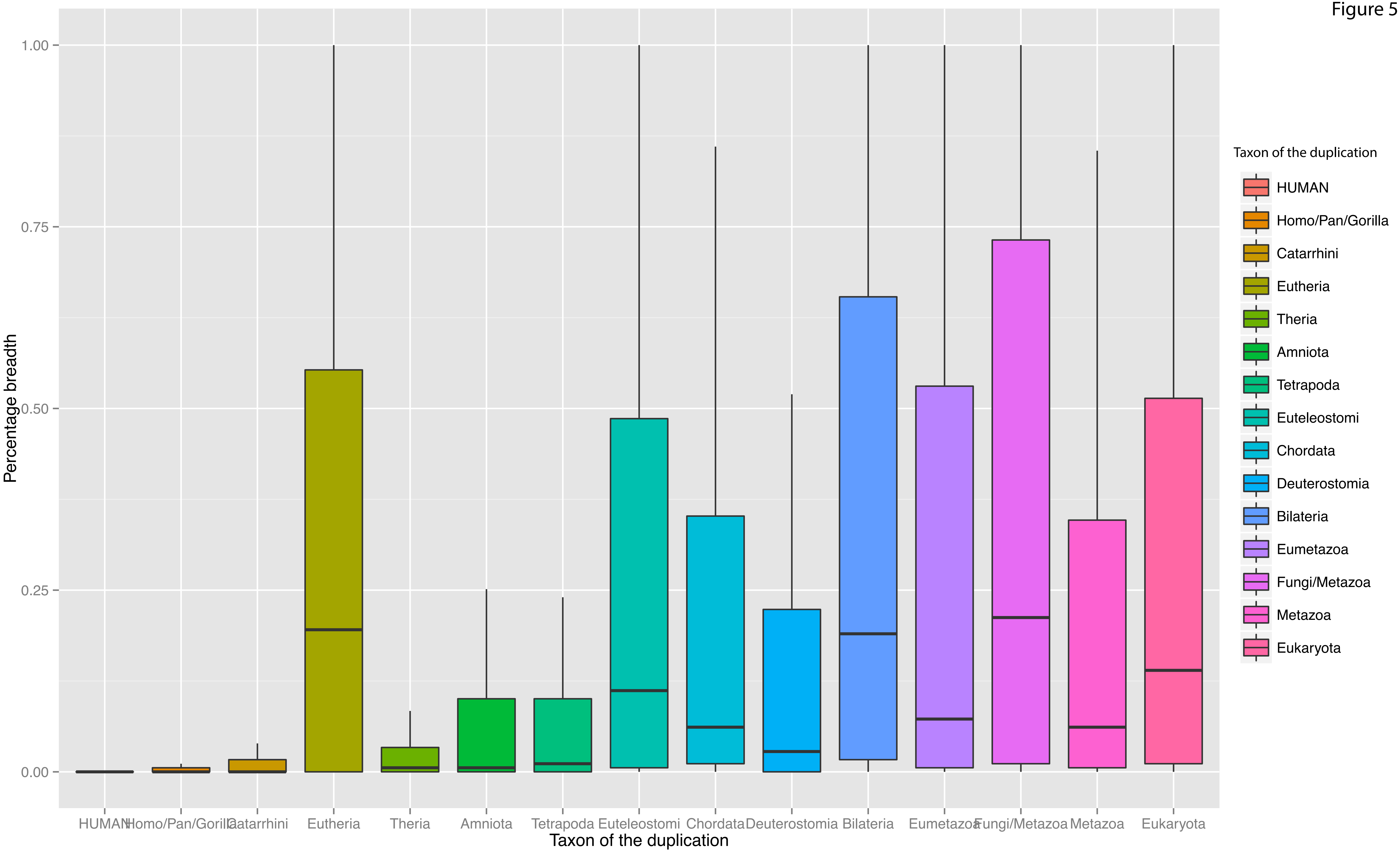
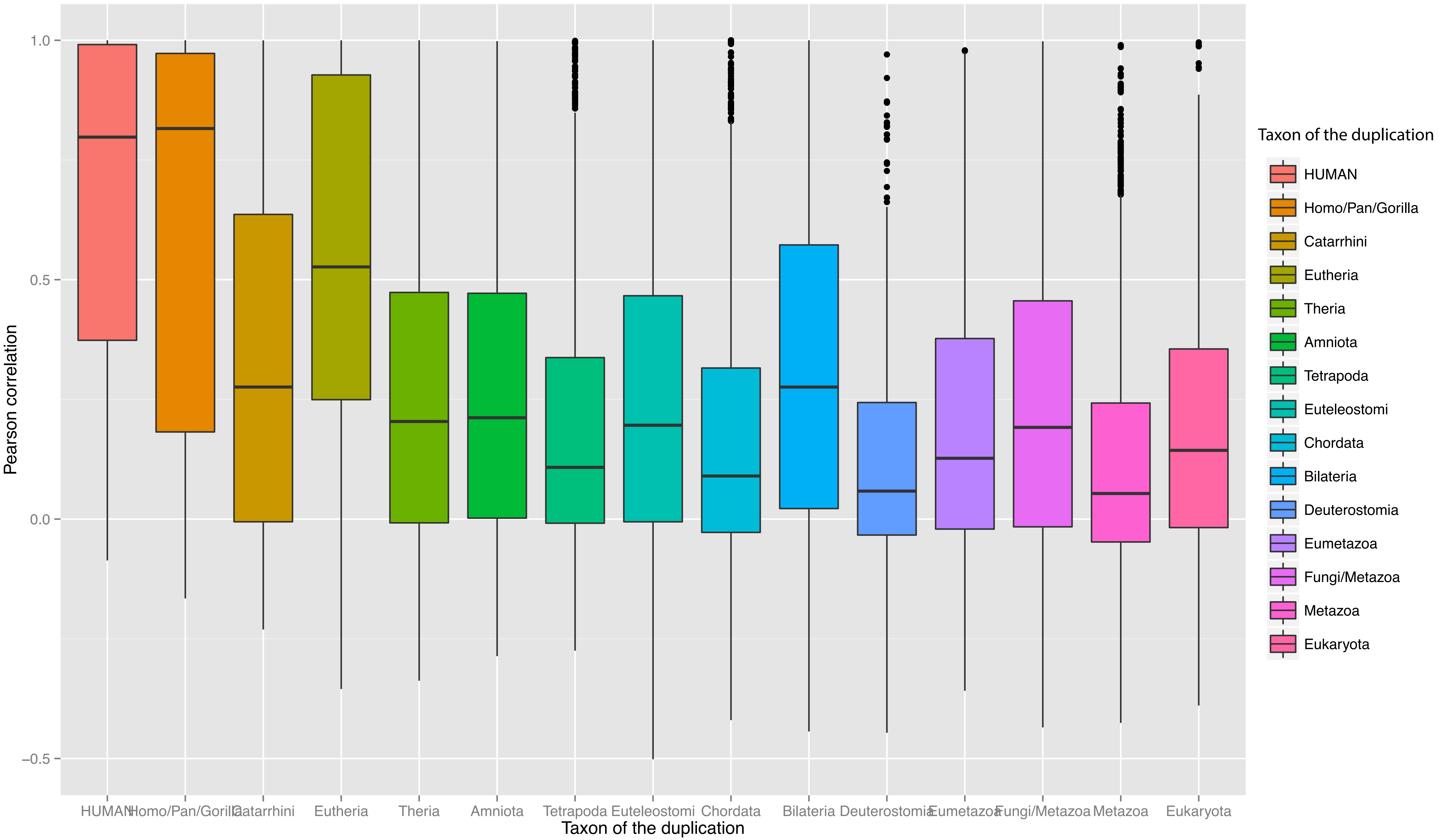

Percentage breadth
1.00
0.75
0.50
0.25
0.00
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
Taxon of the duplication
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota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

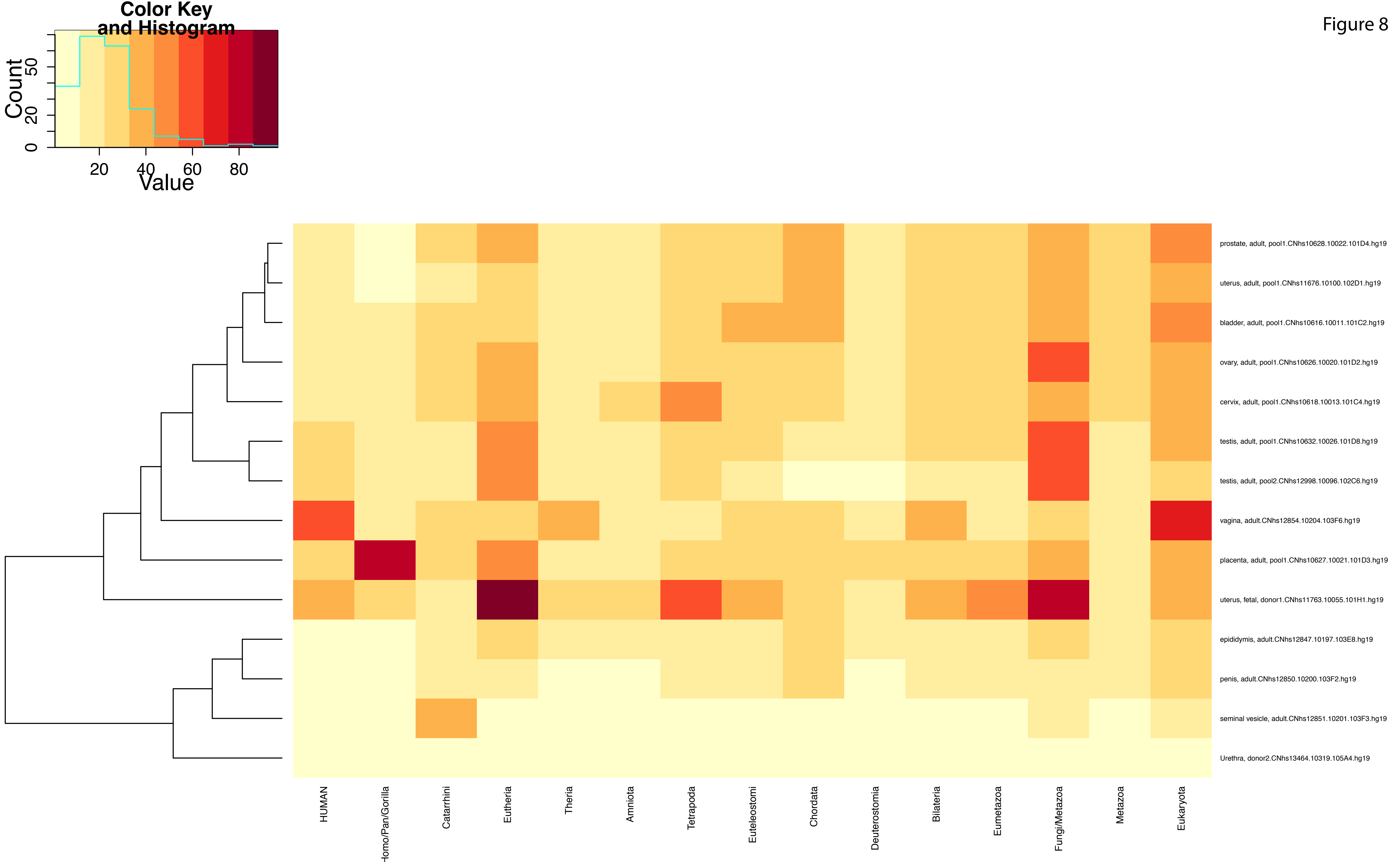

Color Key
and Histogram
Count
0
20
50
20
40
60
80
Value
prostate, adult, pool1.CNhs10628.10022.101D4.hg19
uterus, adult, pool1.CNhs11676.10100.102D1.hg19
bladder, adult, pool1.CNhs10616.10011.101C2.hg19
ovary, adult, pool1.CNhs10626.10020.101D2.hg19
cervix, adult, pool1.CNhs10618.10013.101C4.hg19
testis, 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

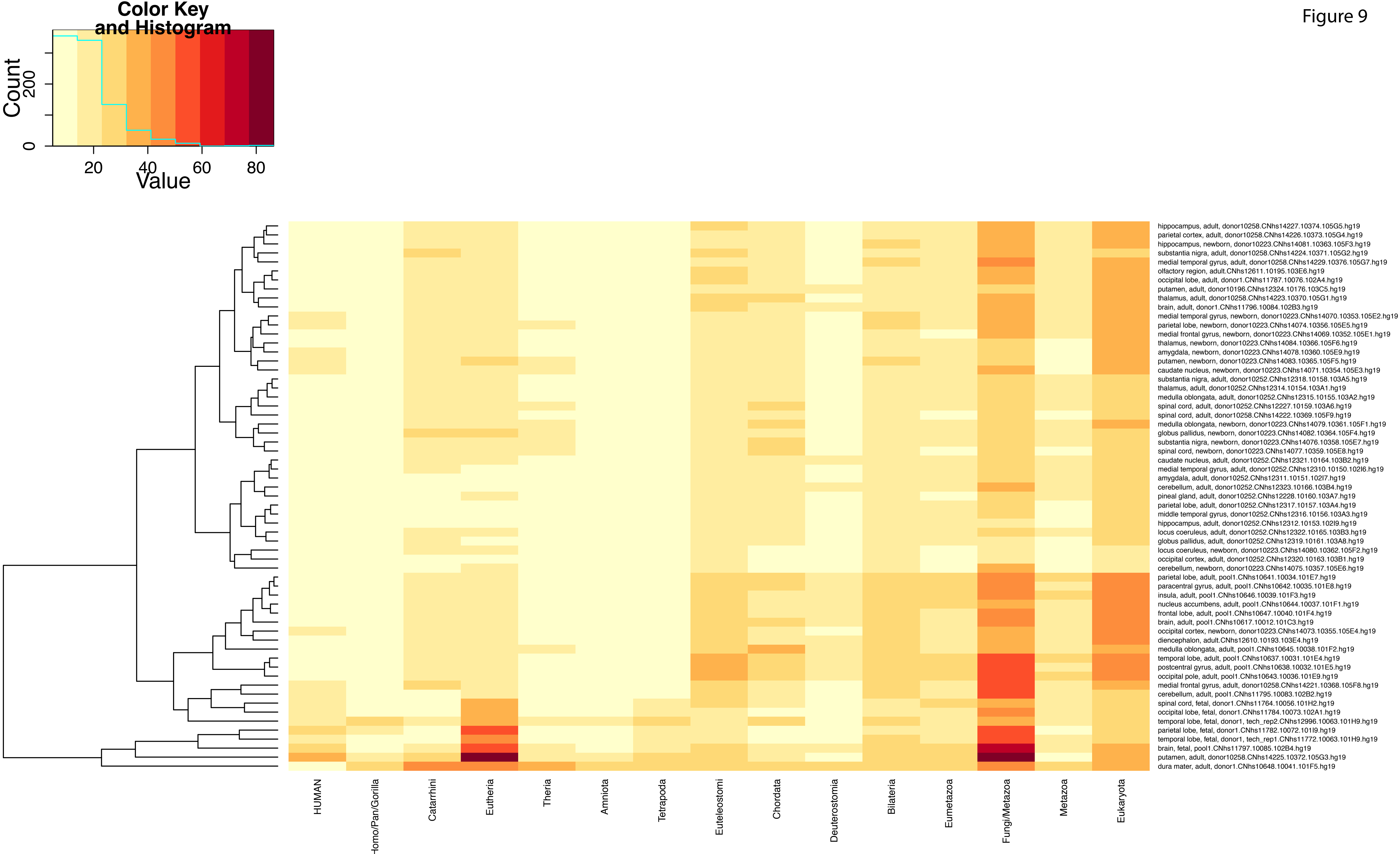

Color Key
and Histogram
Count
0
200
20
40
60
80
Value
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
hippocampus, newborn, donor10223.CNhs14081.10363.105F3.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
medial 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

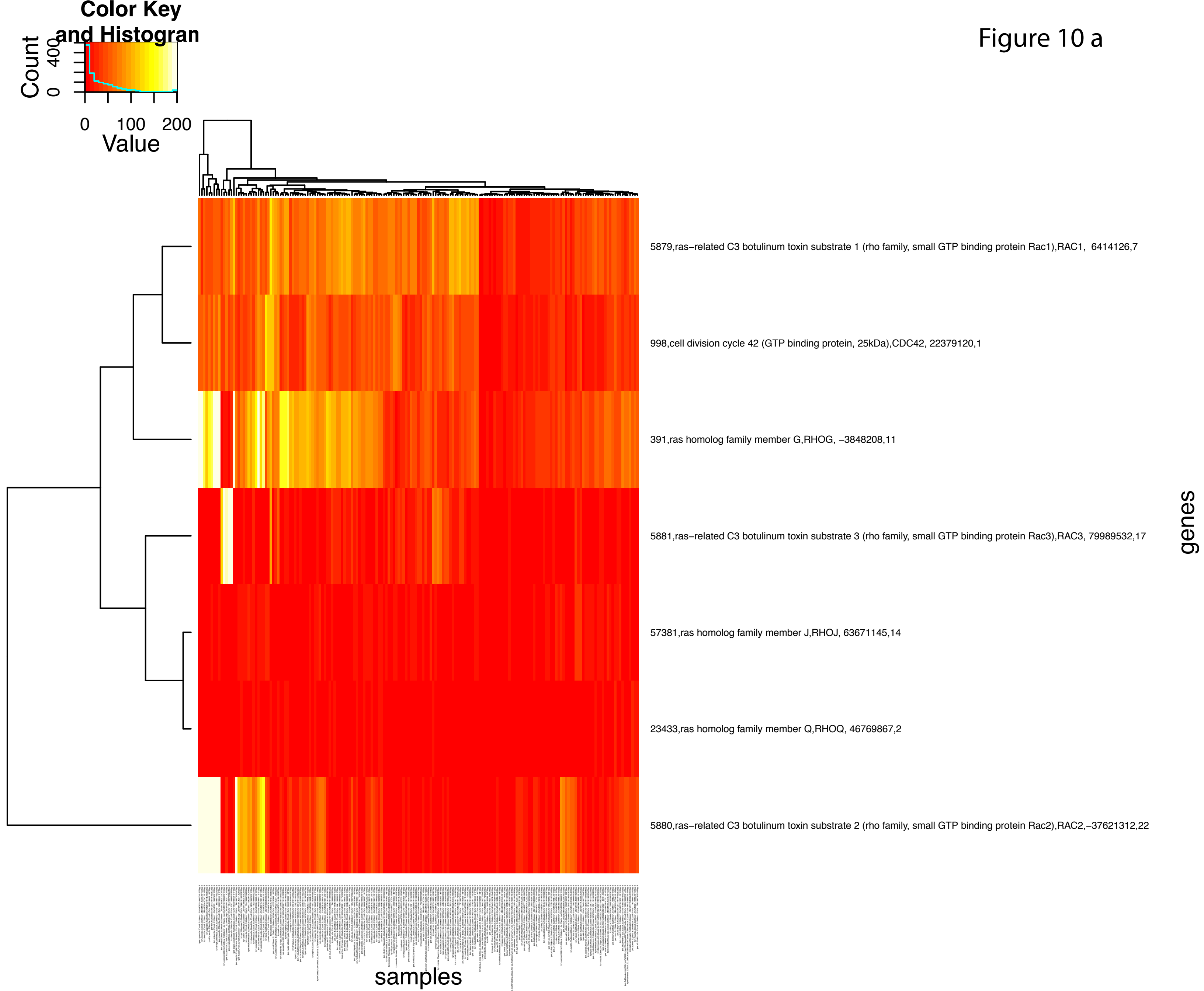


Figure 10b

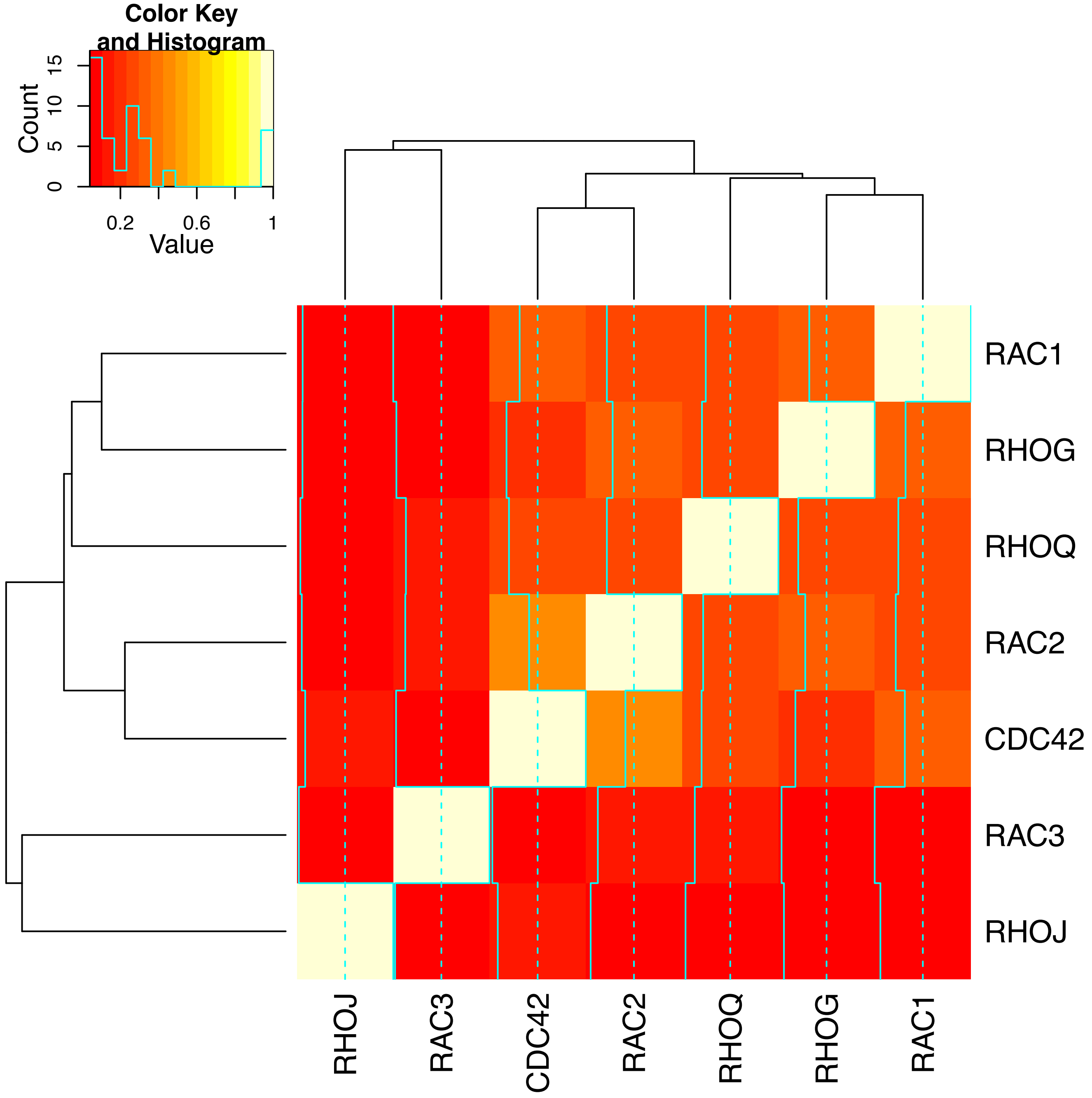


Figure S1 (a)

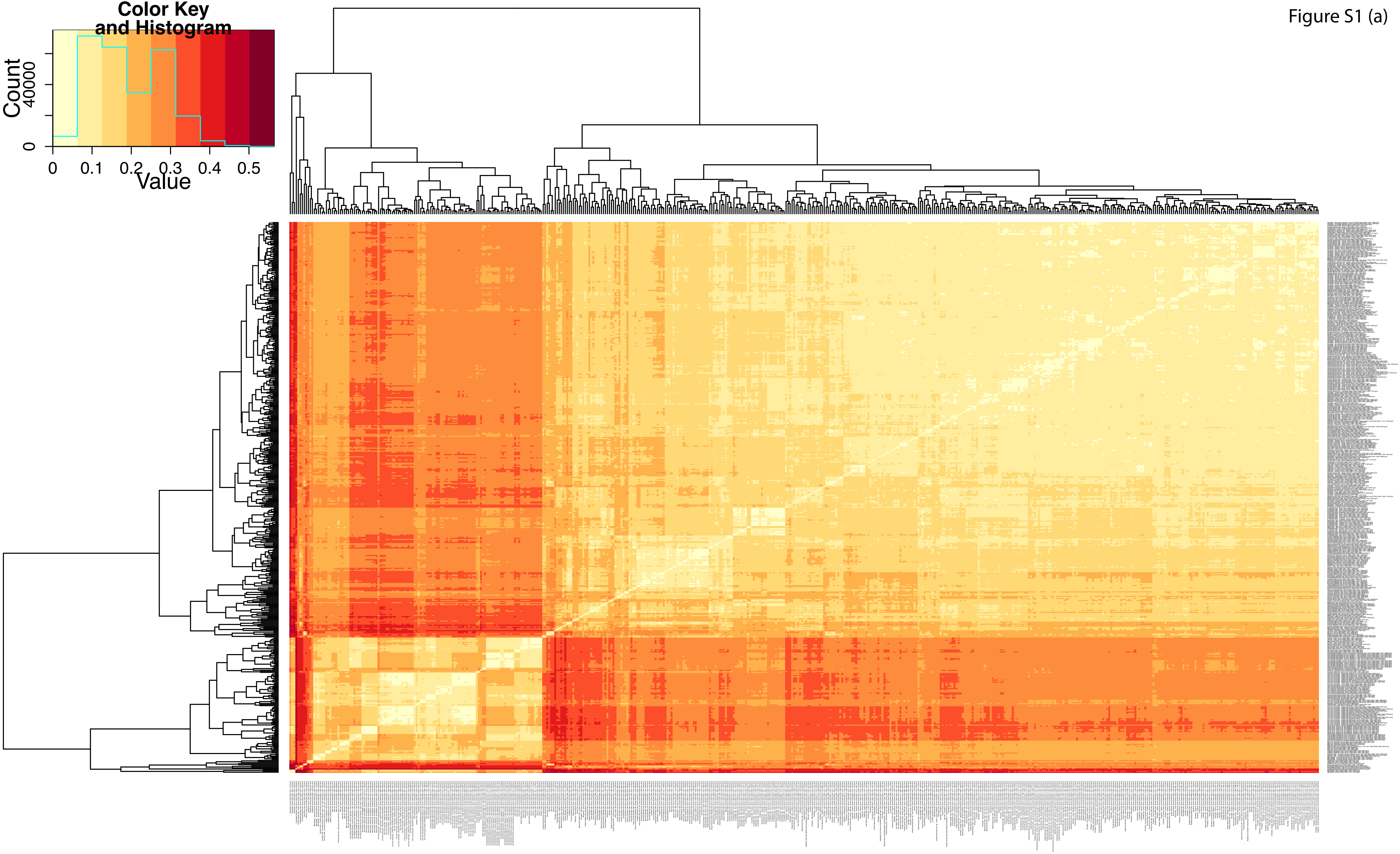
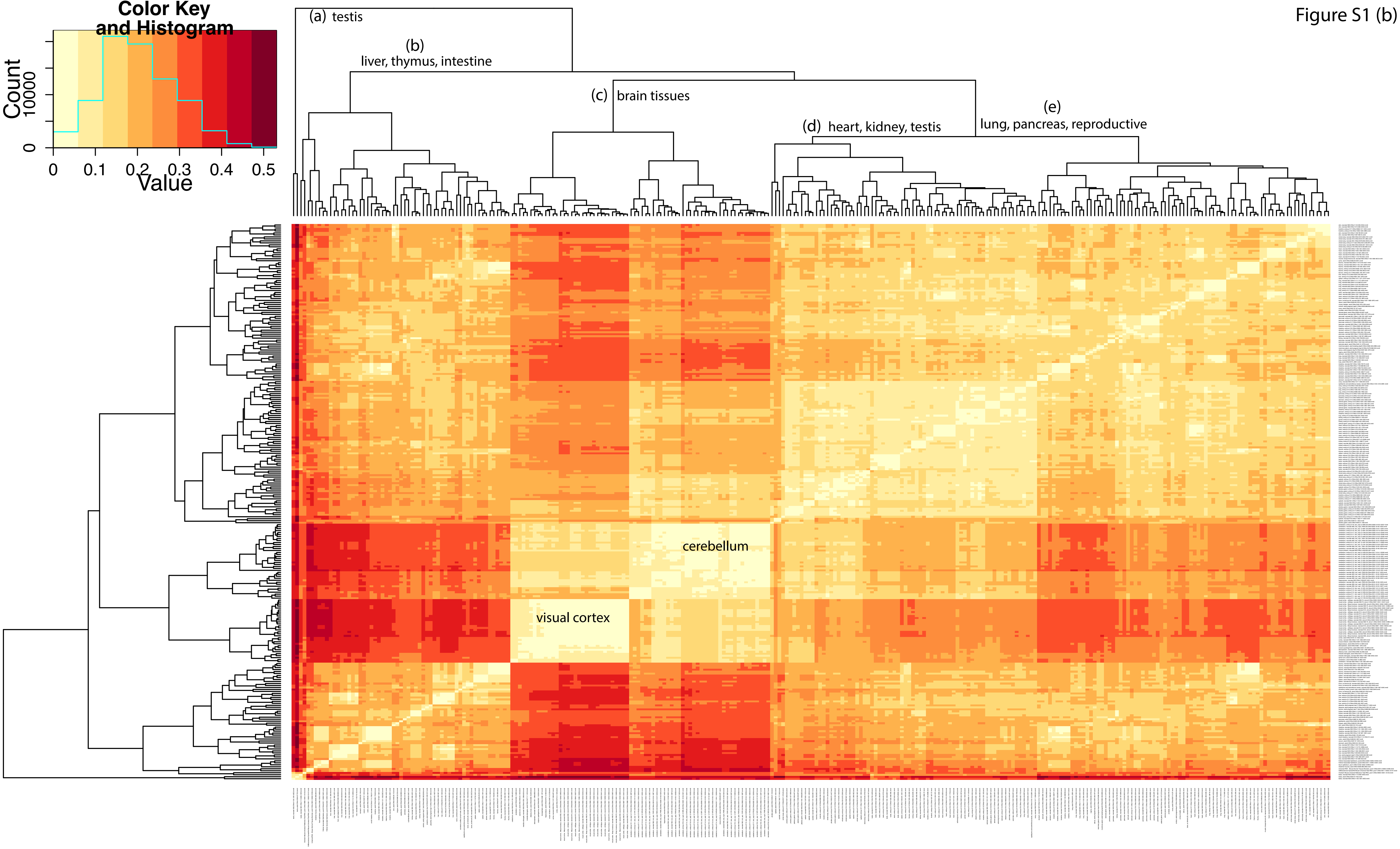


Figure S1 (b)



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)

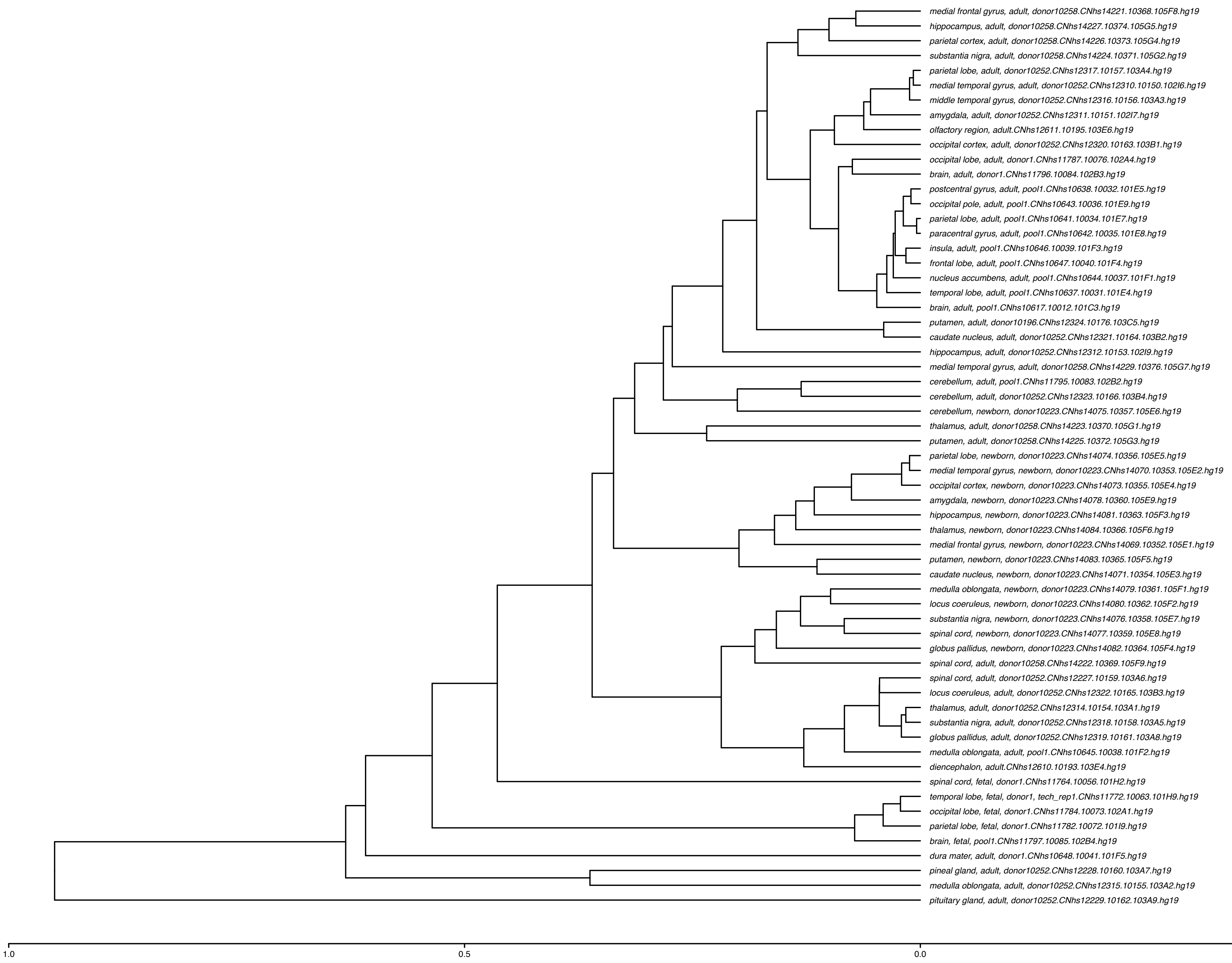
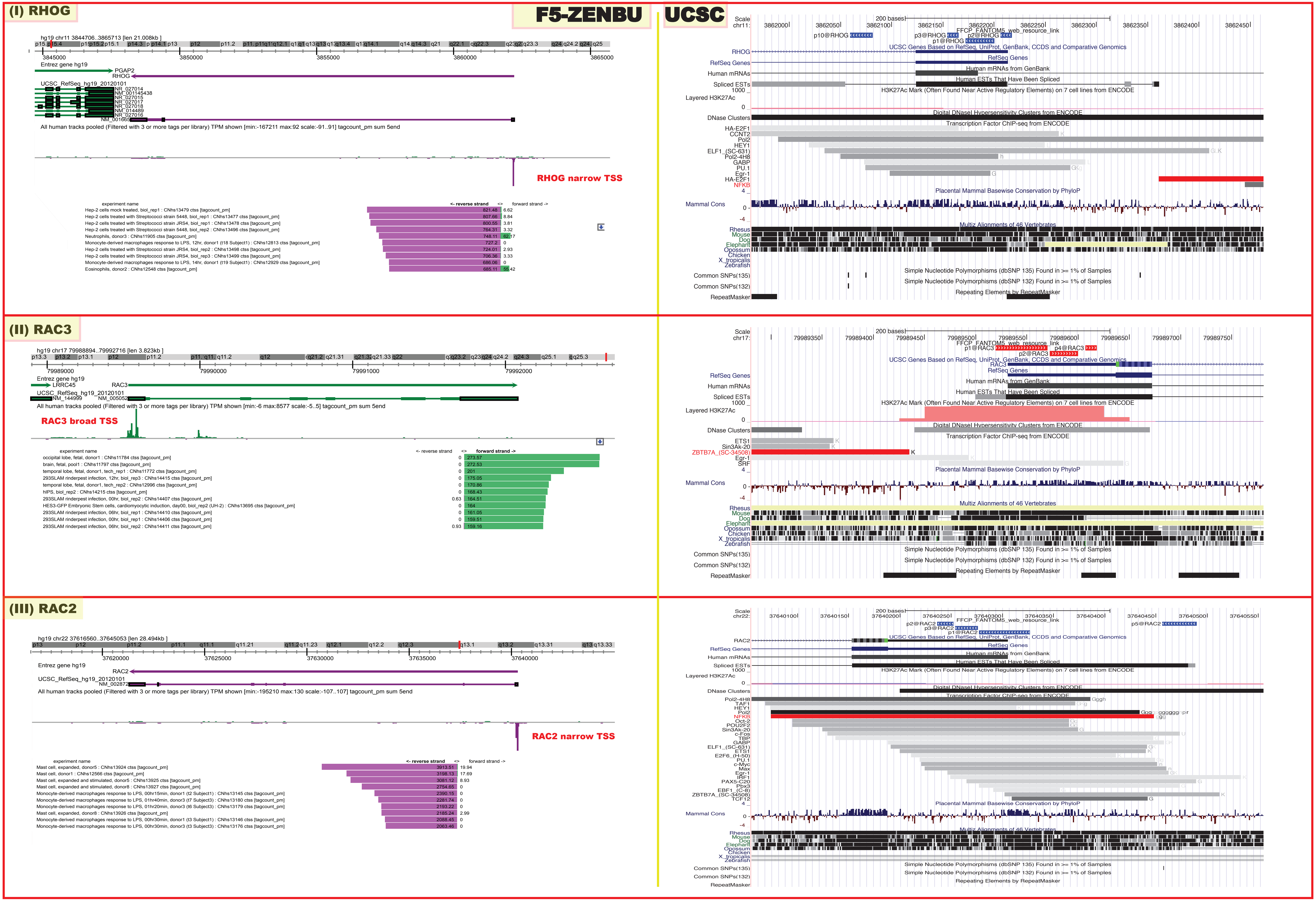

medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
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 2


(I) RHOG
F5-ZENBU
UCSC
hg19 chr11 3844706..3865713 [len 21.008kb]
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
(II) RAC3
hg19 chr17 79988894..79992716 [len 3.823kb]
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
(III) RAC2
hg19 chr22 37616560..37645053 [len 28.494kb]
Entrez gene hg19
RAC2
UCSC_RefSeq_hg19_20120101
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
Scale
200 bases
chr11:
chr17:
chr22:
FFCP_FANTOM5_web_resource_link
p10@RHOG
p3@RHOG
p2@RHOG
p1@RHOG
p1@RAC3
p4@RAC3
p2@RAC3
p2@RAC2
p3@RAC2
p1@RAC2
p5@RAC2
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Common SNPs(135)
Simple Nucleotide Polymorphisms (dbSNP 135) Found in >= 1% of Samples
Common SNPs(132)
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
HA-E2F1
CCNT2
Pol2
HEY1
ELF1_(SC-631)
Pol2-4H8
GABP
PU.1
Egr-1
NFKB
ETS1
Sin3Ak-20
ZBTB7A_(SC-34508)
SRF
TAF1
Oct-2
POU2F2
c-Fos
TBP
E2F6_(H-50)
c-Myc
Max
IRF1
PAX5-C20
Pbx3
EBF1_(C-8)
TCF12

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

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

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