Title page:

Oxana Sachenkova*%, CORE-RIKEN-AUTHORS# and Lukasz Huminiecki*&%

* Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden
& Department of Cell and Molecular Biology, Karolinska Institutet, Sweden
% Science for Life Laboratory, SciLifeLab, Sweden
# RIKEN Omics Centre, Yokohama, Japan

Corresponding author: Lukasz.Huminiecki@scilifelab.se

Keywords: gene expression, CAGE, evolution, cancer

Title:

Patterns of expression change in the F5-CAGE encyclopedia of vertebrate gene expression

Abstract:

F5-CAGE encyclopedia of expression patterns is arguably the most comprehensive and technologically uniform functional genomics dataset ever generated. F5-CAGE includes 952 human and 396 mouse tissues (T), primary cells (PC) and cancer cell-lines (CCL). Here, we use F5-CAGE to explore expression space change in multiple contexts.

Brain exhibits unique transcriptional features, including clustering into fetal, newborn, and adult samples. All samples group into three distinct categories with respect to expression evolution rate. There is trend for young genes to be tissue-specific, with the exception of taxon Eutheria. A major divide between leukemias and solid tumors is seen in CCL. Paralog expression pattern divergence suggests global devolution of expression in CCL. We explore global differences between T and CCL samples further, though family analysis and self-organizing maps. As a focused family evolution example, we use cdc42 family which features many tissue-specific genes and dramatic expression pattern shifts, correlated with ENCODE Tfbs. PhyloSigs suggest novel hypotheses for animal evolution: CNS and reproductive track are discussed as two examples.

Most genes have multiple associated CAGE clusters, with up to 87 for tintin, contributing to multiple isoforms which were previously attributed to alternative splicing alone. TSSes correlate between human and mouse, older genes tend to have more TSSes, and TSS-rich genes are associated with cancer.

Finally, we test the hypothesis of CTCF acting as isolator between paralogs, and instead show its function is more likely in bringing duplicates under the control of the same enhancer. The trend is illustrated with semenogelins and pregnancy specific glycoproteins.

(247 words)

Introduction:

Animals are primarily characterized by multicellularity, and existence of distinct tissue- and cell-types. Exploration of change in the animal expression space is of general interest to broad readership. Here, we present the F5-CAGE expression encyclopedia of vertebrate expression patterns, and use it to explore change in the animal expression space.

F5-CAGE is arguably the most complete functional genomics dataset generated to date (FANTOM5 2013). The major strengths of F5-CAGE are comprehensiveness and technological uniformity. Moreover, F5-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 F5-CAGE resource.

Expression datasets available to date, have been limited in scope to a narrow and biased set of somatic tissues, and not allowing for parallel examination of coding and non-coding transcripts (Su, Cooke et al. 2002). Meta-analysis of expression datasets is 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) and clustering expression profiles (Eisen, Spellman et al. 1998; Quackenbush 2001). However, fairly early in this debate, we signaled that Person R works best for tissue-specific genes, and that conservation of expression profiles is a complex phenomenon, most likely dependent on tissues and functional classes of genes of interest (Huminiecki and Wolfe 2004). Data presented here strengthen these early conclusions.

F5-CAGE should settle heated debates on animal expression pattern evolution. In comparison to ESTs and microarrays, F5-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 protein-coding genes chosen by the chip maker.

For brevity, henceforth we refer to subsets of F5-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 F5-CAGE, while mm-T refers to all murine tissue samples (see also Table 1).

Herein, we scrutinize the theme of expression change in different contexts. We explore animal expression space in the context of tissue-specificity, in context of alternative transcription start sites (TSSes), in context of insulators and enhancers, in context of transcription factor binding sites (Tfbs), in context of differences between normal tissues and cancer cell lines, and finally through characterization of phyloexpression signatures (PhyloSigs).

Results:

Hierarchical clustering of human tissue samples
When Spearman correlation is used as expression distance measure, F5-CAGE WP4 samples samples cluster primarily depending on cell/tissue type of origin, not donor or developmental stage. See heatmaps in Figure 1 for hs-T, FigS1a for hs-PC, FigS1b for hs-CCL; and FigureS1c for mm-T. Clearly, tissue-of-origin differences are more important than individual variability: multiple donors are available for a high proportion of samples. Interestingly, for both hs-T (Figure 1) and mm-T (FigureS1c) brain tissues tend to cluster together, and hs-CCL shows a major divide between leukemias and solid tumors (FigS1b).

Principal component analysis (PCA) exploration of the F5-CAGE encyclopedia
The initial exploration of expression pattern diversity in the F5-CAGE encyclopedia using PCA revealed no 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. However, a PCA focused on human brain revealed unexpected clustering into fetal, newborn and aged adult samples (Figure 2).

Unique expression features of brain samples
Previously, we noted that brain has uniquely complex transcriptional program (Huminiecki, Lloyd et al. 2003). Brain-expressed genes mostly derive from 2R-WGD, and few novel brain-specific genes forming during diversification of mammals (Huminiecki and Heldin 2010). The results presented here, both strengthen and clarify these hypotheses. Brain samples clustered together (Figure 1), and were clear outliers in many different analyses. For example, brain samples exhibited unique clustering by age of donor, forming three clusters which can be attributed to developmental stage: (a) fetal, (b) newborn, and (c) aged adult (Figure 2). No other tissue type demonstrated this effect. Multiple donors and developmental stages were available for many other tissue types, precluding the possibility that observed effect was due to post-mortem delay, or differential pre-processing of embryonic samples during RNA isolation. Take together, these results suggest unique features associated with expression pattern evolution in the CNS.

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 (FigS4). 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). This suggest that conservation of expression profiles is tissue-specific, and might depend on the TF-profile and function of a given cell-type.

Phylogenetic timing of gene duplications using TreeFam8 database
We have previously used TreeFam database (Li, Coghlan et al. 2006) to phylogenetically time gene duplication events (Huminiecki and Heldin 2010). New and much larger release of the database, TreeFam8 (TF8), was used here. TF8 confirmed TF6 results (Huminiecki and Heldin 2010), linking taxa Vertebrata and Bilateria with two greatest waves of gene duplications in the history of animal kingdom.

TF8 trees were linked with F5-CAGE to produce WP4 expression tables: RefSeq transcripts were linked with CAGE tags +/- 500 bps from the TSS. Later analysis stages were performed in R/Bioconductor (v2.11) using among others Biodist, gplot, ggplot, rtracklayer, TxDb.Hsapiens.UCSC.hg19.knownGene, and GOstats.

Assignment of ortholog tissues

To investigate the problem of ortholog tissue assignment systematically, we compared ortholog clusters derived through (a) 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, just like hierarchical clustering on which it is based, is an unsupervised learning approach.

Table 2 summarizes comparison of ortholog human-mouse tissue clusters obtained by simple name matching, with those inferred using the ISHC. Full set of NM- and ISHC-clusters is provided in TableS2. The ISHC clusters tissues according to ortholog gene expression patterns (Fig4a). 8 NM-clusters are recovered by the ISHC, but the majority (19 clusters) are not (Table 2). The NM-clusters which are recovered through the ISHC include: skin, liver, tongue, heart, pancreas, pituitary gland, thymus and “total RNA control”. However, twice as many NM-clusters (namely 19) were not recovered. This effect seems robust to alterations of the ISHC procedure. For example, we have experimented with other expression distance measures, and ISHC based on whole family-averaging rather than ortholog genes, but not higher rates of recovery could be achieved (data not shown).

Histogram in Fig3b contains stack histogram with Pearson R distributions, obtained during the course of the ISHC procedure, for the two intra-species comparisons (hs, mm), and the inter-species comparison (hs-mm). Inter-species distances (hs-mm) are somewhat higher than intra-species distances (hs and mm): 0.78, 0.68 and 0.72 are the respective means.

Phyloexpression signatures (PhyloSigs)

A phyloexpression signatures (PhyloSig) is defined as a strong expression association between individual F5-CAGE samples, and gene duplicates derived from a given taxon. A PhyloSig is a sample+taxon 2-tuple. TableS7 lists top three PhyloSigs for each taxon. For example, genes derived from human specific duplications, tend to be expressed in thymus, liver and adipose tissue. 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 are included in PhyloSigs of many taxa, suggesting immune system innovation as the leading theme throughout animal evolution. Figure 6 focuses on PhyloSigs associated with the reproductive system, while Figures S3 focuses on those for brain tissues. TableS8 lists genes “behind” several PhyloSigs from TableS7 and Figure 6. Chosen sample+taxon associations include: seminal vesicle+Catarrhini, testis+Eutheria, uterus+Eutheria, adipose+human, vagina+human, salivary glad+Catarrhini.

Young duplicates tend to be tissue-specific

Using hs-T data, we observed two broad evolutionary trends related to animal expression patterns: a previously described trend for gradual paralog expression pattern divergence (FigureS2), and a novel trend for young genes to be more tissue-specific in their expression domain (Figure 5). Taxon Eutheria, however, is an outlier both to the gradual divergence trend (FigureS2) and the expression breadth trend. Bilateria is also an outlier to the expression divergence trend (FigureS2), although to a lesser extent. Overall, paralog expression divergence is fast, reaching plateau as soon as taxon Amniota. In contrast, the trend for older genes becoming more and more housekeeping, doesn't reach a plateau until the base of the animal tree of life (taxon Metazoa).

When these analyses were extended into hs-PC and hs-CCL, similar trends could be observed in PC, but not in CCL. In particular, there is no signature for young duplicates to be co-expressed ($0.8 < PC$) in hs-CCL (Figure 4), as they are in hs-T (Figure 4), and hs-PC (data not shown).

We then asked what are functional characteristics of genes lying at the extremes of these two broad trends. Supplementary tables TableS6a and TableS6b summarize the results of GO term and PFAM domain enrichment, for all taxa. TableS6a relates to the fast expression divergence rate. TableS6b relates to high breadth of expression. In both cases, top 0.75 quantile of the distribution was used.

Several observations for highly correlated (PC top quartile: TableS6a), and broadly expressed (EB top quartile: TableS6b) extremes of distribution, at the strict p=0.00001 cut-off, are particularly interesting:

(a) Eutheria, highly significant association of highly correlated genes 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 PF00125 (histone domain; p=1.13e-10) with **human-specific** highly housekeeping genes;

(c) associations of the following cellular location (CC) GO terms with broadly expressed duplicates mapping to **Homo/Pan/Gorilla** and **Catarrhini**: extracellular region (GO: 0005576; p=1.8e-07), extracellular space (GO:0005615; p=5.79e-08);

(d) under-representation of 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 **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. Secondly, the emergence of placental mammals (taxon Eutheria), was associated with a burst of duplications or unusual characteristics in terms their expression patterns, being comparatively widely expressed (Figure 5) and relatively highly co-expressed (FigureS2), with genes involved in chromatin assembly and epigenetic control driving the trend.

Evolutionary history of mammalian tissues.

TableS7 and TableS8 describe strongest associations between taxa of gene duplications, and expression patterns in F5-CAGE (i.e. phyloexpression signatures, or PhyloSigs). To illustrate the power of PhyloSigs, we focus on two examples: reproductive track (Figure 6), and brain (FigureS3).

Rac/cdc42: gene family with dynamic expression pattern evolution

The Rac/cdc42 subfamily of Rho GTPases (TF101109) was chosen to illustrate broader concepts of animal expression pattern evolution (Figure 7). Rac/cdc42 genes regulate a diverse array of cellular events, including cell growth.

A novel phenomenon of mutually exclusive expression of paralogs (MEEP) can be seen in Figure 7. RAC3 is highly expressed in five fetal tissues from which RHOG, and RAC2 are excluded (namely: fetal parietal lobe, fetal temporal lobe, fetal duodenum, fetal occipital lobe, and fetal brain pool). On the other hand, RHOG is highly expressed in adult corpus callosum, where the other two genes are not detectable. 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), expresses RHOG and RAC2, but not RAC3.

Rac/cdc42 TreeFam family tree, weblink referenced in (TF101109), shows that these genes derive from 2R-WGD (Huminiecki and Heldin 2010). In contrast, divergence between RAC2/3 ancestor and RHOG predates the origin of animals. For clarity, it should be mentioned that RHOJ and RHOQ are divergent family members weakly expressed in hs-T. RAC1 and cdc42, also highly diverged, are similar in their expression in hs-T (Figure 7).

What is the most parsimonious scenario for the evolution of MEE in the Rac/cdc42 family? We suggest expression pattern neofunctionalization following RAC2/3 duplication. RAC2 retained most of the ancestral expression characteristics (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 TFs responsible for the MEE phenomenon.

Examples of loci with multiple alternative transcription start sites (TSSes):
F5 Decomposition-based Peak Identification (DPI) procedure (FANTOM5 2013) identifies separate transcription start sites (TTSes), associated with either (a) protein-coding genes, (b) non-coding genes, or (c) novel transcribed loci in the genome. In total, the DPI procedure identified 184,827 robust peaks in human, and 116,277 in mouse. The numbers of robust F5-DPI-CAGE peaks located within +/-500bp of 5'-ends of RefSeq annotated transcripts were 93,614 and 61,072, in human and mouse, respectively (FANTOM5 2013). Those 5'-proximal peaks were used to infer transcription patters of protein-coding genes in the F5-CAGE protein-coding gene expression tables (so-called WP4 tables).

Histogram in Figure 9a shows frequencies of genes (NCBI Gene collection) with increasing number of 5'-proximal F5-DPI-CAGE peaks (i.e. putative TSSes). There are 24677 RefSeq human genes in total in this dataset, with 91671 TSSes mapped. 8181 genes are associated with exactly one TSS, 4617 with two, and 3275 with three. There is a long tail of genes associated with high number of TSSes. At the extreme end of this TSS-rich distribution lie 19 human loci associated with more than 40 TSSes (namely MAP1B with 45, IGF2 64, COL1A1 58, DST 51, IGFBP5 43, APOB 64, AKAP13 46, ARPP21 44, TCF4 51, ALB 48, FN1 81, LEF1 42, TTN 87, SORBS2 47, XIRP2 49, CALD1 49, NTRK2 43, GFAP 46, and PDE4D with 60 TSSes). Many of these TSS-rich genes are structural proteins, building blocks of muscle fibers, cytoskeleton or extracellular matrix. For example: collagen, type I, alpha 1 (COL1A1), fibronectin 1 (FN1), caldesmon 1 (CALD1), dystonin (DST), albumin (ALB), titin (TTN), glial fibrillary acidic protein (GFAP), and apolipoprotein B (APOB). Many of TSS-rich genes were previously known to produce multiple protein isoforms, but these were attributed to alternative-splicing, or RNA editing (in case of APOB), not alternative transcription start sites.

We have examined 5 high-TSS loci in the human genome, with the F5-UCSC GB (alternative TSSes are annotated as p1, p2, ..., pX@geneX), or F5-ZENBU GB (F5 custom browser developed to visualize TSS peaks and CAGE tag contributions from different samples). While some of these loci big (for example SEPT9, APOB), others are very compact (for example, UBE2D3, and IGF2). Distribution of TSSes along chromosomes is obviously non-random, and there is no simple correlation between locus size and the number of mapped TSSes.

a) **septin 9 locus** (SEPT9, chr17: 75,222,695-75,551,472; **329kb size locus**) shown in Figure 8 (composite of multiple ZENBU GB views) and FigS12 (locus overview with the UCSC GB). Multiple alternative SEPT9 isoforms have been presumed to result from alternative splicing (McDade, Hall et al. 2007). However, Figure 8 shows that alternative isoforms differ mostly in the 5'-terminus, and alternative 5'-exons align perfectly with F5-CAGE TSSes. This locus also demonstrates well different levels of TSS clustering: starting from all TSSes associated with a locus, through TSSes associated with a single protein isoform, down to base pair level resolution, where the shape and individual CAGE tag library contributions can be measure using the F5-ZENBU genome browser. Unlike for yeast where septins were studied in context of filament formation and polymerization, the function of mammalian septins is largely unknown .

b) **ubiquitin-conjugating enzyme E2D 3** (UBE2D3, chr4: 103,746,336-103,751,072; **5kb**) shown in FigureS13, has multiple alternatively structured 5'-UTRs, which are built using alternative 1st exons, and a fixed 2nd exon (with alternative 5'-boundaries). The structure of alternative UBE2D3 isforms, along with the localization of p1-p22@UBE2D3 ,suggest strongly that alternative TSSes are implicated in biogenesis of UBE2D3 isoforms.
c) **endothelin converting enzyme 1** (ECE1, chr1: 21,595,010-21,704,874; **110kb**) locus is shown in FigureS15. The ECE1 locus spans over 100kb of genomic space, with four known isoforms: endothelin-converting enzyme 1 (NM_001397), ECE1-2 (NM_001113349), ECE1-3 (NM_001113347), and ECE1-4 (NM_001113348).
d) **apolipoprotein B** (APOB, chr2:21,213,641-21,277,606; **64kb**) is associated with 64 TSSes, and described in FigureS16. However, APOB locus, also shows features associated with classic alternative splicing, such as at least three internal exons surrounded by Alu elements. Intronic Alu pairs are known to be associated with alternatively spliced exons. (Lev-Maor, Ram et al. 2008). Significantly, however, no F5-TSSes are associated with internal Alu-flanked introns, proving that F5-CAGE can distinguishe between alternative splicing and alternative TSSes.
e) **Insulin-like growth factor 2 locus** (IGF2, chr11:2,148,248-2,175,234; **27kb**) associated with 64 TSSes and shown in FigureS17. This is a particularly complex locus, which includes two bi-directional promoters, and many additional opposite-strand TSSes associated with transcripts BC017277, BC021076, CU674261.

Finally, we performed gene set enrichment analysis (GSEA) with the MSigDB database for genes associated with most TSSes: 58 genes with more than 30 TSSes (set A), and 19 genes with more than 40 TSSes (set B). Set A if, of course, a superset of B. GSEA-MSigDB analysis revealed strong associations for both sets, including associations with cancer, growth factor pathways, and specific promoter nucleotide motifs.

Composition on sets A and B:

Set A consists of MAP1B, CFLAR, IGF2, COL1A1, KIAA1217, DST, IGFBP3, IGFBP5, FOXP1, APOB, AKAP13, PRRX1, PDE4DIP, RABGAP1L, ARPP21, LMO7, TCF4, RTN4, MUC6, ZEB2, ALB, NDRG2, ANK3, FN1, LEF1, MTUS1, THBS1, TTN, SORBS3, SORBS2, XIRP2, TG, MAP2, COL6A2, GNAS, SERPINA1, SYNE1, H19, ZBTB38, LPAR1, PAPPA, CTNND2, DLG2, CALD1, CD36, CCDC80, ANK2, MYLK, NRXN1, CLU, SEPT9, MBP, NTRK2, GFAP, TYRP1, MYH11, GPM6A, PDE4D; and includes 2 growth factors (IGF2 and TG), two homeodomain proteins (PRRX1 and ZEB2), 1 cell differentiation marker (CD36), 3 kinases (MYLK, NTRK2, and TTN), 7 TFs (FOXP1, LEF1, LMO7, PRRX1, TCF4, ZBTB38, ZEB2), as well as 7 known oncogenes (COL1A1, FOXP1, GNAS, MYH11, PDE4DIP, PRRX1, SEPT9)

Set B consists of MAP1B, IGF2, COL1A1, DST, IGFBP5, APOB, AKAP13, ARPP21, TCF4, ALB, FN1, LEF1, TTN, SORBS2, XIRP2, CALD1, NTRK2, GFAP, PDE4D; and includes one growth factor (IGF2), two TFs (LEF1 and TCF4), two protein kinases (NTRK2 and TTN), and one known oncogene (COL1A1).

GSEA resutls for set A and B:

Top ten MSigDB sets associated with test Set A are (FigureS7 and FigureS8): genes in cancer modules 66, 12, 100, 137, 11, 47, and 2; TSSes containing the motifs TTGTTT and SMTTTTGT, and genes upregulated in angioimmunoblastic lymphoma (AILT).

Top ten MSigDB sets associated with the test Set B are (FigureS9 and FigureS10): genes in cancer modules 12, 23 and 47, genes associated with TSS motifs TTGTTT, TGCCAAR, as well as as genes up- or down regulated in a number of experimental conditions

involving retroviral transduction, RNAi knock-out (PDEF) or growth factor stimulation (e.g. progesterone, PDEF).

Pairs of paralogs overlapping CTCF sites and F5 enhancers (F5-en)

We constructed a dataset of same chromosome paralog pairs (SCPPs). Out of the total number of 71,716 duplication pairs linked to expression data, 1788 pairs were located closer than 10 kb from each other (10kb-SCPPes). 10kb-SCPPes were balanced in respect to their overlap with CTCF sites, i.e. among 10kb-SCPP pairs the ratio of the pairs overlapping a CTCF binding site (10kb-SCPP+CTCF), versus those which didn't (10kb-SCPP-CTCF), was 48% to 51% (864 pairs against 924). SCPPs located further apart were by-chance biased towards CTCF-overlap, and any functional signal was likely to be lost in noise. For example, for the paralogs in the distance of 10-100kbps (Figure 10b), 87% overlapped a CTCF. For pairs located further than 100kb, as many as 99% overlapped a CTCF. We focused, therefore, on 10kb-SCPPes as these pairs were CTCF-overlap-balanced, and CTCF effects might be expected in that genomic range.

10kb-SCPPs overlapping a CTCF (Figure 10a) have a bimodal Pearson correlation (PC) distribution, with the majority of paralogs either strongly correlated in expression (PC in the 0.5-1 interval), or anti-correlated (PC in the -0.5-0 interval). In contrast, the majority 10kb-SCPPes not overlapping a CTCF lie in the low correlation spectrum (PC in the interval between 0 and 0.5).

Next, we controlled for expression breadth (EB) as a possible source of analysis bias. Tissue-specific paralogs can appear more strongly co-expressed when PC is used as expression distance measure (Huminiecki and Wolfe 2004). However, EB of 10kb-SCPP +CTCF is higher (average EB of 0.44) than for 10kb-SCPP-CTCF (0.13). Thus, differences in tissue-specificity cannot account for the observed effects, and controlling for EB would in fact increase the difference between to 10kb-SCPP subsets.

Discussion:

Hierarchical clustering of samples according to their protein-coding gene expression profiles provided first overview of the F5-CAGE comprehensive encyclopedia of vertebrate gene expression (Figure 1, FigureS1a, FigureS1b, FigureS1c). In addition to hierarchical clustering, to examine expression change, we used principal component analysis (PCA). However, no clear overall pattern of clustering was revealed using PCA. This underlines intrinsic variability of expression patterns in different cell lineages, and argues against well controlled global blocks of transcriptional regulation common to sets of tissues, for example from three distinct embryonic layers.

PCA, however, classified human brain libraries form three distinct clusters: (a) fetal, (b) neonatal, and (c) adult (Figure 2). In this context, one ought to consider lengthy brain development and maturation: brains of newborns are arguably the most immature of all babies organs (Figure 2). No similar clustering by developmental stage could be observed for two other human tissues examined in detail (lung and liver), where multiple donors and developmental stages were available (data not shown). This initial brain PCA analysis was followed by hierarchical clustering of human brain samples (dendrogram shown in FigureS1d), which confirmed age-related clusters described in TableS1.

FigureS1e shows comparative-expression-PCA (CE-PCA) performed jointly for both human and mouse brain samples (where human and mouse data were clustered on ortholog genes, following the principle developed for the ISHC). In the CE-PCA, human and mouse samples form two distinctly demarcated species-specific clusters. However, there are no mouse fetal brain samples, and mouse brain data are biased towards cerebellum and visual cortex. Given available data, no developmental stage related clustering can be observed in mouse (Figure S1e), suggesting that murine brain development is different in that respect from hominid's.

Previous studies comparing evolutionary profiles between species assumed a very simple model for assignment of orthology between tissues: automatically assuming that samples with matching names in different species are ortholog tissues. Here, we suggest that this is over-simplification of a complex problem. Name clusters are not easily recovered by unsupervised learning approaches, such as the ISHC (Table 2). What are the reasons? Firstly, as demonstrated by rapid expression divergence between paralogs, expression pattern evolution is fast (FigS2). Lineage-specific expression pattern shifts and tissue-specific evolutionary novelties put into question the very assumption of the existence of ortholog tissues. Conceptually it may simply be wrong to assume that human brain sublocations correspond directly to those in mouse brain. Behavioral, reproductive, and ecological differences between these two species are profound. Secondly, anatomical differences between species with different body size, make it difficult to dissect tissues, and isolate RNAs in a reproducible fashion, especially for organ systems with complex anatomy (such as the CNS, heart, or reproductive system).

Furthermore, 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.

A number of observations emerged through the comparison of several broad human F5-CAGE sample subtypes (tissues - T, primary cells - PC, and cancer cell-lines - CCL). For example, the peak of highly co-expressed paralogs (top quartile of PC distribution) is missing in CCL, in contrast to T and PC (Figure 4). In other words, while recent closely related paralogs tend to be expressed in the same T samples, they are not co-expressed

in the same CCL samples. This suggest that normal conditions of promoter regulation do not exist in CCL The effect cannot be attributed to cell culture conditions alone, as paralog co-expression signature is 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. This is indicative of global distortion of cancer expression space, as paralogs from the top quartile of PC distribution tend to be recent gene duplicates (the majority no older than taxon “Eutheria”: Figure 4a), with similar promoters and in tandem arrangement on the genome.

We followed on these findings with more detailed examination of gene family expression in T, PC, and CCL. The rationale for focusing on gene families, instead of individual genes, was to allow for epistatic and pleiotropic functions common in paralogs. Across family members, average expression (AVG), and preferential expression measure (PEM) were calculated (supplementary tables S3, S4, and S5). Data in TableS5 can be queried in multiple additional ways, depending on the biological question. For example, one can identify 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.

FigureS5 shows scatterplot pairs for seven variables (7V) associated with families in T, PC, and CCL: family size, AVG and PEM in each sample subgroup (T, PC, and CCL). While AVG correlate strongly between the three sample types, PEM in T does not correlate with PC and CCL. This is explained by the fact that tissues are mixtures of different cell-types. Clearly, fundamentally different regulatory phenomena are responsible for tissue-specific expression versus cell-line specific expression.

Finally, we have used self-organizing map (SOM) to cluster gene families into groups of similar 7V characteristics in hs-T, hs-PC, and hs-CCL (FigureS6a and FigureS6b, TableS9a and TableS9b). The SOM proved efficient in clustering gene families into groups with similar 7V profiles. Several, prototypes warrant particular attention (**prototype No**: *size*): (**1**: *2*) high average expression in all three sample subtypes (gene families of SH3 domain-binding glutamic acid-rich-like protein, and tumor rejection antigen); (**4**: *3*) high expression in T but low in PC and CCL (including natriuretic peptide precursor A/B, quinoid dihydropteridine reductase, kinesin family member 5); (**30**: *5*) high PEM in PC and CCL, all other variables low (superoxide dismutase 2, RAB7, ATP synthase, RAN, protein disulfide isomerase family A); (**35**: *4*) high PEM in T, lower in PC and CCL, all other variables low (damage-specific DNA binding protein, replication factor C, cleavage and polyadenylation specific factor 3, replication protein A1); (**36**: *2*) very high PEM in PC and CCL, all other variables low (heat shock 27kDa protein, and superoxide dismutase families).

In Results, the Rac/cdc42 family was chosen as a focused example for application of F5-CAGE (Figure 7). Here, we ask whether dramatic spatial expression domain shifts, and mutually exclusive expression of paralogs (MEEP) observed in this family, can be correlated with promoter structure as revealed by ENCODE transcription binding sites (Tfbs). RAC2, RAC3, and RHOG) using pooled ENCODE Tfbs data. Table 3 lists all Tfbs associated with TSSes of these three genes, while Table 4 calculates JI for pairwise promoter comparisons (visualized in Figure 7b). JI has bimodal distribution with peaks in two intervals: 0-0.2 and 0.2-0.5 (Figure 7b-key and Table 4).

We have also examined promoter regions of these three genes manually with the UCSC-ENCODE genome browser (UCSC GB on hg19, FigS14). Narrowly expressed, RAC3 (promoter region at chr17: 79988532 - 79990531) features one strong TF binding site: ZBTB7A (ZBTB7A_(SC-34508); cluster score (out of 1000): 1000; K562 cell line; chr17:79989226-79989435), a weak and narrow DNASEI site (score 24), and a weak H3K27AC signal. In contrast, liberally expressed RAC2 (Promoter = chr22:

37639306-37641305) and RHOG (Promoter = chr11:3861214-3863213) have broad and high-scoring DNASEI sites (DNASEI scores of 121 and 136 respectively), and very strong H3k27AC signals. RAC2 and RHOG have many Tfbs within narrow 500 bps window of their TSSes, with strong NFKB signature for both RAC2 (cluster score (out of 1000): 852; chr22:37640075-37640448; cell line GM12891), and RHOG (cluster score (out of 1000): 716; chr11:3862446-3862758; cell line GM12891). Taken together, these results suggested a scenario where ancestral NFKB binding site was replaced with ZBTB7A in RAC3.

Published literature supports 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 (ZBTB7A): zinc finger and BTB domain containing 7A, had been shown to be an oncogenic TF (Maeda, Hobbs et al. 2005; Maeda, Hobbs et al. 2005), implicated in glioma (Rovin and Winn 2005). Interestingly, ZBTB7A promoter region itself features three ZBTB7A binding sites, suggesting an autoregulatory positive feedback loop (UCSC GB on hg19: chr19:4,066,216-4,068,615).

Overall, there is substantial correlation between co-expression (Figure 7a) and JI (Figure 7b), 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 (FigS14). This analysis is now being extended to global scale, with JI for all TF families.

Our PhyloSigs 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. 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 CTCF sites, yet adjacent genes both down-stream (PI3) and up-stream (SLPI) are also co-expressed in semen (UCSC GB 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 GB on hg19: chr19: 43,172,185 - 43,493,834).

Semenogelins and PSGs provide focused examples of broader genomic trends reported here: the trend for young genes to be specific in their expression domain, and for the potential role of CTCF in coordinating such domain-specific expression in clustered loci.

A sideline to the examination of the Rac/cdc42 family, is that the phyloexpression tree (i.e. the gene tree derived from F5-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 FigureS2: high PC variability for younger taxa, and plateau in expression divergence reached as soon as taxon Amniota.

To examine the landscape of alternative transcription start sites at multi-TTS loci, we have examined a number of them using the FANTOM5-UCSC GB (FigureS12-21). Several of these loci are muscle protein, and some were known to generate alternative isoforms. However, these alternative isoforms were presumed to result from alternative splicing; here we show they derive from alternative TSSes and alternative promoters. In many cases, these TSSes align well with Tfbs from the ENCODE dataset.

To verify biological significance of multiple mapping TSSes, we compared the correlation between the number of mapped TSSes for human and mouse, for 1:1 ortholog genes (using matching gene symbol names). R value of 0.502 for Pearson R was obtained for this correlation (Figure 9b).

CTCF is a chromatin binding factor that was proposed to function in transcriptional regulation as insulator, by binding and looping chromatin, and preventing interaction between promoter and nearby enhancers or silencers (Phillips and Corces 2009). Our goal was to test the hypothesis of CTCF insulator effects using pairs of paralogs as the tool. In this in silico experiment, paralog pairs are equivalent to molecular probes, uniquely calibrated by nature to measure the effects of overlapping CTCF sites and enahncers.

In Results, we showed evidence for CTCF in regulating co-expression of paralogs on short genomic distances (10kb-SCPPes). We examined the overlap between 10kb-SCPPes and F5-enhancers(Sandelin). Figure 11 shows PC density plots for three subsets of 10kb-SCPPes: 10kb-SCPP+CTCF+F5en (red); 10kb-SCPP+CTCF-F5en (green); 10kb-SCPP-CTCF-F5en (blue). Interestingly, out of 50 identified 10kb-SCPP+F5en pairs, 46 were also overlapping CTCF. Only 4 pairs had exclusive overlap with enhancers. Taken together, these results suggest that CTCF sites and enhancers co-localize and cooperate in regulation of expression of paralogs (Figure 11). These results do not fit directly with the hypothesis of CTCF acting as expression insulator, blocking interactions between promoters and enhancers: however, this effect was frequently reported for isolated loci, in specific conditions of X inactivation (Filippova, Cheng et al. 2005), using artificial linear insulator reporter constructs (Degner, Verma-Gaur et al. 2011), or transgenic loci with ectopic enhancers or insulators (Hou, Zhao et al. 2008). In a recent global study (Schmidt, Schwalie et al. 2012), CTCF's role as expression insulator seemed weakly confirmed, but disappointingly only Manhattan distances from liver data were used. Furthermore, recent studies attribute to CTCF function in building complex chromatin structures, in cooperation with cohesin (Lee and Iyer 2012). Finally, elegant studies reporting CTCF's role in promoting enhancer-promoter interactions have already been published in embryonic stem cells (Liu, Scannell et al. 2011) and the protocadherin-alpha gene cluster (Kehayova, Monahan et al. 2011). A model was proposed in which CTCF/cohesin-bound promoter are activated by interaction with the distal CTCF/cohesin-bound enhancers (Dekker 2012). Perhaps, time is due for reassessment of the function of CTCF, a factor which can both block and promote enhancer-promoter interactions, depending on genomic context, tissue of expression, and modulating Tfbs.

Future work. We are currently using the ENCODE dataset in conjunction with F5-CAGE to examine the over-all correlation between promoter structure as defined by ENCODE Tfbs and DNASEI footprints and F5-CAGE expression patterns. The relationship

between TSS divergence between paralogs and their expression pattern divergence is also investigated. Finally, we are also using F5-CAGE to test the ortholog conjecture.

Data access:

Access to F5-CAGE is provided at the FANTOM5 public website, including the UCSC GB mirror, and F5's own CAGE-focused ZENBU 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.

Expression distances and hierarchical clustering
Expression distances were calculated using Bioconductor package bioDist Release (2.11). PC was used for Figure 4, 10 and 11, as it works well for tissue-specific genes. Between-paralog Euclidean distances were used for FigS4, where within-tissue expression divergence rates are measured. Spearman distances were used for T, PC, and CCL heatmaps. Spearman correlation is rank-based and less affected by strong outliers.

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

ENCODE data
Multi cell-line, clustered 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 GB on the GRCh37/hg19 assembly, which includes tracks with subset of ENCODE data from TfbsClusteredV2.

Gene Set Enrichment Analysis (GSEA)
To investigate functional enrichment among genes with most TSSes, we have used the GSEA and the online MSigDB database (database v3.1, JavaScript-based webportal v3.85 **(MSigDB)**.

Conservation of the number of mapped TSSes between human and mouse
Out of total of 24677 human and 20320 mouse genes, 13844 ortholog pairs for human and mouse were identified by matching gene symbol names. These orthologs were linked with 67932 TSSes in human, and 46420 in mouse. For the ortholog-TSS dataset, the median was 4 TSSes linked to one human protein-coding gene, and 75% quantile was 6 TSSes. Maximum was 87. For mouse genes, the median was 3, 75% quantile was 4, and maximum 41. Lower values observed for mouse, could be explained either as the properties of the dataset (CAGE coverage for mouse is lower than for human), or as reflecting genuine biological differences in transcription initiation between these two species.

It should be noted that there were statistically significant differences between TSS coverage for 1:1 orthologs, and whole human and mouse sets. For example, the average

number of TSSes for all human genes was 3.714, while for human genes with a 1:1 mouse ortholog, it was 4.906 (p-value < 2.2e-16 in Welch Two Sample t-test). Similarly, the average number of TSSes for all mouse genes was 2.9004 , while for mouse genes with a 1:1 mouse ortholog, it was 3.353 (p-value < 2.2e-16 in Welch Two Sample t-test). These differences cannot affect the conclusion about overall conservation of TSS numbers between human and mouse, but significantly higher numbers of TSSes associated with 1:1 orthologs in both human and mouse, suggest that genes which underwent duplication in any these lineages have a lower number of associated TSSes.

CTCF overlap

For performing the analysis on CTCFs location and it's relationship with the expression levels of the paralogs we used the CTCF data provided by the ENCODE project. First, CTCF binding sites were extracted from wgEncodeRegTfbsClusteredV2.bed. This file combines data from many different cell lines and transcription-factor targeting antibodies. The paralogous genes were mapped to their locations using Refseq ids and the known gene promoters data set from the package FunciSNP.data (Coetzee, Rhie et al. 2012). Only the pairs that are located on the same chromosome were considered and the Genomic Range for each pair was constructed from the position of the promoter of the first gene to the position of the promoter of it's duplicate partner. Same chromosome duplicate TreeFam pairs of all ages were considered.

Self-organizing map (SOM)

Bioconductor package Kohonen was used to construct a rectangular 6 by 6 SOM. The resulting SOM had 36 prototypes, which through the process of competitive and cooperative learning clustered observations (i.e. gene families) into groups with similar expression characteristics. The family expression characteristics are described using seven variables (7V) describing family size, as well as average expression and PEM in T, PC, and CCL.

Acknowledgments:

We would like to acknowledge the FANTOM5 consortium.

Disclosure declaration:

The authors declare no competing interests.

FIGURE LEGENDS:

Fig1. 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.

Fig2. Brain samples group into age-related groups.
Human brain samples cluster in a PCA analysis (PC1 versus PC2) into three age related groups. Samples are visualized using colored marks (adult - red, fetal - green, newborn - blue). These clusters are illustrated using oval shapes with boundaries hued in corresponding colors. Few samples are outliers to the overall PCA clustering pattern, for example: putamen, cerebellum, spinal cord, medulla oblongata. Interestingly, most outlying samples correspond to the brain stem, suggesting distinct developmental pattern for hindbrain.

Fig3. 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.

Fig4. 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.

Fig5. Young genes 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.

Fig6. 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.

Fig7. 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.

Figure 8. Alternative TSSes in the SEPT9 locus.
SEPT9 locus illustrates how alternative TSSes cluster at different levels of resolution, and how they drive expression of distinct protein isoforms in different tissues. Here, we look at three levels of resolution: (I) Locus level, (II) Promoter level, and (III) TSS level. Analysis of

top sites of expression (green bars, TPM values), suggests that alternative TSSes have differential expression profiles.

TSS cluster1 is enriched in epithelial expression. cluster2 appears cancer-specific. cluster3 is subdivided into three very distinct subregions, broadly expressed (a), (b) and melanocyte/melanoma-specific subregion (c). Cluster4 shows strong expression in pericytes: endothelial-associated cells which play an important regulatory role in control of vascular permeability and the formation of blood-brain barrier.

Figure 9. Most genes have multiple F5-TSSes, and tendency to be associated with high number of F5-TSSes is conserved between human and mouse.
Panel (a) shows the histogram of the number of F5-TSSes per gene in human, and panel (b) shows the scatterplot of the number of F5-TSSes in human and mouse for pairs of orthologs.

Figure 10. Close paralog pairs overlapping a CTCF site are more similar in expression profile. Density plots of Pearson correlations (PC) between pairs of same chromosome paralogs, depending on distances: (a) distance less than 10kb; (b) distance between 10kb-100kb; (c) more than 100kb. Panel (d) shows histograms of the distances between the paralogs, with (green) or without CTCF overlap (red).

Figure 11. Close paralogs which overlap both an enhancer and a CTCF site are highly correlated in expression. Density plots of Pearson correlation (PC) between expression levels in pairs of same chromosome paralogs (distance less than 10kb) which overlap both an enhancer and a CTCF (turquoise line); overlap only a CTCF (green line); overlap none of those (red line). Number of pairs divided only by the enhancer is very low, and insufficient to draw a density plot.

Supplementary figures:

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

FigS1b. Hierarchical clustering of hs-CCL samples (Spearman correlation).

FigS1c. Hierarchical clustering of mm-T samples (Spearman correlation).

FigS1d. Dendrogram of human brain samples.

FigS1e. PCA of human and mouse brain samples.

FigS2. Expression pattern divergence bar-plot between paralogs.

FigS3. Evolutionary history of mammalian tissues: brain samples.

FigS4. 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).

FigS5. Comparative multivariate analysis of gene family expression patterns
To examine overall differences in average gene family expression characteristics between F5-CAGE tissues, PC, and CCL, we performed multivariate analysis of the following 7 variables associated with TreeFam families in the F5-CAGE dataset (Figure 12a): number of human family members (family_size), average expression in human tissues (AVG_tissues), average expression in cells (AVG_cells), average expression in cancer (AVG_cancer), preferential expression measure in tissues (PEM_tissues), preferential expression measure in cells (PEM_cells), and preferential expression measure in cancer (PEM_cancer).

FigS6. Self-organizing map (SOMs) identifies clusters of families with similar characteristic expression characteristics in the three human sample types
6 by 6 rectangular SOM (R package Kohonen) was used to classify 682 TreeFam manually-annotated familyA families into groups with similar size, average expression and preferential expression in the three human sample types. Classification of all families into a prototype number is given in TableS9a (prototypes are numbered/ordered left to right first, and bottom to top second). “Counts plot” in panel (a) illustrates final prototype sizes, while panel (b) illustrates average variable values for all TreeFam families clustering into a given prototype. Exact prototype sizes are as follows (in ascending numerical order): 2, 7, 8, 3, 3, 5, 7, 12, 16, 7, 10, 6, 13, 9, 28, 9, 23, 16, 12, 37, 55, 4, 19, 22, 28, 26, 56, 38, 8, 5, 57, 78, 32, 14, 4, 2. Prototype 32 is the most populated one, with 78 TreeFam families (top row, 2nd from the left). Exact average variable values in each prototype are given in TableS9b.

FigS7. Functional enrichment. GSEA in top F5-TSS genes.

FigS8. Functional enrichment. GSEA in top F5-TSS genes.

FigS9. Functional enrichment. GSEA in top F5-TSS genes.

FigS10. Functional enrichment. GSEA in top F5-TSS genes.

FigS11. Number of F5-TSSes and gene taxon of origin.

FigS12. F5-TTSes in the septin 9 locus.

FigS13. F5-TTSes in the ubiquitin-conjugating enzyme E2D 3 locus.

FigS14. RAC2, RAC3 and RHOG loci in the UCSC and Zenbu GBs.

FigS15. F5-TTSes in the endothelin converting enzyme 1 locus.

FigS16. F5-TTSes in the apolipoprotein B locus.

FigS17. F5-TTSes in the Insulin-like growth factor 2 locus.

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. Brain samples in human cluster according to developmental stage.

TableS2. The full F5-CAGE clustering dataset is shown: (a) NM-dataset; (b) ISHC clusters.

TableS3 Gene families with differential tissue-specific expression in T versus PC, and CCL.

TableS4. Families with differential average expression in T versus PC and CCL.

TableS5. Gene family expression status in tissues, PC and CCL.

TableS6a. GO term and PFAM domain enrichment, for all taxa, in genes with unusually fast expression divergence rate.

TableS6b. GO term and PFAM domain enrichment, for all taxa, in genes with unusually high breadth of expression.

TableS7. Strongest associations between timing of gene duplication and expression site in F5-CAGE.

TableS8. Example genes behind strongest associations between timing of gene duplication and expression site in F5-CAGE.

TableS9a. SOM analysis of gene families: family size, average expression in T, PC and CCL samples; average PEM in T, PC and CCL samples. Assignment to one of 36 SOM clusters/prototypes.

TableS9b. Characterization of 36 SOM clusters/prototypes: SOM prototype number; average family size for all families assigned to the prototype; average expression in T, PC and CCL samples; average PEM in T, PC and CCL samples.

Table 1. Number of samples in distinct F5-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	1.00	0.24	0.07	0.28	0.25	0.17	0.23
RHOG	0.24	1.00	0.10	0.35	0.31	0.07	0.21
RHOJ	0.07	0.10	1.00	0.10	0.09	0.06	0.12
RAC1	0.28	0.35	0.10	1.00	0.25	0.04	0.34
RAC2	0.25	0.31	0.09	0.25	1.00	0.16	0.43
RAC3	0.17	0.07	0.06	0.04	0.16	1.00	0.07
CDC42	0.23	0.21	0.12	0.34	0.43	0.07	1.00

Table 4. Jaccard index for the cdc42 family

TableS1. Brain samples in human cluster according to developmental stage.

fetal	newborn	adult
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	**1st Cluster:** 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 putamen, adult, donor10258.CNhs14225.10372.105G3.hg19 **2nd Cluster:** 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 substantia nigra,adult, donor10196.CNhs13803.10178.103C7.hg19 spinal cord, fetal, donor1.CNhs11764.10056.101H2.hg19	**1st Cluster:** medial temporal gyrus,adult, donor10196.CNhs13809.10183.103D3.hg19 medial frontal gyrus,adult, donor10196.CNhs13796.10170.103B8.hg19 occipital cortex,adult, donor10196.CNhs13798.10172.103C1.hg19 parietal lobe,adult, donor10196.CNhs13797.10171.103B9.hg19 medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19 hippocampus,adult, donor10196.CNhs13795.10169.103B7.hg19 amygdala,adult, donor10196.CNhs13793.10167.103B5.hg19 hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19 parietal cortex, adult, donor10258.CNhs14226.10373.105G4.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.CNhs12 . 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 caudate nucleus,adult, donor10196.CNhs13802.10177.103C6.hg19 thalamus, adult, donor10258.CNhs14223.10370.105G1.hg19 substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19 hippocampus, adult, donor10252.CNhs12312.10153.102I9.hg19 medial temporal gyrus, adult, donor10258.CNhs14229.10376.105G7.hg19 cerebellum, adult, donor10252.CNhs12323.10166.103B4.hg19 cerebellum,adult, donor10196.CNhs13799.10173.103C2.hg19 cerebellum, adult, pool1.CNhs11795.10083.102B2.hg19 cerebellum, newborn, donor10223.CNhs14075.10357.105E6.hg19

fetal	newborn	adult
		2nd Cluster: spinal cord,adult, donor10196.CNhs13807.10181.103D1.hg19 locus coeruleus,adult, donor10196.CNhs13808.10182.103D2.hg19 thalamus,adult, donor10196.CNhs13794.10168.103B6.hg19 globus pallidus,adult, donor10196.CNhs13801.10175.103C4.hg19 diencephalon, adult.CNhs12610.10193.103E4.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 globus pallidus, newborn, donor10223.CNhs14082.10364.105F4.hg19 spinal cord, adult, donor10258.CNhs14222.10369.105F9.hg19 **Outliers:** dura mater, adult, donor1.CNhs10648.10041.101F5.hg19 pineal gland, adult, donor10252.CNhs12228.10160.103A7.hg19 pineal gland,adult, donor10196.CNhs13804.10179.103C8.hg19 medulla oblongata, adult, donor10252.CNhs12315.10155.103A2.hg19 medulla oblongata,adult, donor10196.CNhs13800.10174.103C3.hg19

TableS3. Gene families with differential tissue-specific expression in tissues versus PC and CCL.

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

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

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

TableS4. 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

TableS6. See accompanying file ---.

TableS7. F5-CAGE phyloexpression signatures
Strongest associations between timing of gene duplication and expression site in F5-CAGE.

Taxon	Short name	Full name	value
HUMAN	thymus liver adipose	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 liver, fetal, pool1.CNhs11798.10086.102B5.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19	245.550373474054, 113.830800845426, 108.231343517192
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	334.943127801149, 324.458197889213, 320.092508641026
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	243.468270217443, 124.307182830041, 70.8639227847559
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	566.323654743391, 225.788115451309, 208.247650380362
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	139.028270874016, 134.020715393351, 104.065647888749
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	116.93878540317, 50.2908907452444, 45.6543158685601

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	185.17845759193, 149.778206000194, 139.671894228099
Euteleostomi	thymus adipose thymus	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 thymus, fetal, pool1.CNhs10650.10043.101F7.hg19	78.6409808377661, 48.7050213293096, 42.7078793767614
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	114.337614853883, 74.9230098351664, 58.7017523063515
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	50.0273404519974, 31.0416435172888, 28.2508159948265
Bilateria	thymus adipose pancreas	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 pancreas, adult, donor1.CNhs11756.10049.101G4.hg19	108.48916971048, 73.8207660128358, 56.0027402697822
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	117.867716198574, 51.5054249037973, 50.8844737140422
Fungi/ Metazoa	thymus adipose adipose	thymus, adult, pool1.CNhs10633.10027.101D9.hg19 adipose, donor1.CNhs13972.10184.103D4.hg19 adipose, donor3.CNhs13974.10186.103D6.hg19	450.200187073667, 232.438498377191, 177.112712684284

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	42.7661408222488, 40.9127053558847, 37.0478158146202
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	100.769739509574, 75.1287161094575, 72.4926635448796

TableS8. See accompanying file TableS8.pdf.

TableS9a. See accompanying file TableS9a.txt.

TableS9b. See accompanying file TableS9a.txt.

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.

Coetzee, S. G., S. K. Rhie, et al. (2012). "FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs." Nucleic acids research **40**(18): e139.

Degner, S. C., J. Verma-Gaur, et al. (2011). "CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells." Proceedings of the National Academy of Sciences of the United States of America **108**(23): 9566-9571.

Dekker, J. (2012). "CTCF and cohesin help neurons raise their self-awareness." Proceedings of the National Academy of Sciences of the United States of America **109**(23): 8799-8800.

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

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

Filippova, G. N., M. K. Cheng, et al. (2005). "Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development." Developmental cell **8**(1): 31-42.

Hou, C., H. Zhao, et al. (2008). "CTCF-dependent enhancer-blocking by alternative chromatin loop formation." Proceedings of the National Academy of Sciences of the United States of America **105**(51): 20398-20403.

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

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

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

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

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

Kehayova, P., K. Monahan, et al. (2011). "Regulatory elements required for the activation and repression of the protocadherin-alpha gene cluster." Proceedings of the National Academy of Sciences of the United States of America **108**(41): 17195-17200.

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

Lee, B. K. and V. R. Iyer (2012). "Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation." The Journal of biological chemistry **287**(37): 30906-30913.

Lev-Maor, G., O. Ram, et al. (2008). "Intronic Alus influence alternative splicing." PLoS genetics **4**(9): e1000204.

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.
Liu, Z., D. R. Scannell, et al. (2011). "Control of embryonic stem cell lineage commitment by core promoter factor, TAF3." Cell **146**(5): 720-731.
McDade, S. S., P. A. Hall, et al. (2007). "Translational control of SEPT9 isoforms is perturbed in disease." Human molecular genetics **16**(7): 742-752.
MSigDB "MSigDB."
Mungall, C. J., C. Torniai, et al. (2012). "Uberon, an integrative multi-species anatomy ontology." Genome biology **13**(1): R5.
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.
Quackenbush, J. (2001). "Computational analysis of microarray data." Nature reviews. Genetics **2**(6): 418-427.
Sandelin, A. e. a. Systematic in-vivo characterization of active enhancers across the human body.
Schmidt, D., P. C. Schwalie, et al. (2012). "Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages." Cell **148**(1-2): 335-348.
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. "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.
ZBTB7A. "ZBTB7A." from http://www.factorbook.org/mediawiki/index.php/ZBTB7A.

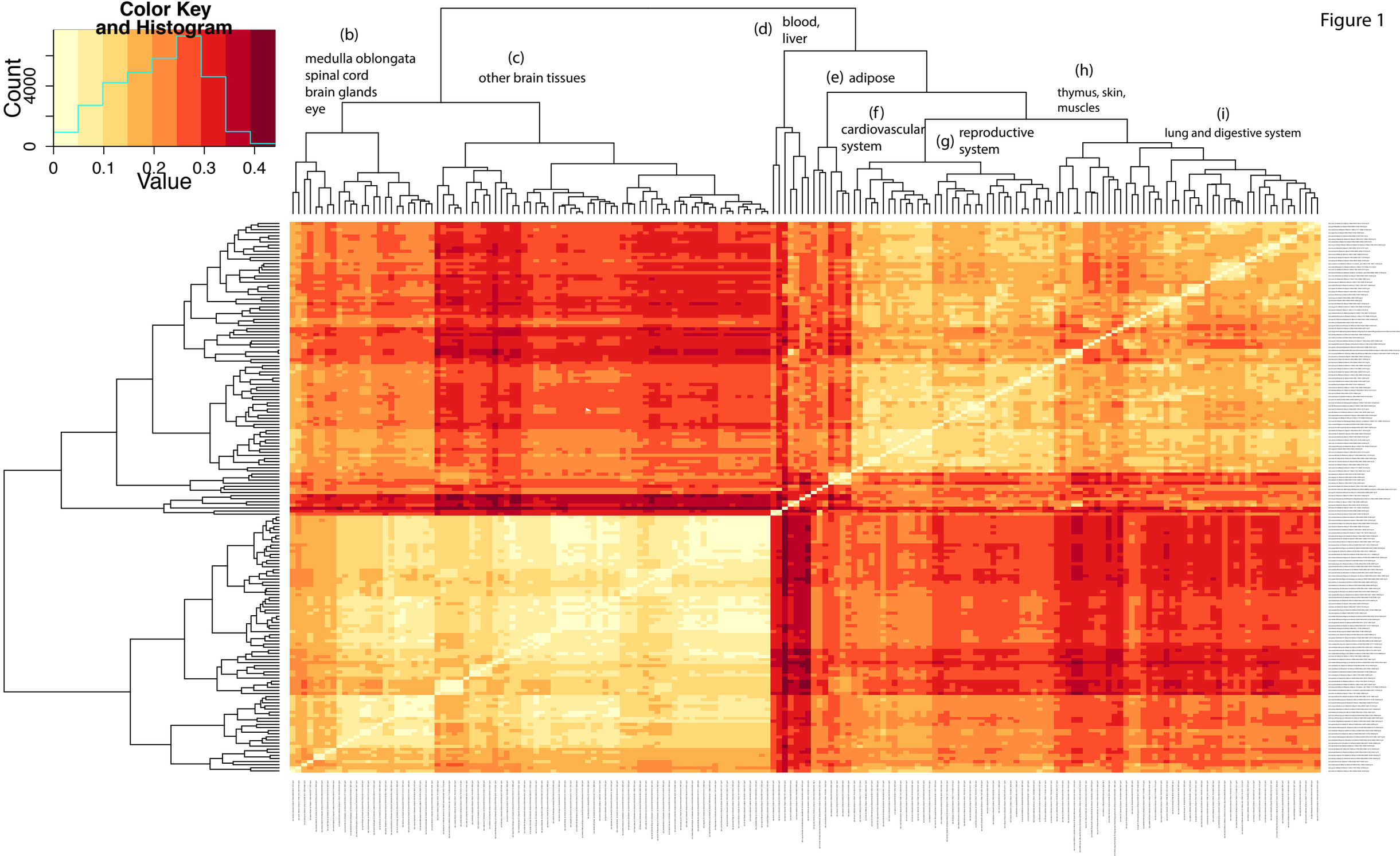
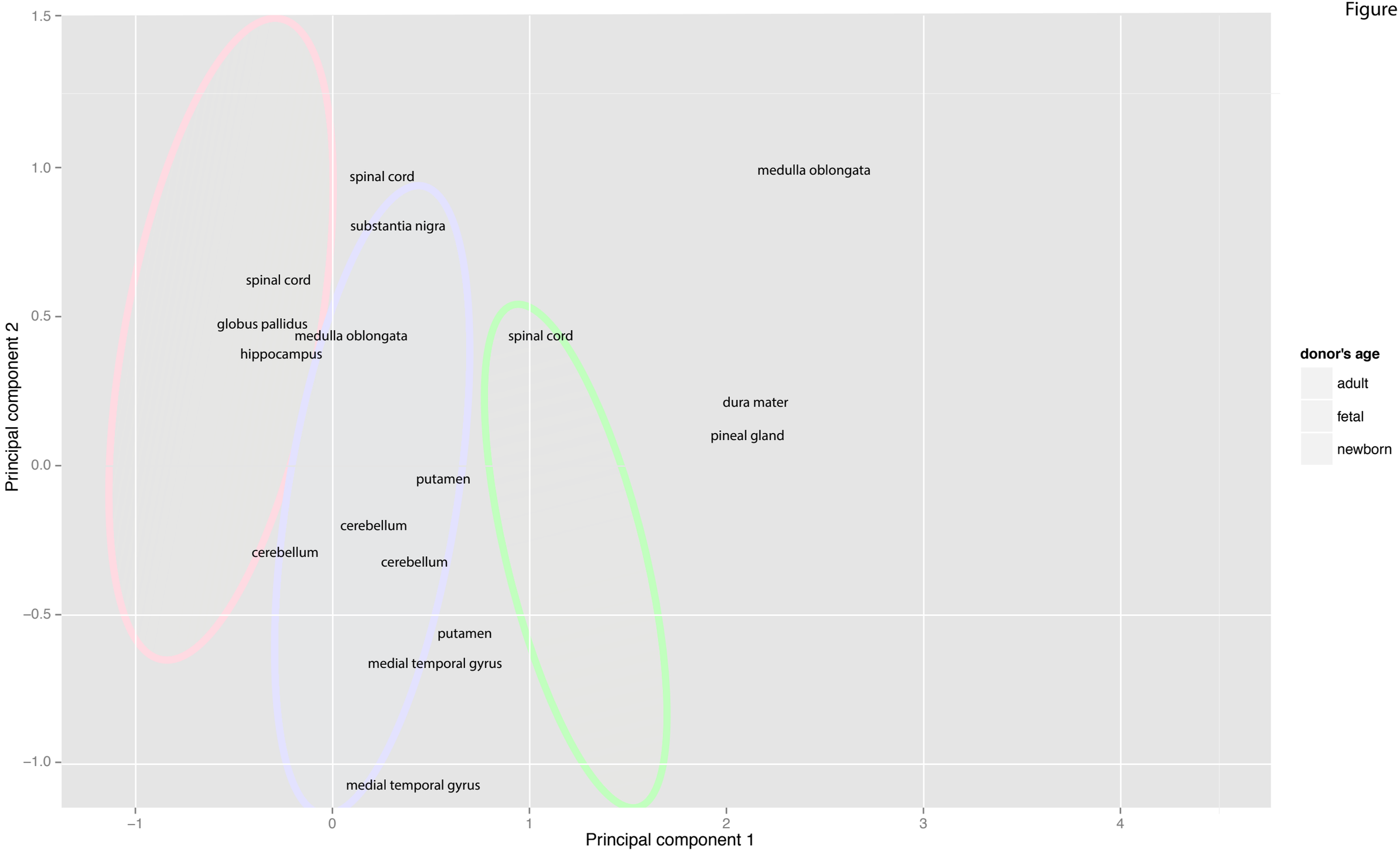

Color Key and Histogram
Count
4000
0
0
0.1
0.2
0.3
0.4
Value
(b)
medulla oblongata
spinal cord
brain glands
eye
(c)
other brain tissues
(d)
blood,
liver
(e) adipose
(f)
cardiovascular
system
(g)
reproductive
system
(h)
thymus, skin,
muscles
(i)
lung and digestive system

Figure 1

Figure 2


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

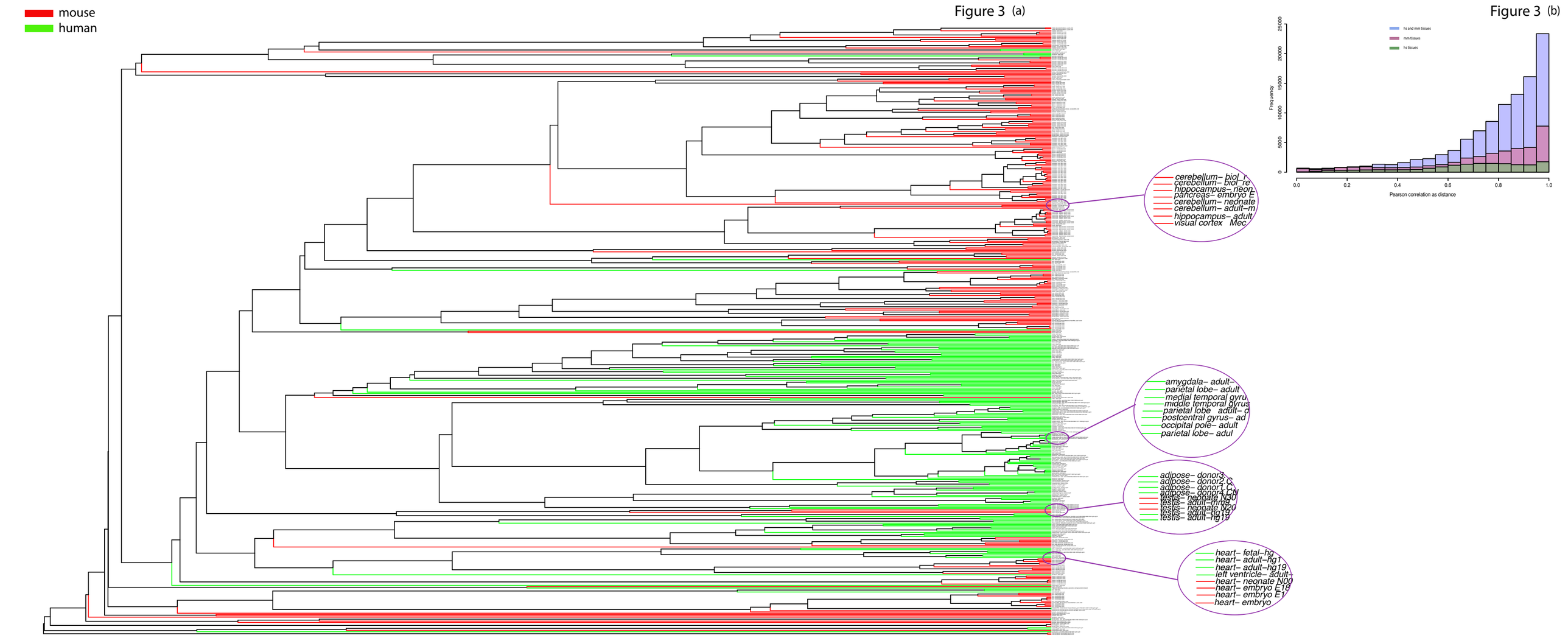

mouse
human
Figure 3 (a)
cerebellum– biol_1
cerebellum– biol_re
hippocampus– neon
pancreas– embryo E
cerebellum– neonate
cerebellum– adult–m
hippocampus– adult
visual cortex Mec
amygdala– adult–
parietal lobe– adult
medial temporal gyru
middle temporal gyrus
parietal lobe adult– d
postcentral gyrus– ad
occipital pole– adult
parietal lobe– adul
adipose– donor3 C
adipose– donor2 C
adipose– donor1 C
adipose– donor4 CN
testis– neonate N30
testis– adult–hrs2
testis– neonate N20
testis– adult–hg19
testis– adult–hg19
heart– fetal–hg
heart– adult–hg1
heart– adult–hg19
left ventricle– adult–
heart– neonate N00
heart– embryo E18
heart– embryo E1
heart– embryo
Figure 3 (b)
hs and mm tissues
mm tissues
hs tissues
Frequency
0
5000
10000
15000
20000
25000
0.0
0.2
0.4
0.6
0.8
1.0
Pearson correlation as distance

Figure 4

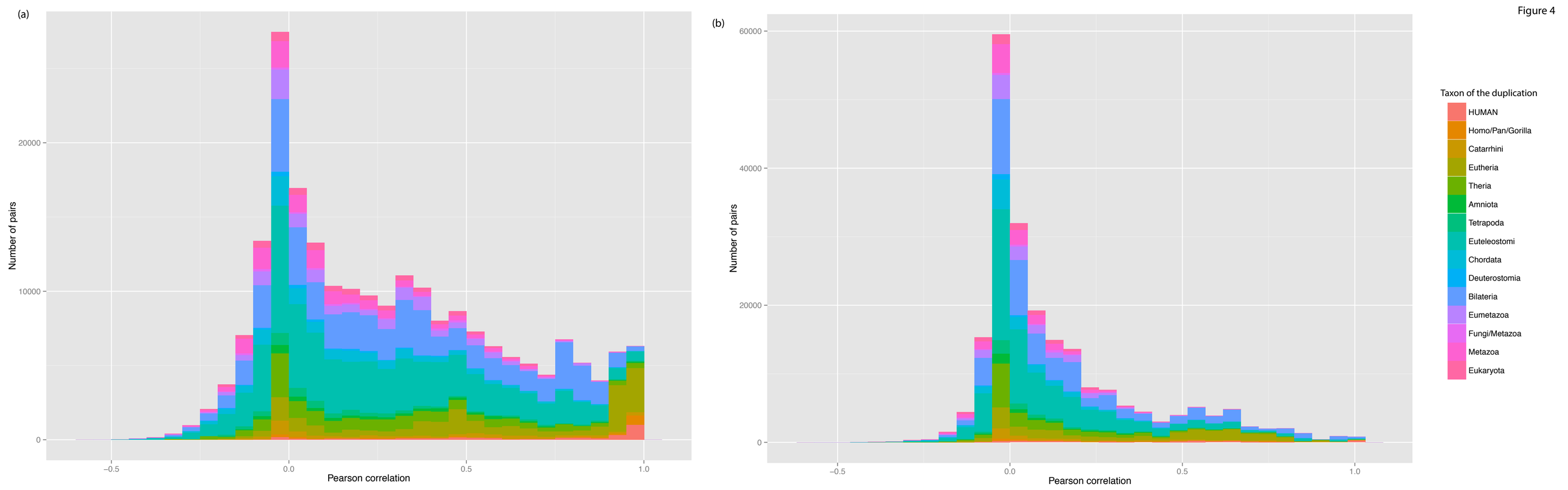

(a)
Number of pairs
20000
10000
0
−0.5
0.0
0.5
1.0
Pearson correlation
(b)
Number of pairs
60000
40000
20000
0
−0.5
0.0
0.5
1.0
Pearson correlation
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa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

Figure 5

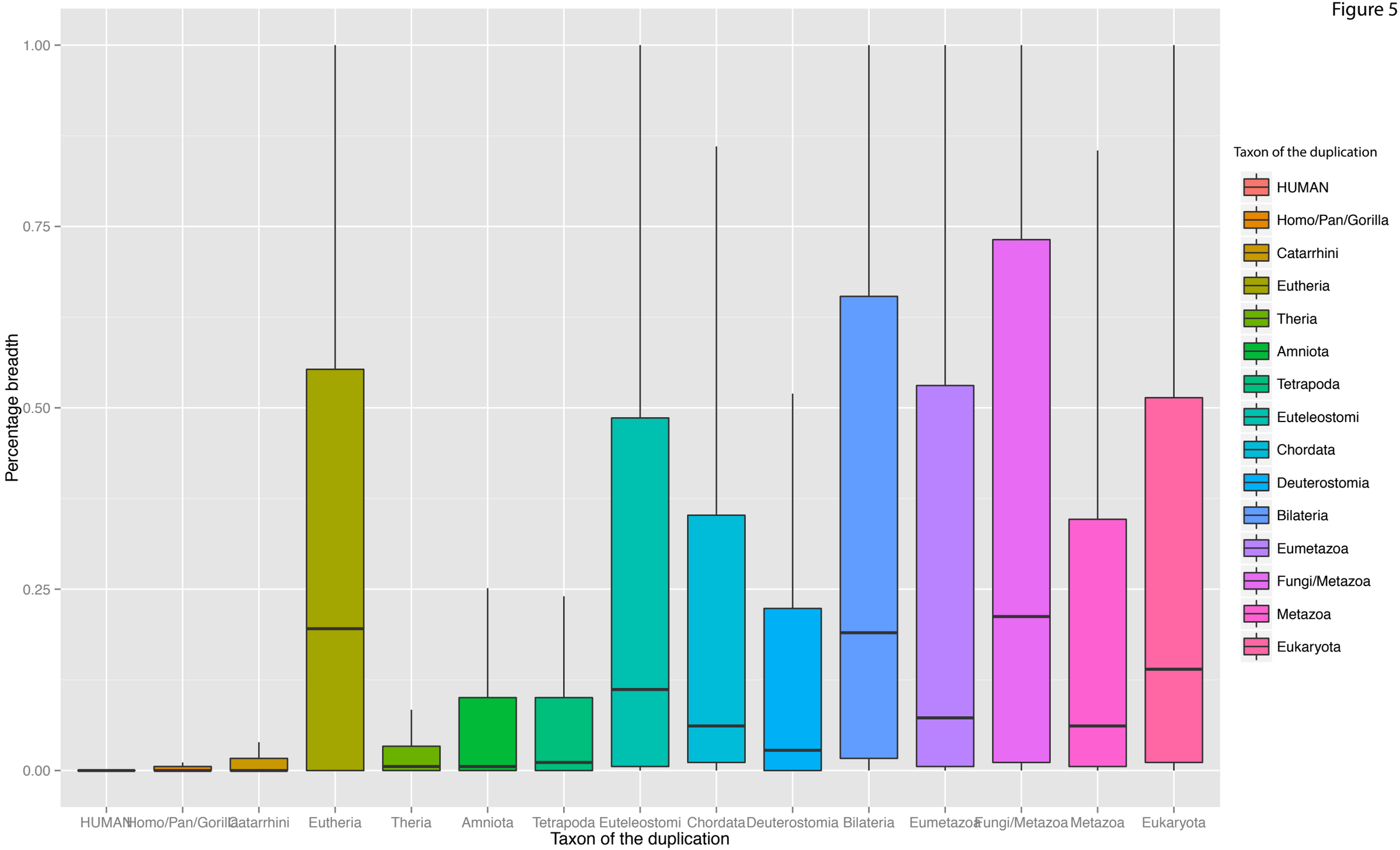

1.00
0.75
0.50
0.25
0.00
Percentage breadth
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
Taxon of the duplication
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota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

Figure 6

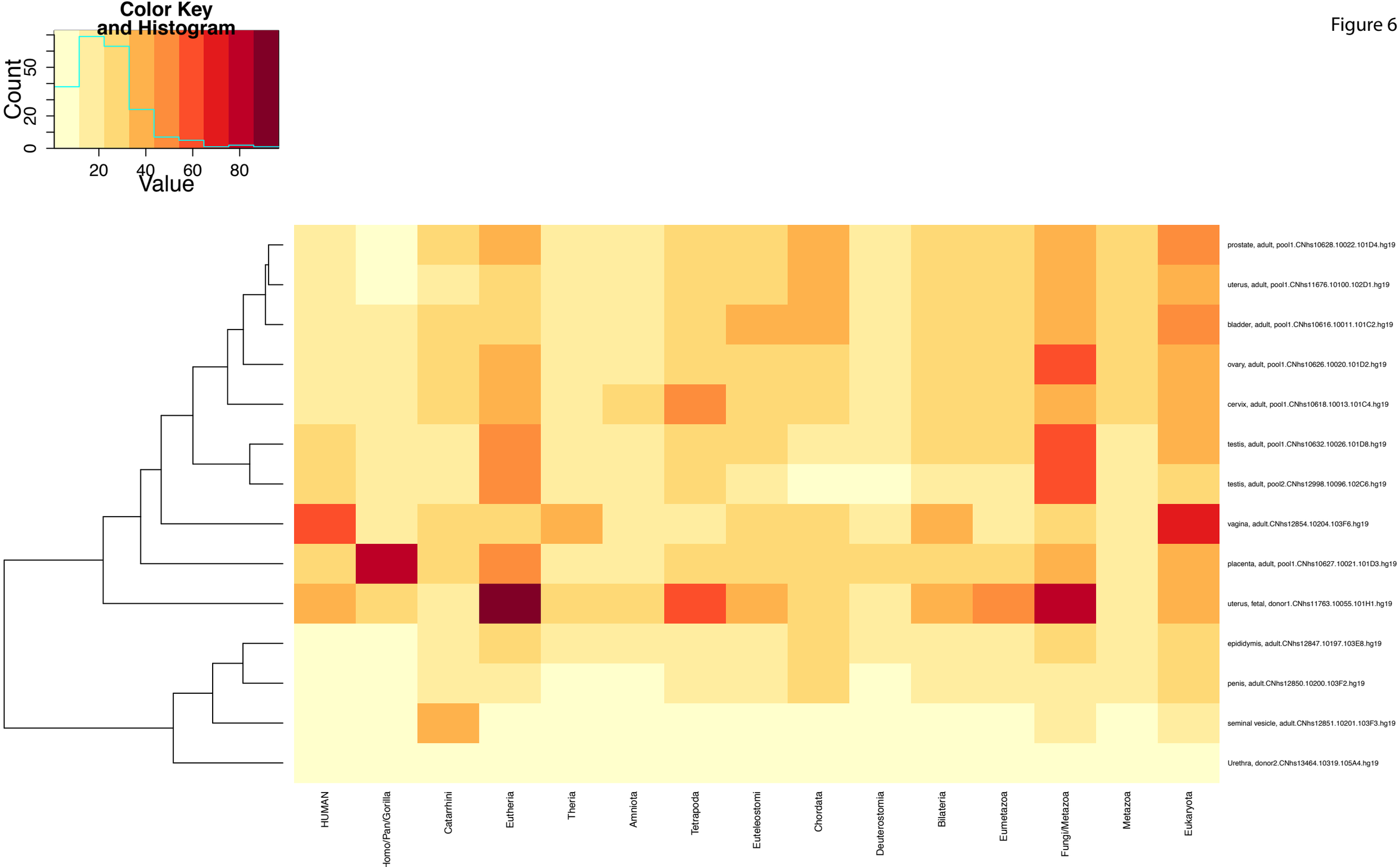

Color Key
and Histogram
Count
0
20
50
20
40
60
80
Value
prostate, adult, pool1.CNhs10628.10022.101D4.hg19
uterus, adult, pool1.CNhs11676.10100.102D1.hg19
bladder, adult, pool1.CNhs10616.10011.101C2.hg19
ovary, adult, pool1.CNhs10626.10020.101D2.hg19
cervix, adult, pool1.CNhs10618.10013.101C4.hg19
testis, 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 7 (a)

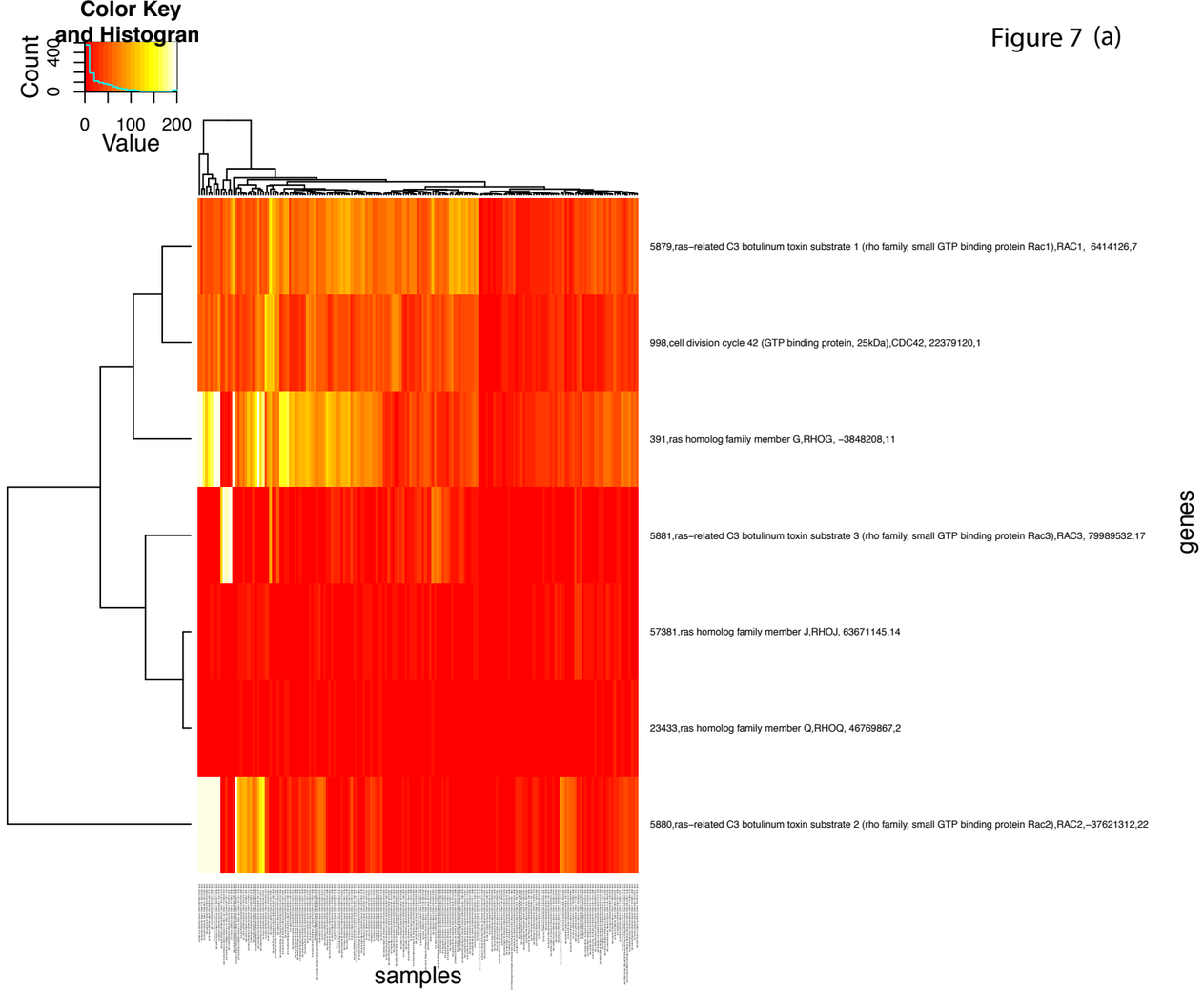
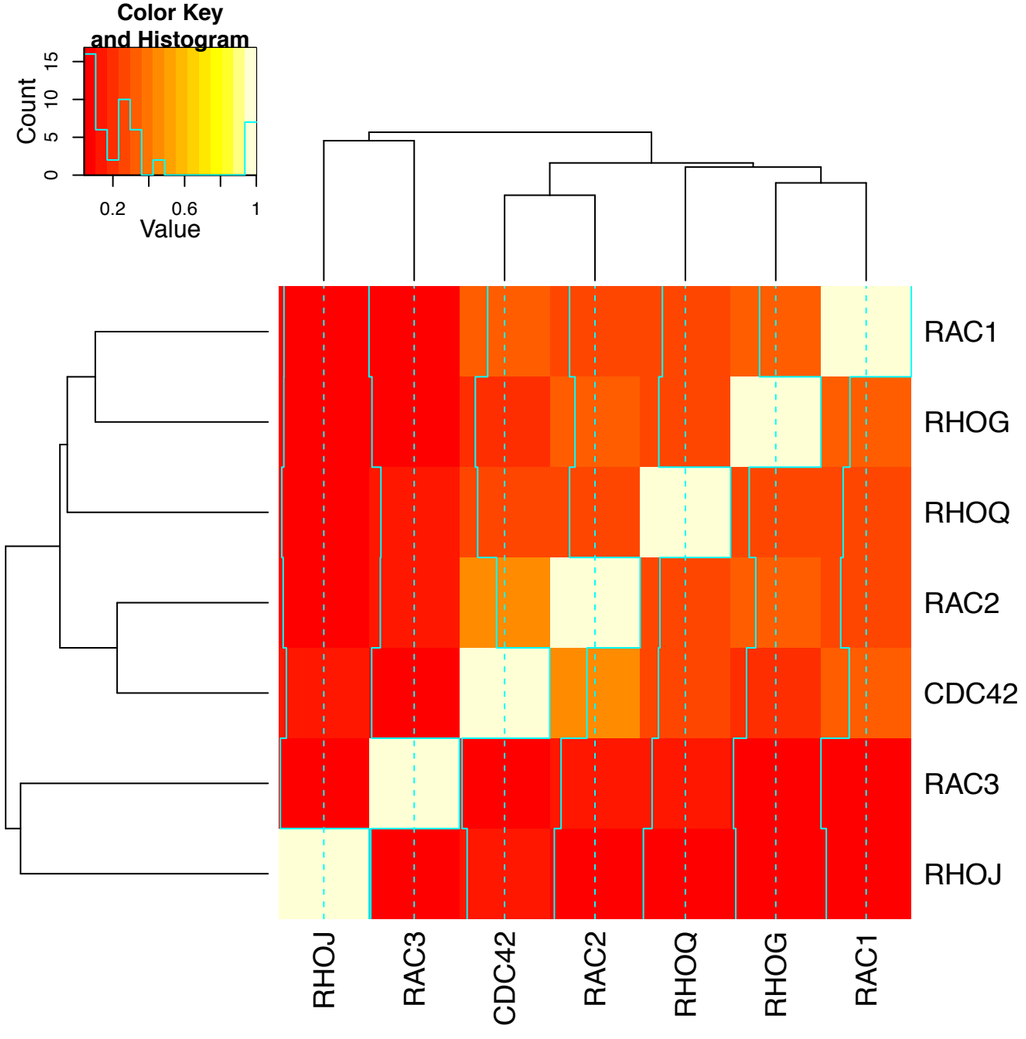
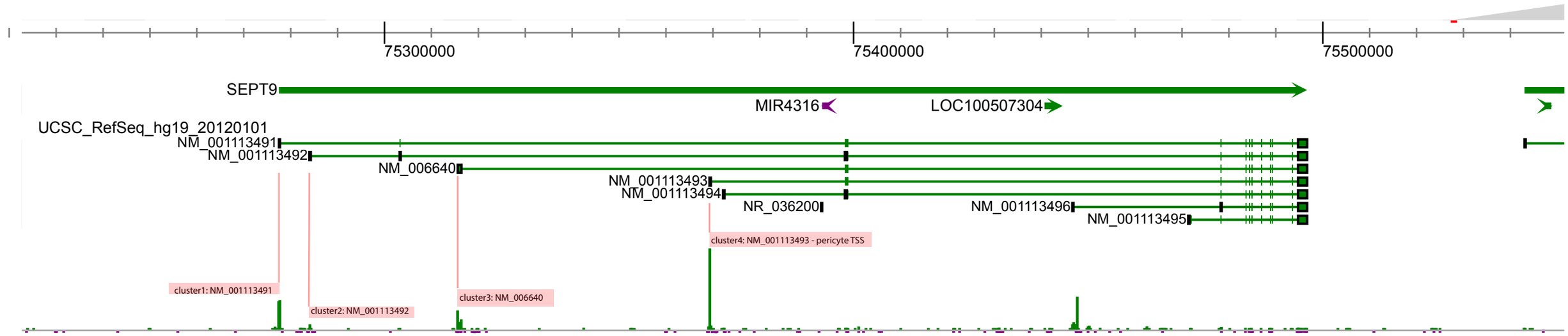
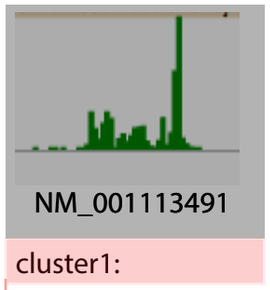


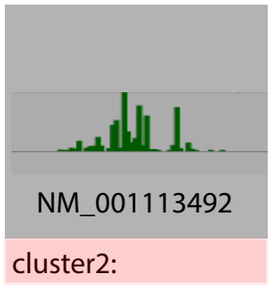
Figure 7 (b)



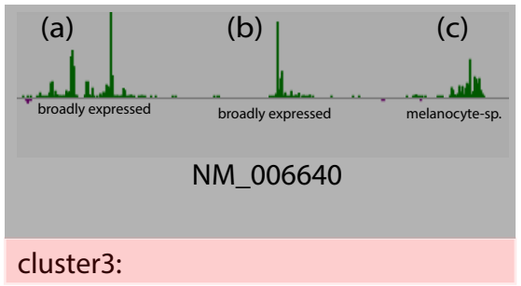




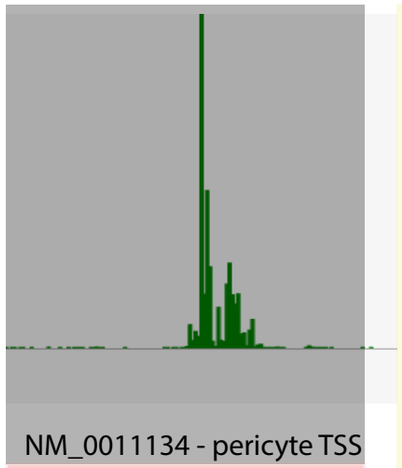
cluster1:



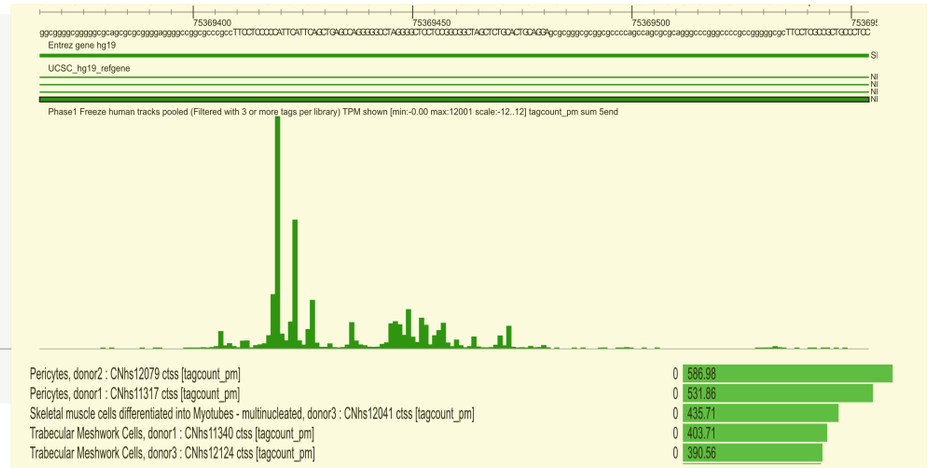
cluster2:



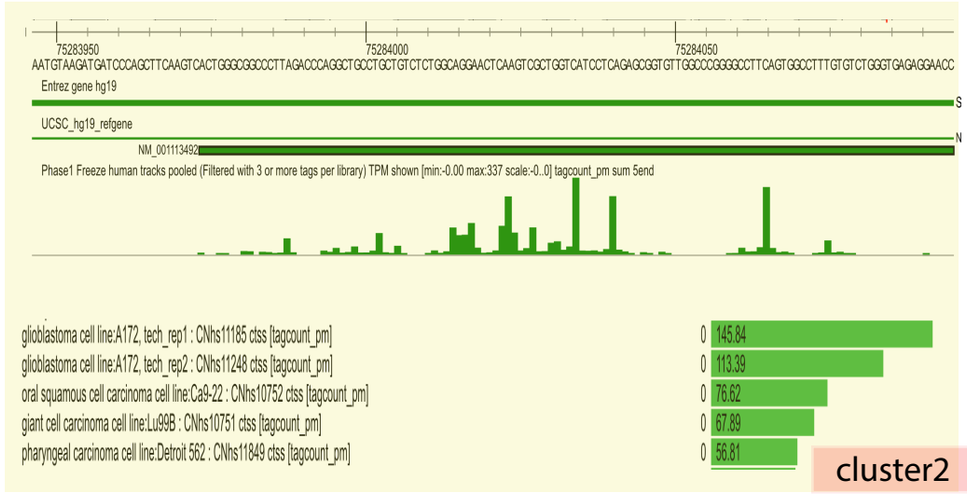
cluster3:



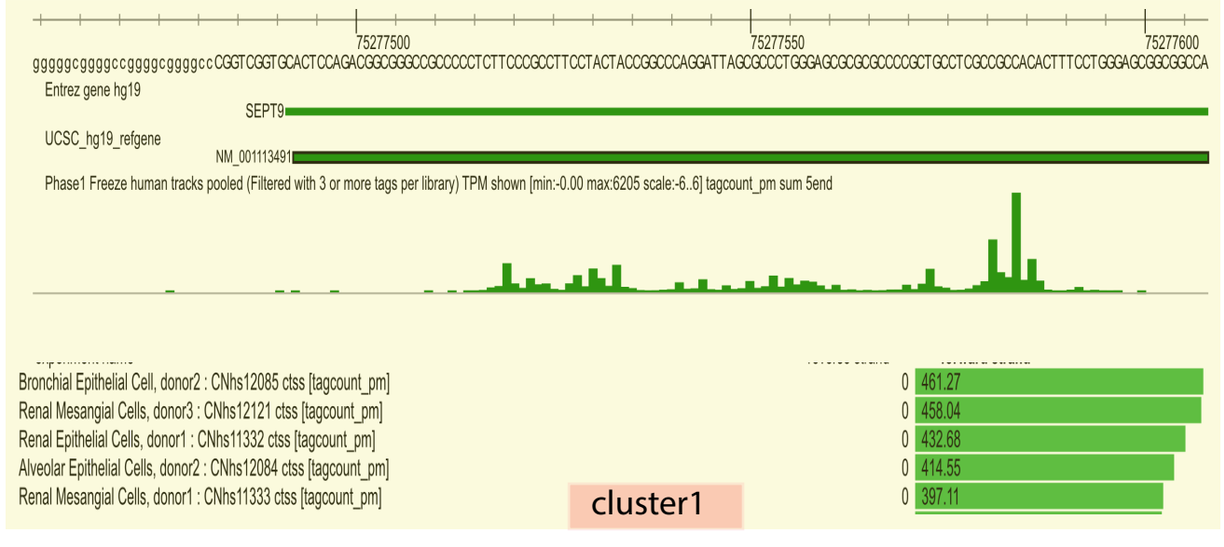
cluster4:



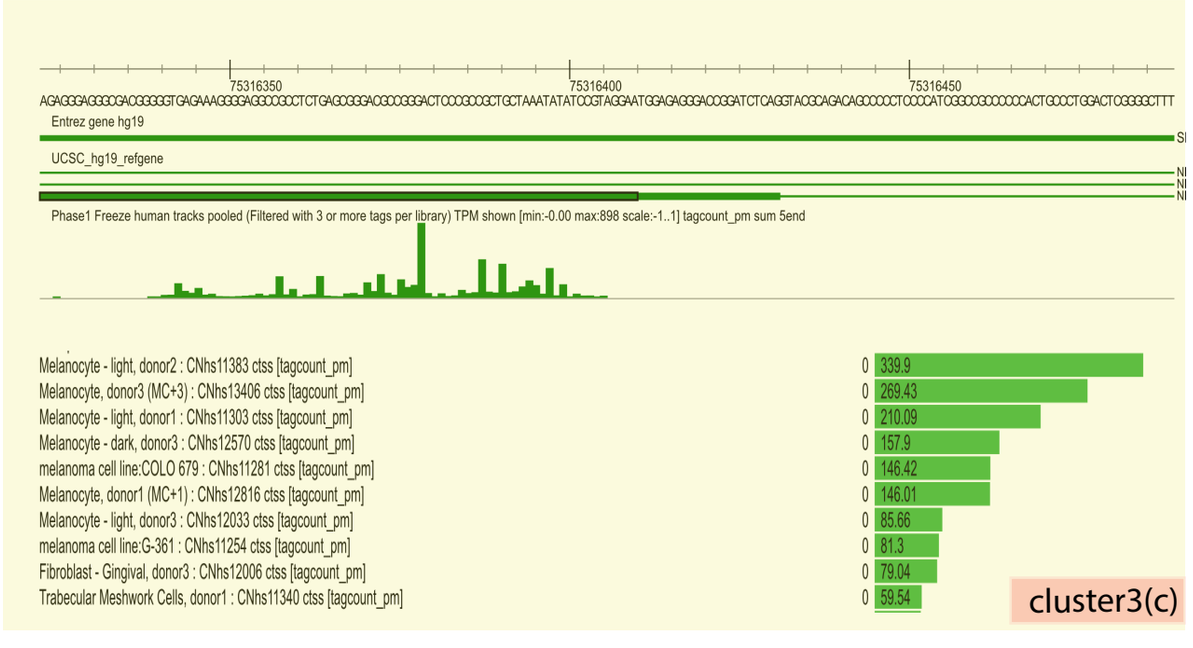
cluster4



cluster2



cluster1



cluster3(c)

Figure 8

Figure 9 (a)

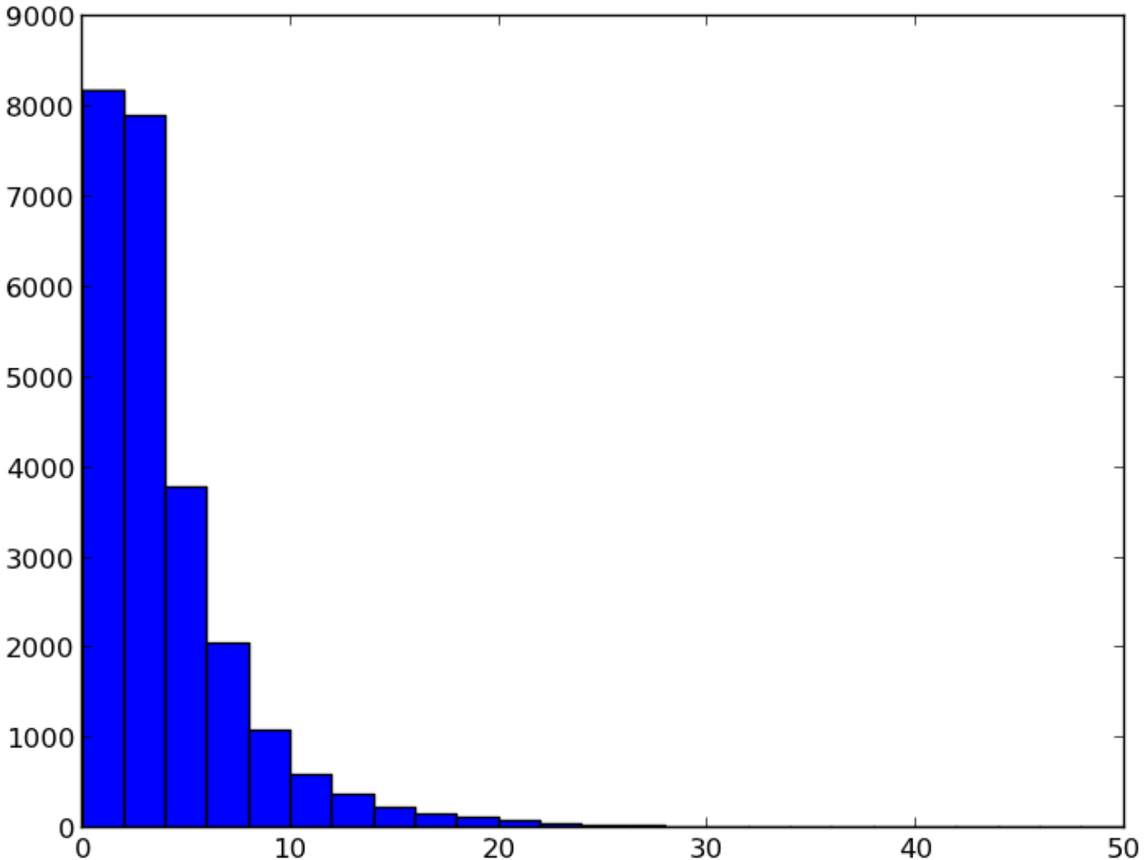


Figure 9 (b)

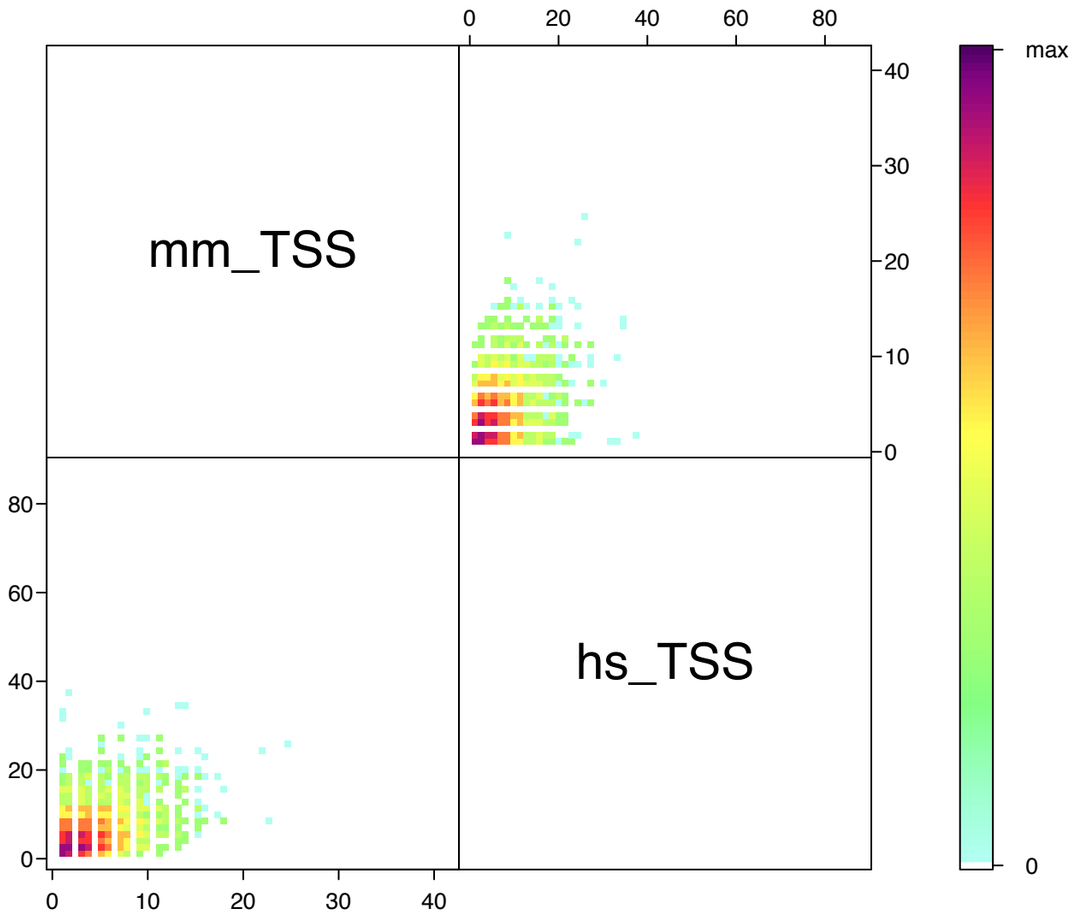
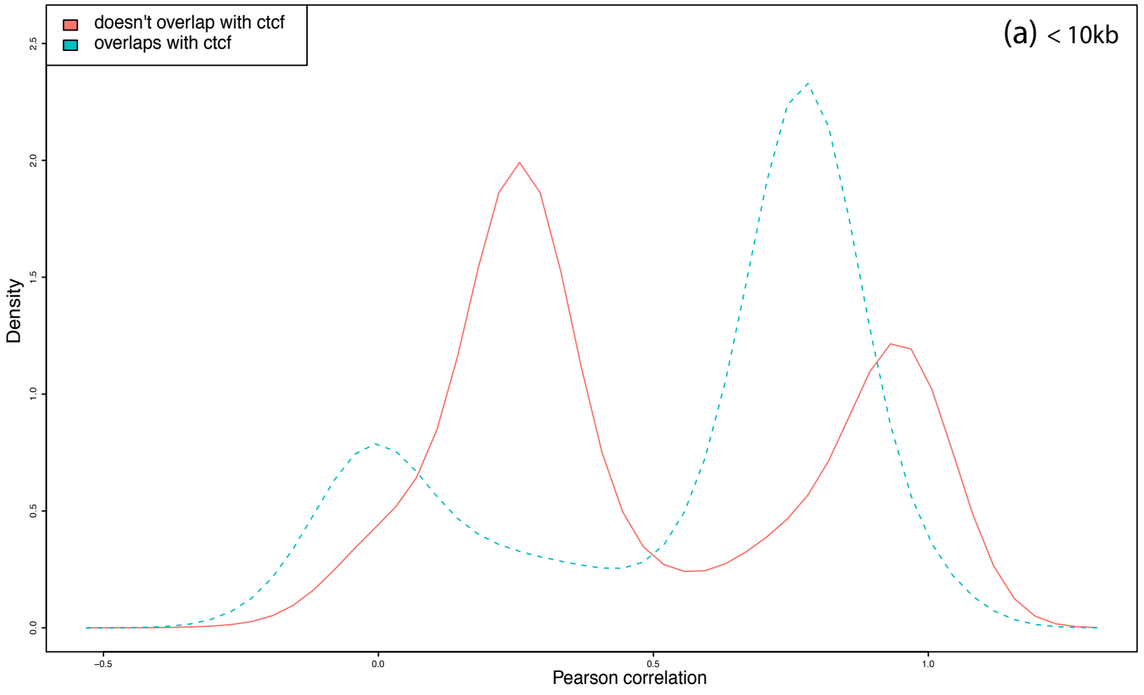
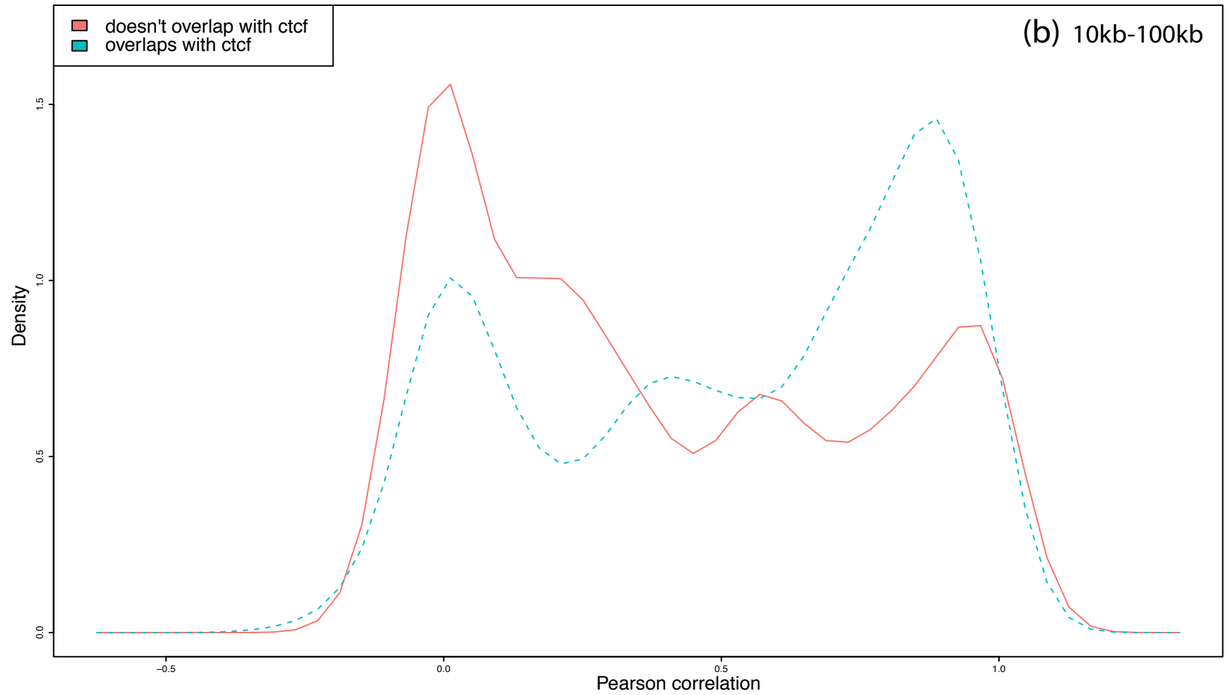
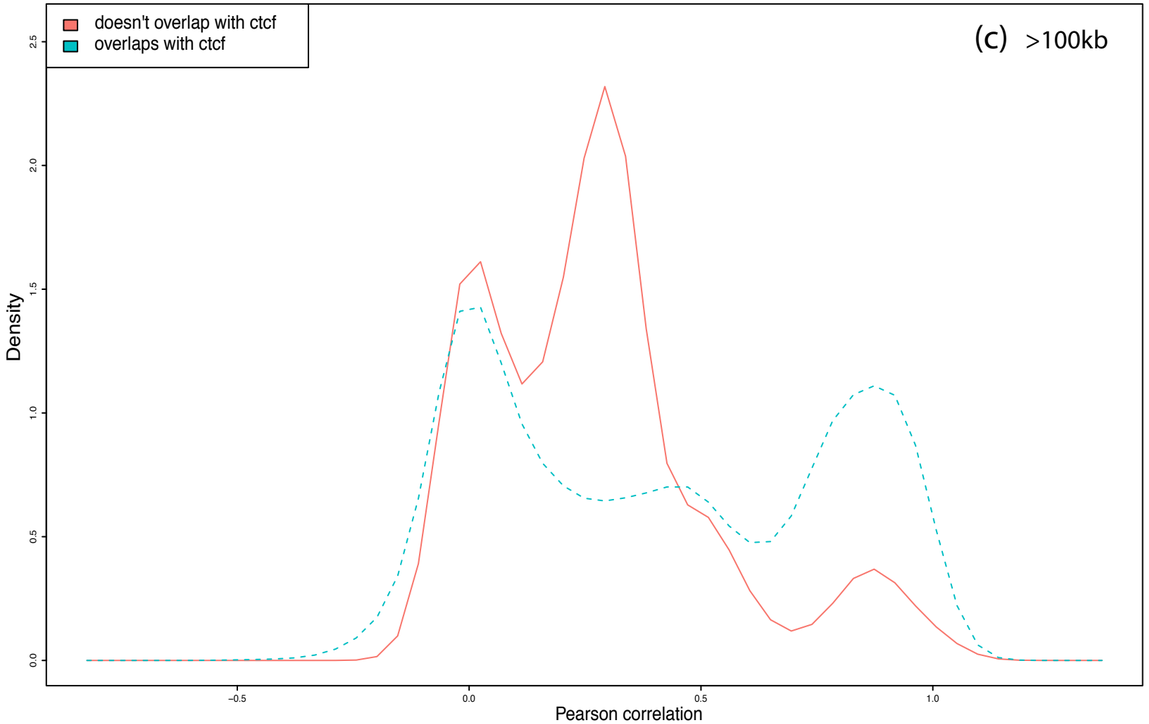


Figure 10


doesn't overlap with ctcf
overlaps with ctcf
(a) < 10kb
Density
Pearson correlation


doesn't overlap with ctcf
overlaps with ctcf
(b) 10kb-100kb
Density
Pearson correlation


doesn't overlap with ctcf
overlaps with ctcf
(c) >100kb
Density
Pearson correlation

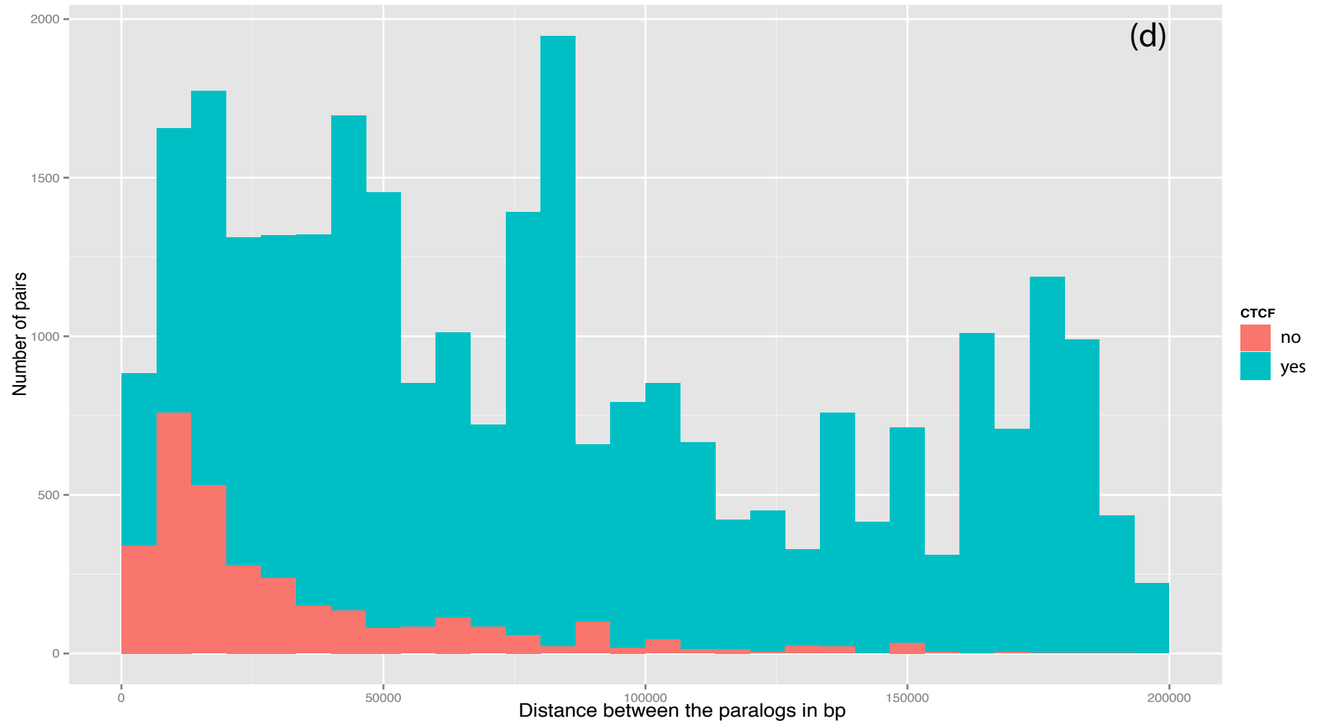

(d)
Number of pairs
Distance between the paralogs in bp
CTCF
no
yes

Figure 11

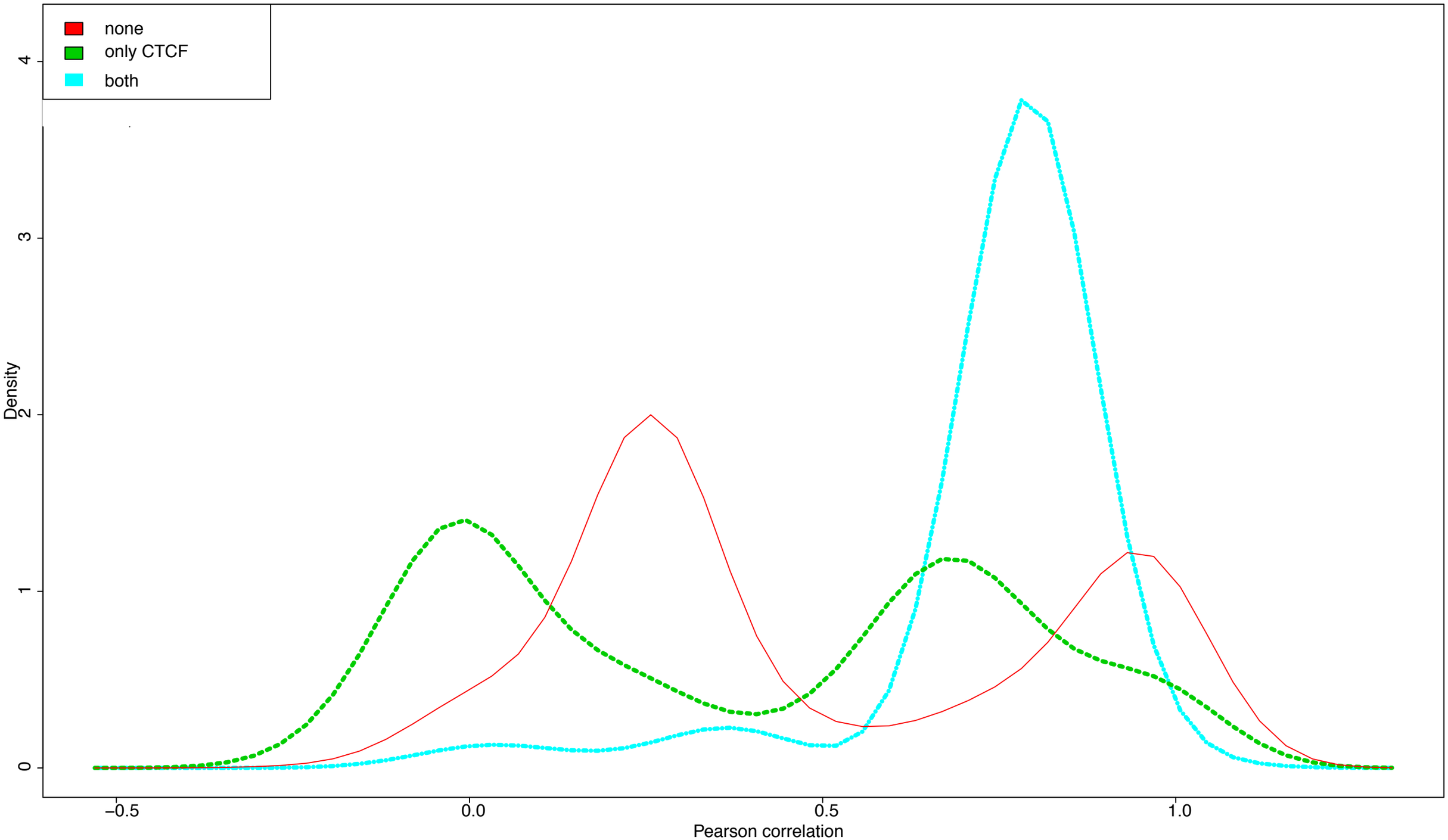

none
only CTCF
both
Density
0
1
2
3
4
−0.5
0.0
0.5
1.0
Pearson correlation

Figure S1 (a)

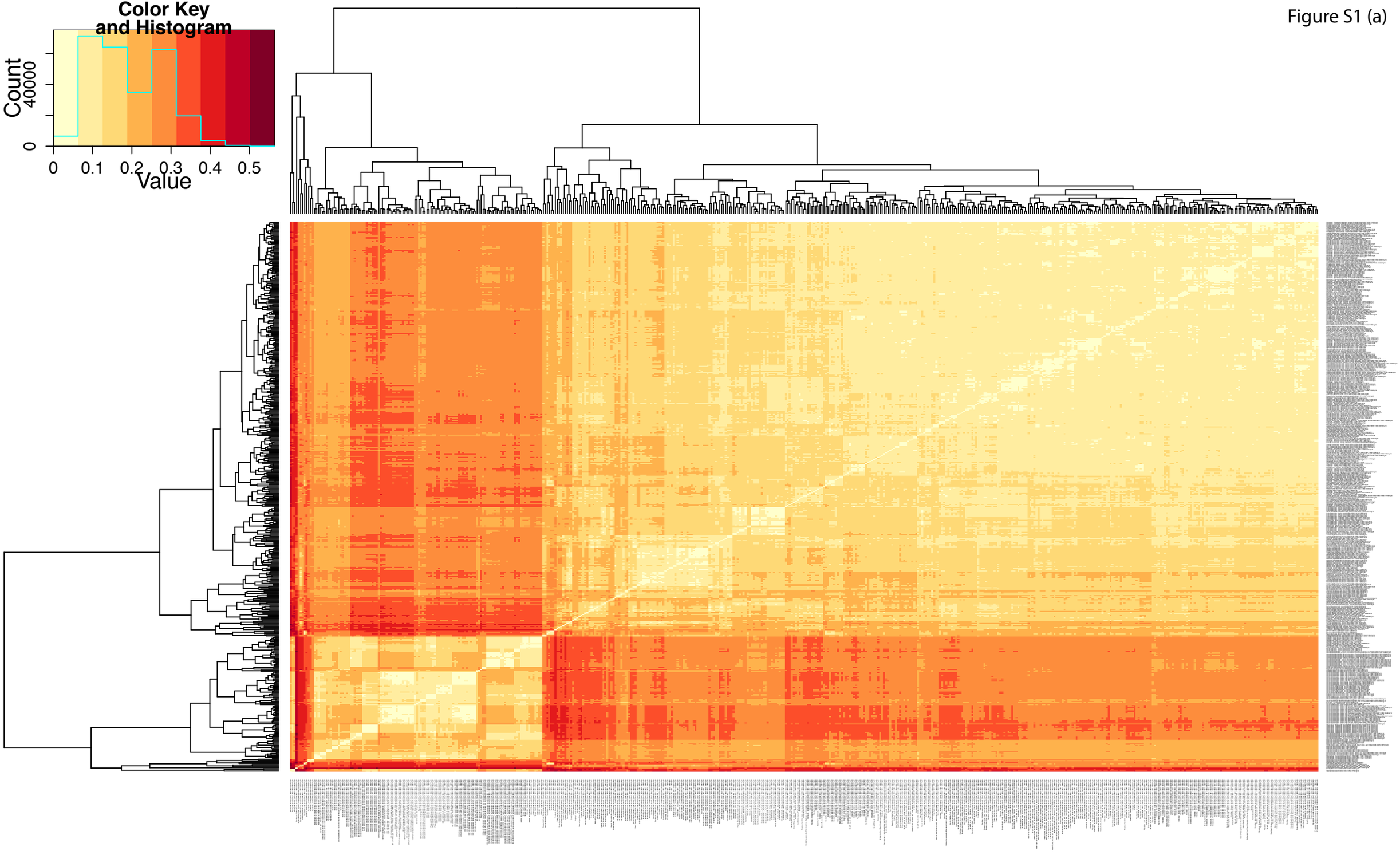


Figure S1 (b)

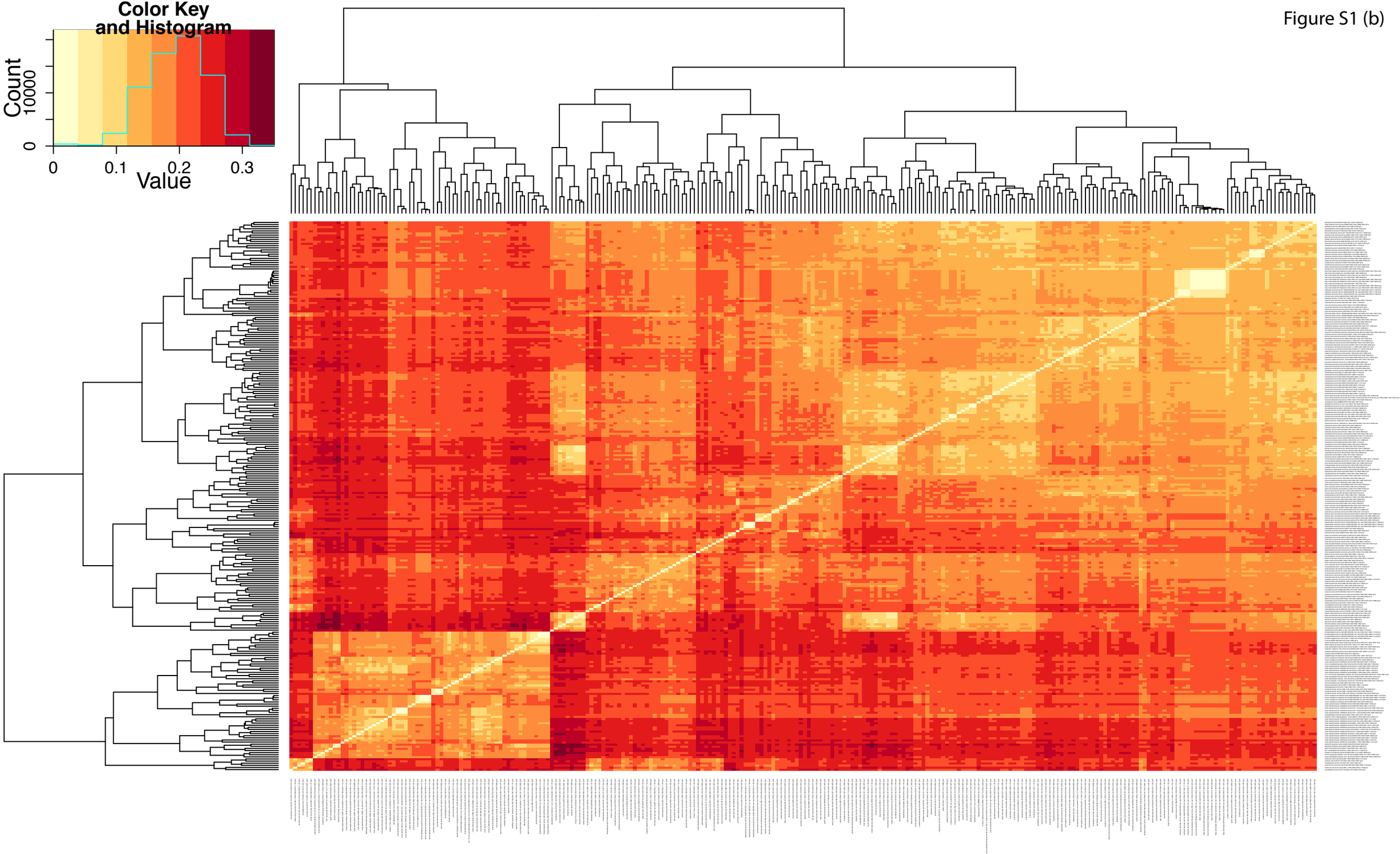


Figure S1 (c)

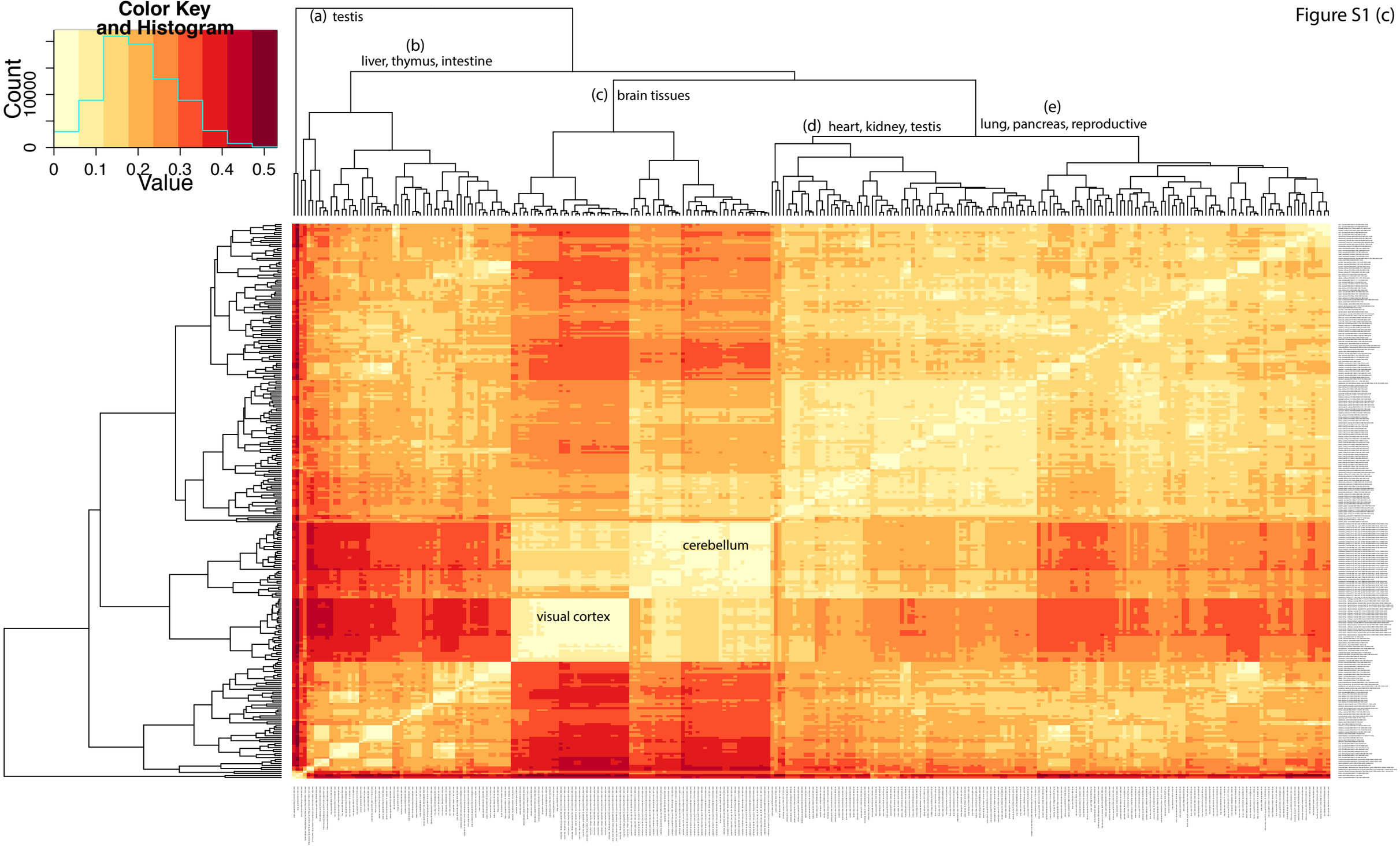

Color Key and Histogram
Count
10000
0
0
0.1
0.2
0.3
0.4
0.5
Value
(a) testis
(b)
liver, thymus, intestine
(c) brain tissues
(d) heart, kidney, testis
(e)
lung, pancreas, reproductive
cerebellum
visual cortex

Figure S1 (d)

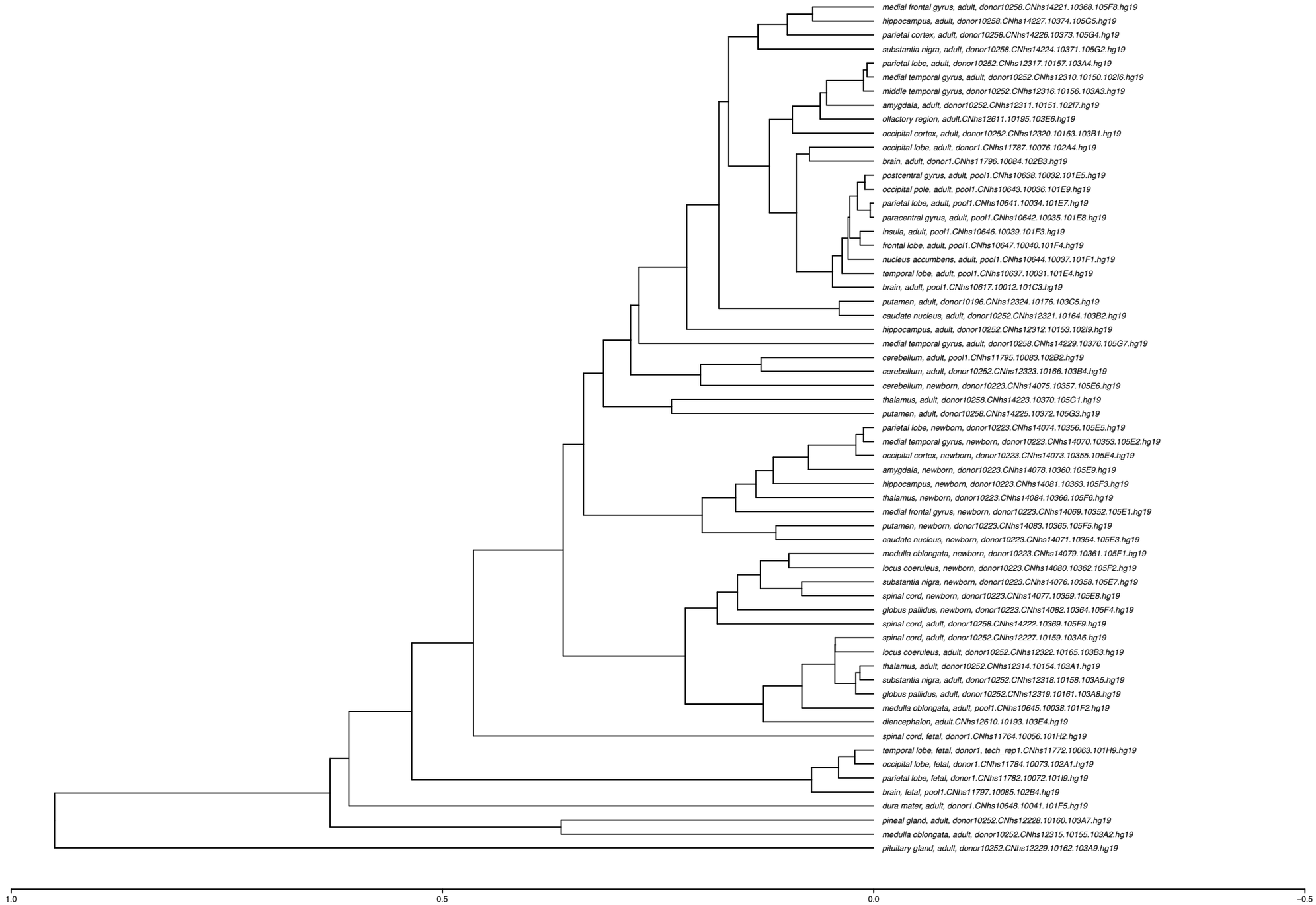

medial frontal gyrus, adult, donor10258.CNhs14221.10368.105F8.hg19
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
parietal lobe, adult, donor10252.CNhs12317.10157.103A4.hg19
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 S1 (e)

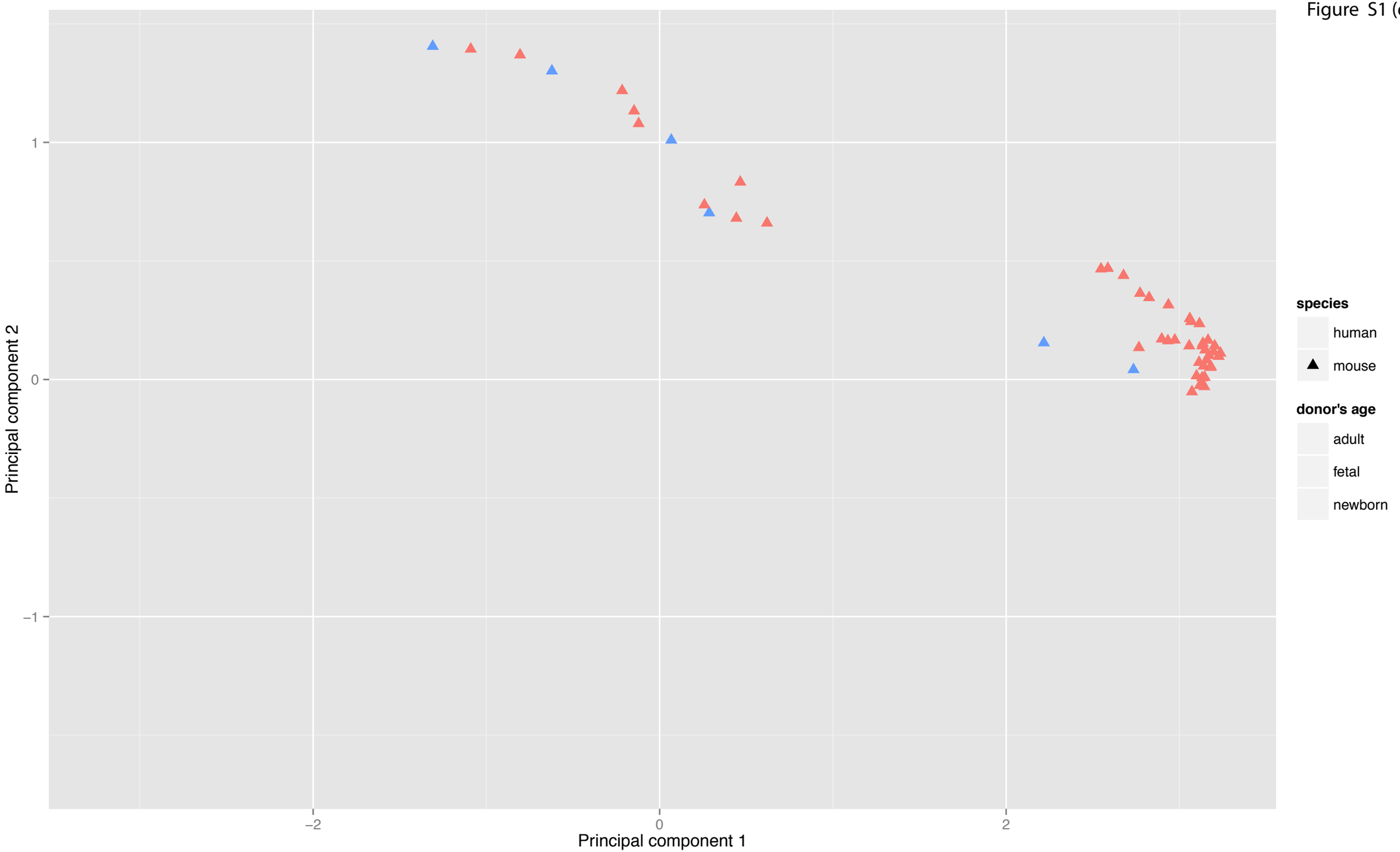

Principal component 2
1
0
−1
−2
0
2
Principal component 1
species
human
mouse
donor's age
adult
fetal
newborn

Figure S2

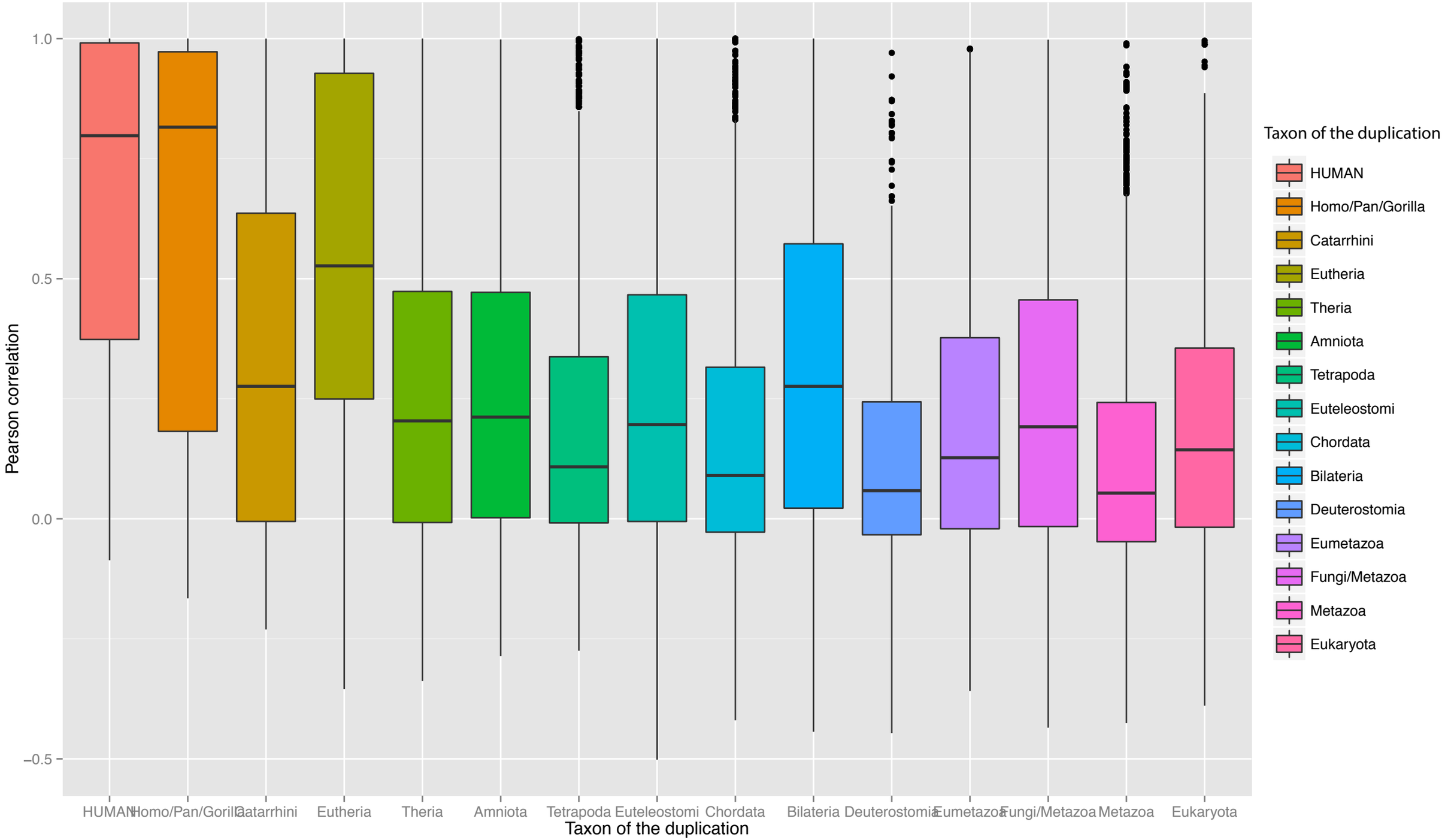

1.0
0.5
0.0
−0.5
Pearson correlation
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Bilateria
Deuterostomia
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota
Taxon of the duplication
Taxon of the duplication
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota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

Figure S3

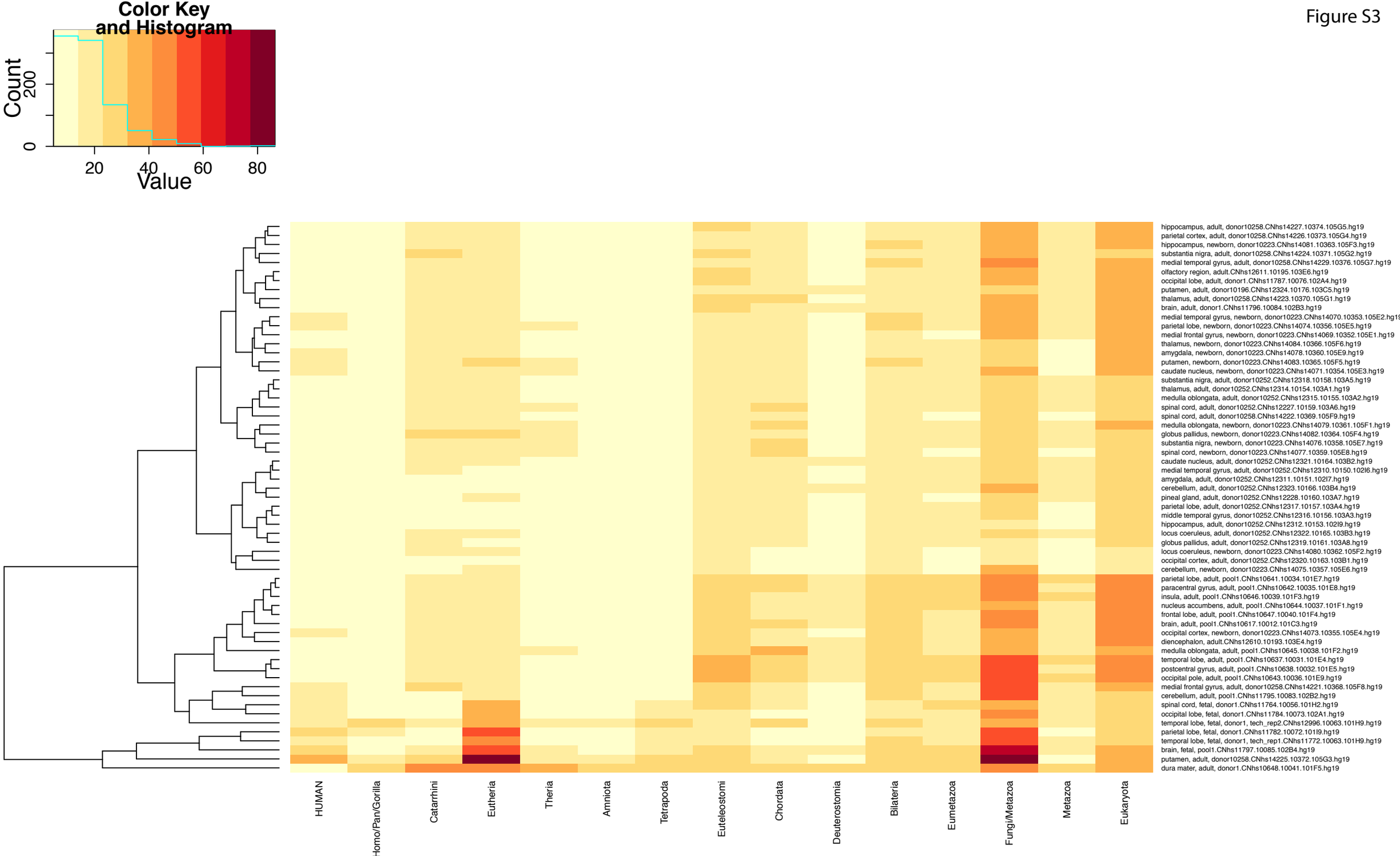

Color Key
and Histogram
Count
0
200
20
40
60
80
Value
hippocampus, adult, donor10258.CNhs14227.10374.105G5.hg19
parietal cortex, adult, donor10258.CNhs14226.10373.105G4.hg19
hippocampus, newborn, donor10223.CNhs14081.10363.105F3.hg19
substantia nigra, adult, donor10258.CNhs14224.10371.105G2.hg19
medial 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 S4

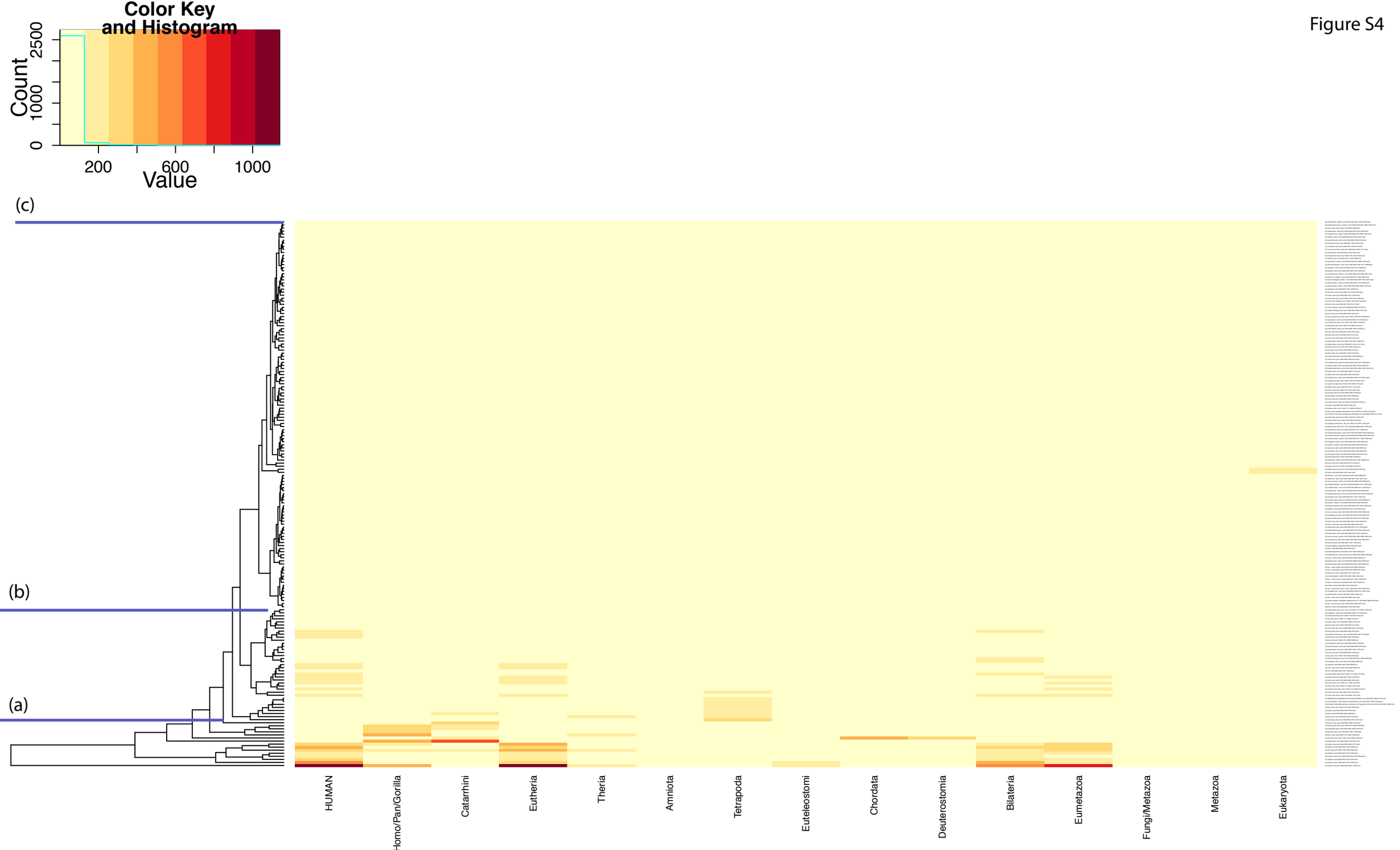

Color Key
and Histogram
Count
0
1000
2500
200
600
1000
Value
(c)
(b)
(a)
HUMAN
Homo/Pan/Gorilla
Catarrhini
Eutheria
Theria
Amniota
Tetrapoda
Euteleostomi
Chordata
Deuterostomia
Bilateria
Eumetazoa
Fungi/Metazoa
Metazoa
Eukaryota

Figure S5

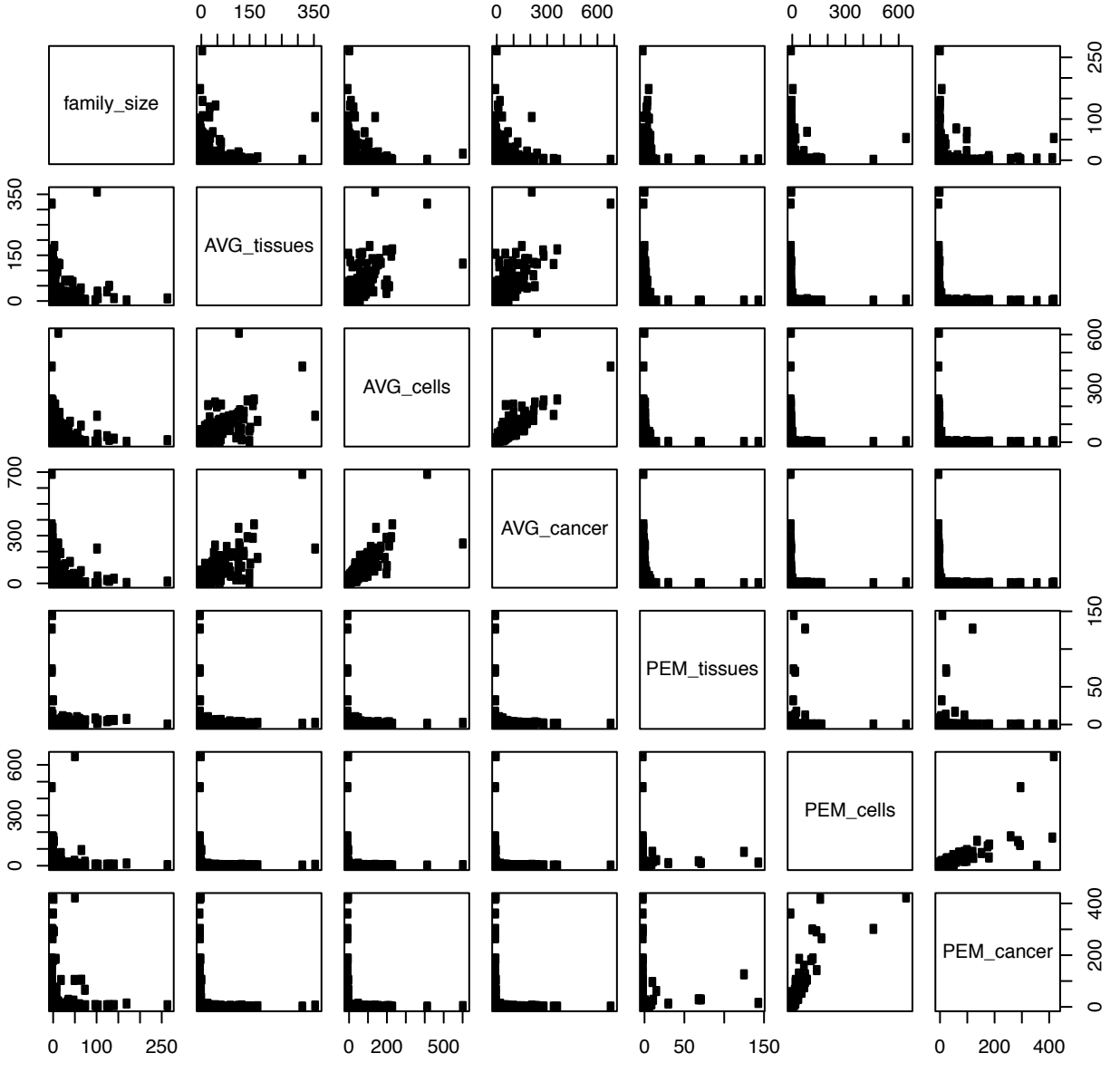

family_size
AVG_tissues
AVG_cells
AVG_cancer
PEM_tissues
PEM_cells
PEM_cancer
0 150 350
0 300 600
0 300 600
0 100 250
0 150 350
0 300 600
0 300 700
0 50 150
0 300 600
0 200 400
0 100 250
0 200 500
0 50 150
0 200 400

Figure 6 (a)

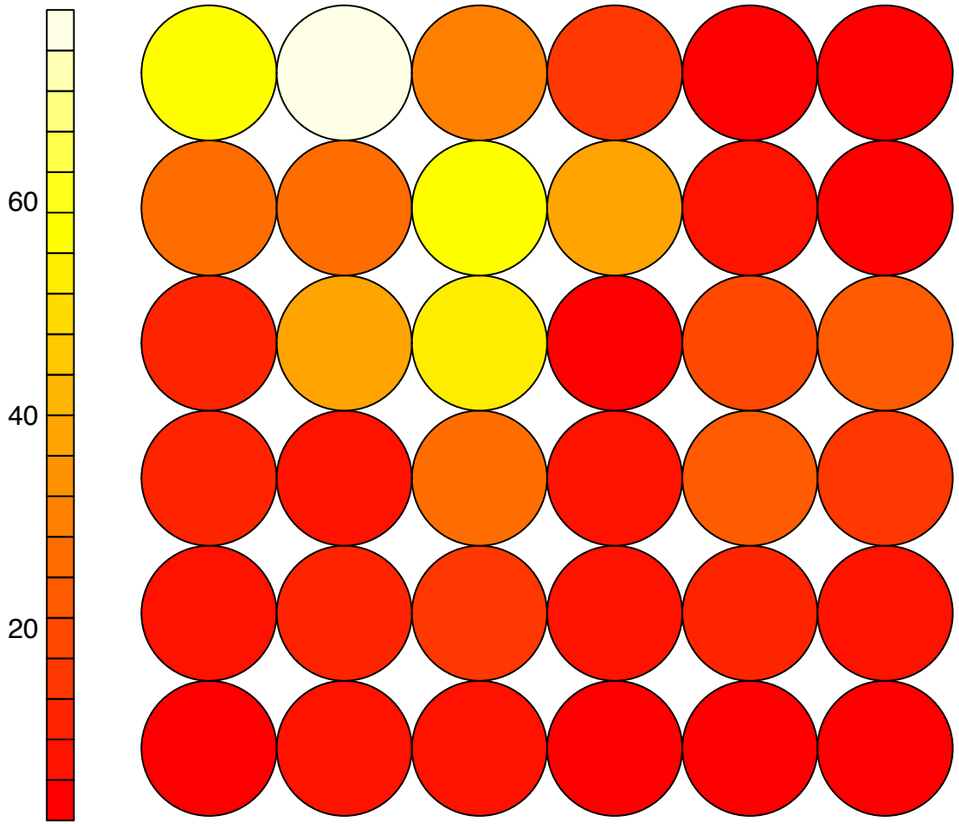
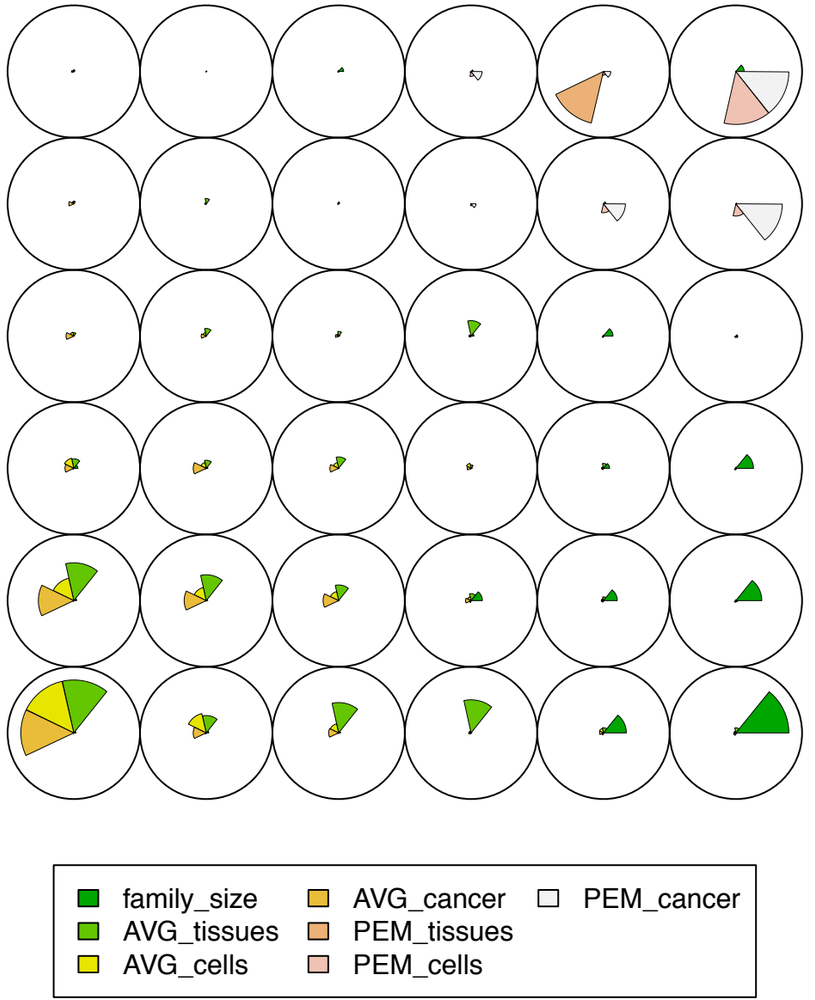
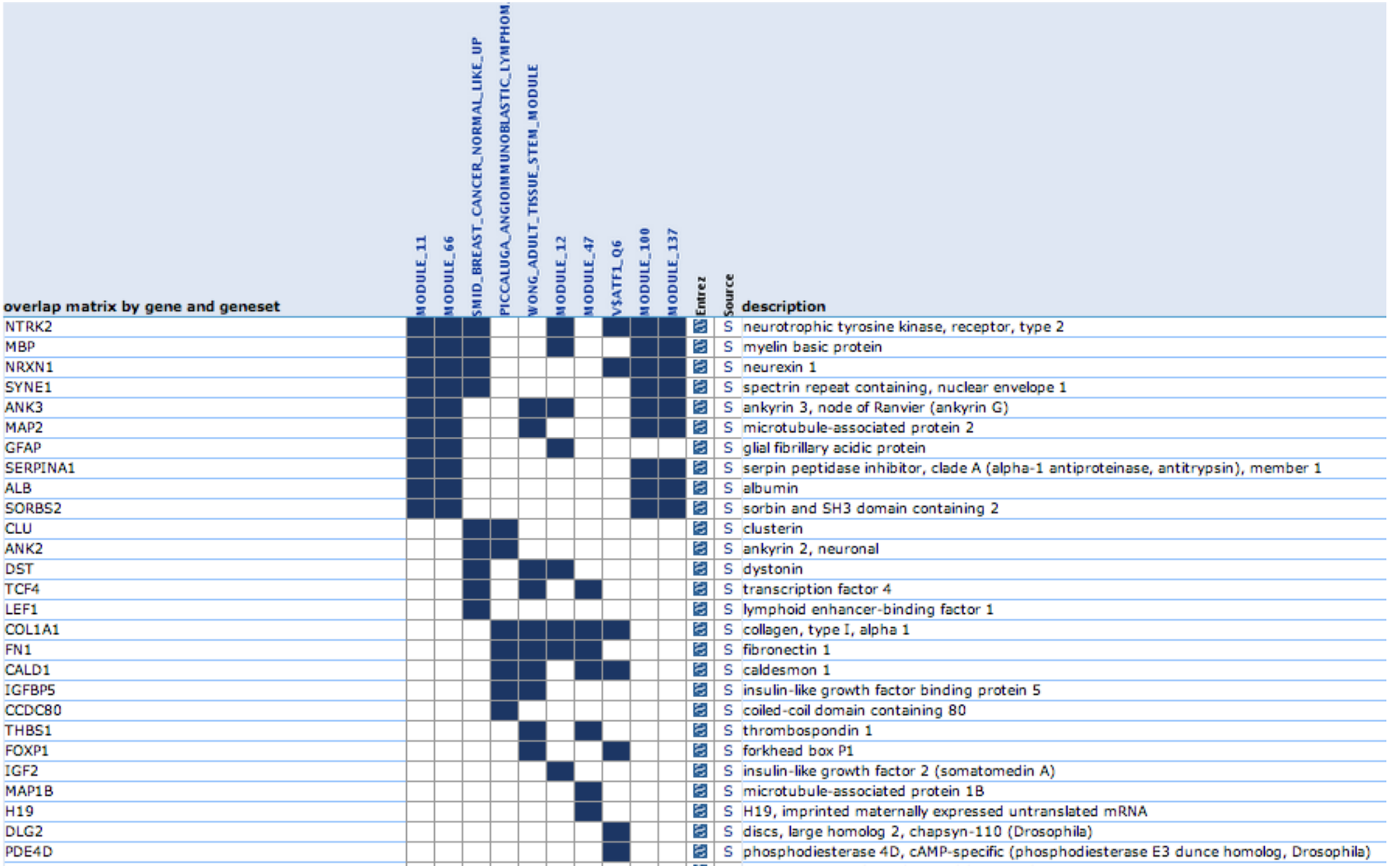


Figure 6 (b)



Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p value
MODULE_11 [540]	Genes in the cancer module 11	10		$2.27\ e^{-10}$
MODULE_66 [552]	Genes in the cancer module 66	10		$2.81\ e^{-10}$
SMID_BREAST_CANCER_NORMAL_LIKE_UP [476]	Genes up-regulated in the normal-like subtype of breast cancer.	9		$1.73\ e^{-9}$
PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOMA_MA_UP [205]	Up-regulated genes in angioimmunoblastic lymphoma (AILT) compared to normal T lymphocytes.	7		$2.38\ e^{-9}$
WONG_ADULT_TISSUE_STEM_MODULE [721]	The 'adult tissue stem' module: genes coordinately up-regulated in a compendium of adult tissue stem cells.	10		$3.66\ e^{-9}$
MODULE_12 [360]	Genes in the cancer module 12	8		$4.45\ e^{-9}$
MODULE_47 [225]	Genes in the cancer module 47	7		$4.55\ e^{-9}$
V$ATF1_Q6 [231]	Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif CYYTGACGTCA which matches annotation for ATF1: activating transcription factor 1	7		$5.46\ e^{-9}$
MODULE_100 [544]	Genes in the cancer module 100	9		$5.51\ e^{-9}$
MODULE_137 [546]	Genes in the cancer module 137	9		$5.69\ e^{-9}$



overlap matrix by gene and geneset	MODULE_11	MODULE_66	SMID_BREAST_CANCER_NORMAL_LIKE_UP	PICCALUGA_ANGIOIMMUNOBLASTIC_LYMPHOM	WONG_ADULT_TISSUE_STEM_MODULE	MODULE_12	MODULE_47	V$ATF1_Q6	MODULE_100	MODULE_137	Entrez	Source	description
NTRK2	■	■	■			■		■	■	■		S	neurotrophic tyrosine kinase, receptor, type 2
MBP	■	■	■			■			■	■		S	myelin basic protein
NRXN1	■	■	■					■	■	■		S	neurexin 1
SYNE1	■	■	■						■	■		S	spectrin repeat containing, nuclear envelope 1
ANK3	■	■			■	■			■	■		S	ankyrin 3, node of Ranvier (ankyrin G)
MAP2	■	■			■				■	■		S	microtubule-associated protein 2
GFAP	■	■				■						S	glial fibrillary acidic protein
SERPINA1	■	■							■	■		S	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
ALB	■	■							■	■		S	albumin
SORBS2	■	■							■	■		S	sorbin and SH3 domain containing 2
CLU			■	■								S	clusterin
ANK2			■	■								S	ankyrin 2, neuronal
DST			■		■	■						S	dystonin
TCF4			■		■		■					S	transcription factor 4
LEF1			■									S	lymphoid enhancer-binding factor 1
COL1A1				■	■	■	■	■				S	collagen, type I, alpha 1
FN1				■	■	■	■					S	fibronectin 1
CALD1				■	■		■	■				S	caldesmon 1
IGFBP5				■	■							S	insulin-like growth factor binding protein 5
CCDC80				■								S	coiled-coil domain containing 80
THBS1					■		■					S	thrombospondin 1
FOXP1					■			■				S	forkhead box P1
IGF2						■						S	insulin-like growth factor 2 (somatomedin A)
MAP1B							■					S	microtubule-associated protein 1B
H19							■					S	H19, imprinted maternally expressed untranslated mRNA
DLG2								■				S	discs, large homolog 2, chapsyn-110 (Drosophila)
PDE4D								■				S	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p value
WANG_SMARCE1_TARGETS_UP [280]	Genes up-regulated in BT549 cells (breast cancer) by expression of SMARCE1 [GeneID=6605] off a retroviral vector.	6		$5\ e^{-9}$
TTGTTT_V$FOXO4_01 [2061]	Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif TTGTTT which matches annotation for MLLT7: myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 7	10		$1.61\ e^{-8}$
MODULE_12 [360]	Genes in the cancer module 12	6		$2.23\ e^{-8}$
WILCOX_PRESPONSE_TO_ROGESTERONE_DN [66]	Genes down-regulated in primary cultures of ovarian surface epithlium cells exposed to progesterone [PubChem=5994] for 5 days.	4		$3.92\ e^{-8}$
TGCCAAR_V$NF1_Q6 [722]	Genes with promoter regions [-2kb,2kb] around transcription start site containing the motif TGCCAAR which matches annotation for NF1: neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)	7		$4.77\ e^{-8}$
GU_PDEF_TARGETS_UP [71]	Integrin, VEGF, Wnt and TGFbeta signaling pathway genes up-regulated in PC-3 cells (prostate cancer) after knockdown of PDEF [GeneID=25803] by RNAi.	4		$5.28\ e^{-8}$
MODULE_47 [225]	Genes in the cancer module 47	5		$9.72\ e^{-8}$
PILON_KLF1_TARGETS_UP [504]	Genes up-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 [GeneID=10661] knockout compared to those from the wild type embryos.	6		$1.63\ e^{-7}$
MODULE_23 [565]	Genes in the cancer module 23	6		$3.18\ e^{-7}$
HOOI_ST7_TARGETS_DN [123]	Genes down-regulated in PC-3 cells (prostate cancer) stably expressing ST7 [GeneID=7982] off a plasmid vector.	4		$4.85\ e^{-7}$

overlap matrix by gene and geneset	WANG_SMARCE1_TARGETS_UP	TTGTTT_V$FOX04_01	MODULE_12	WILCOX_PRESPONSE_TO_ROGESTERONE_DN	TGCCAAR_V$NF1_Q6	GU_PDEF_TARGETS_UP	MODULE_47	PILON_KLF1_TARGETS_UP	MODULE_23	HOOI_ST7_TARGETS_DN	Entrez	Source	description
COL1A1	■	■	■	■		■	■	■	■	■		S	collagen, type I, alpha 1
FN1	■	■	■	■		■	■	■	■	■		S	fibronectin 1
NTRK2	■	■	■									S	neurotrophic tyrosine kinase, receptor, type 2
MAP1B	■	■					■					S	microtubule-associated protein 1B
IGFBP5	■			■						■		S	insulin-like growth factor binding protein 5
CALD1	■				■		■	■				S	caldesmon 1
DST		■	■	■	■							S	dystonin
TCF4		■			■	■	■					S	transcription factor 4
ALB		■			■			■	■			S	albumin
SORBS2		■			■				■			S	sorbin and SH3 domain containing 2
TTN		■								■		S	titin
PDE4D		■										S	phosphodiesterase 4D, cAMP-specific (phosphodiesterase E3 dunce homolog, Drosophila)
GFAP			■		■							S	glial fibrillary acidic protein
IGF2			■					■	■			S	insulin-like growth factor 2 (somatomedin A)
LEF1					■	■						S	lymphoid enhancer-binding factor 1
APOB								■	■			S	apolipoprotein B (including Ag(x) antigen)
AKAP13												S	A kinase (PRKA) anchor protein 13

Figure S11

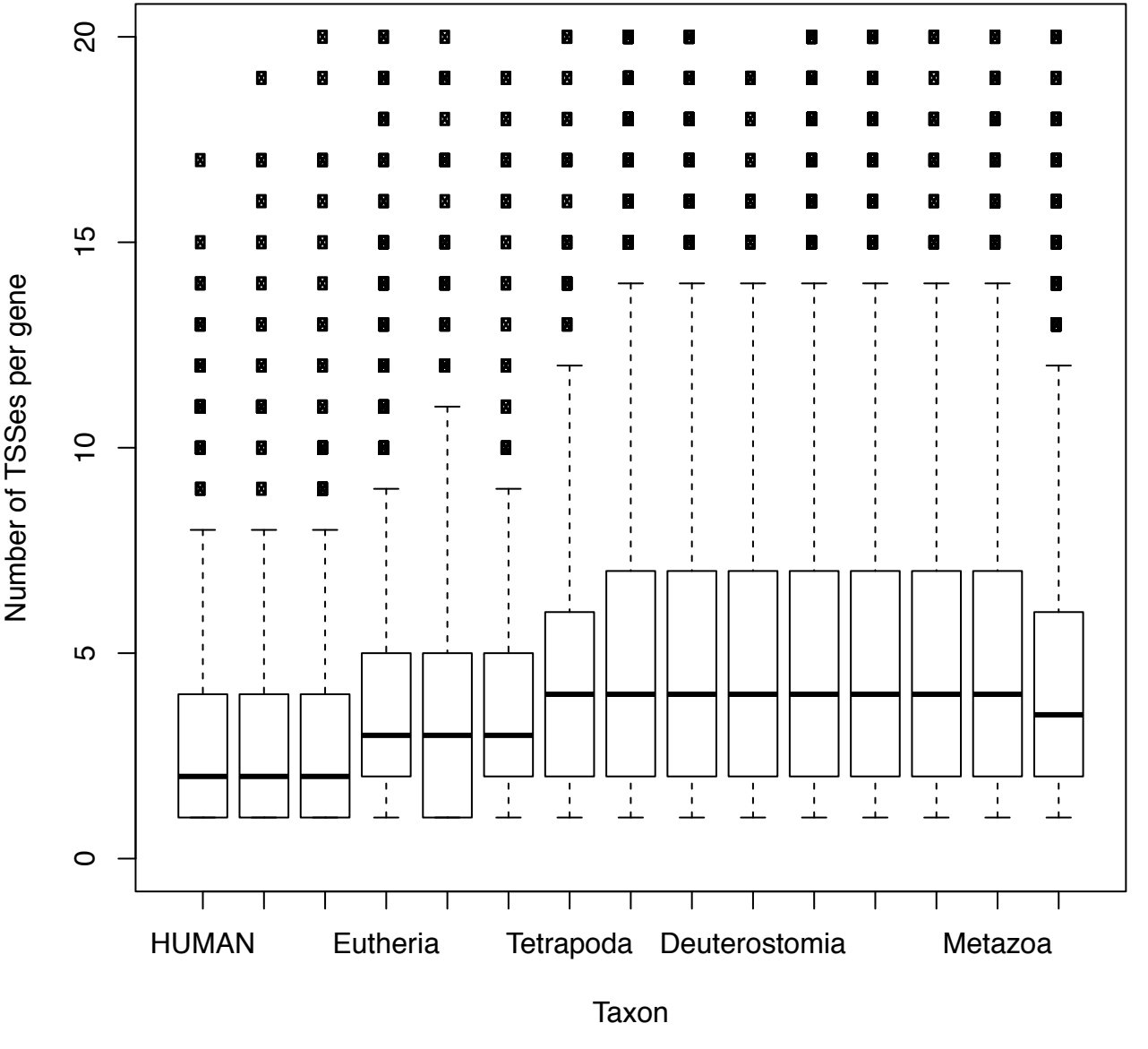

Number of TSSes per gene
0
5
10
15
20
HUMAN
Eutheria
Tetrapoda
Deuterostomia
Metazoa
Taxon

Figure S12

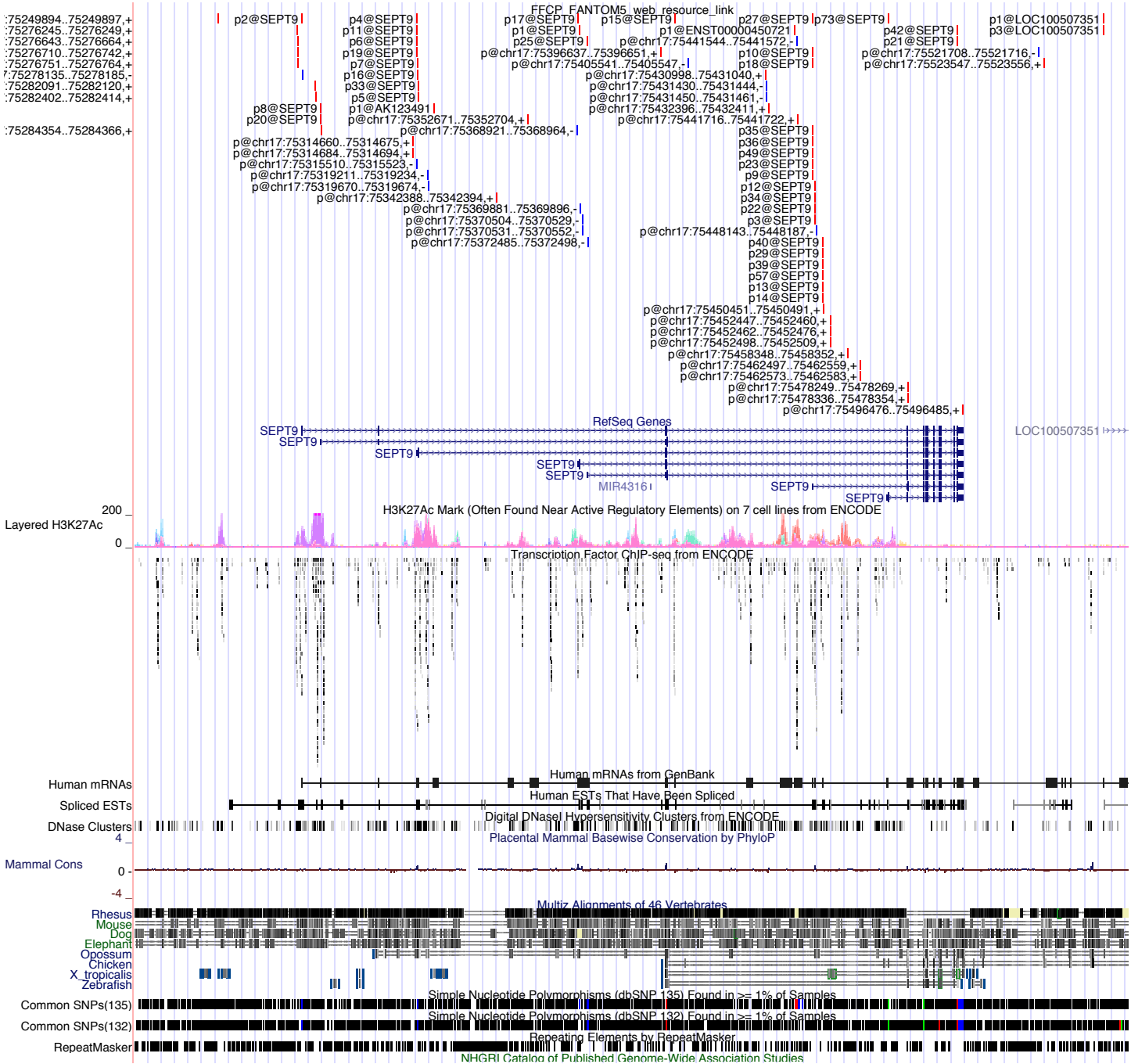

FFCP_FANTOM5_web_resource_link
RefSeq Genes
SEPT9
LOC100507351
MIR4316
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
Transcription Factor ChIP-seq from ENCODE
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
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
NHGRI Catalog of Published Genome-Wide Association Studies

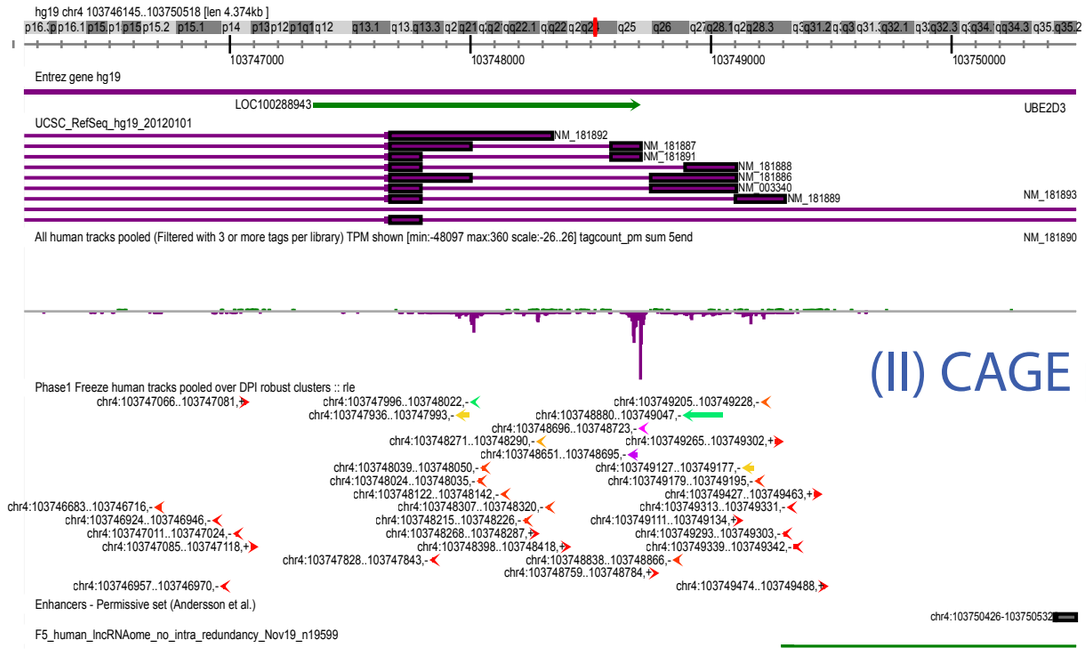
Figure S13

(a)

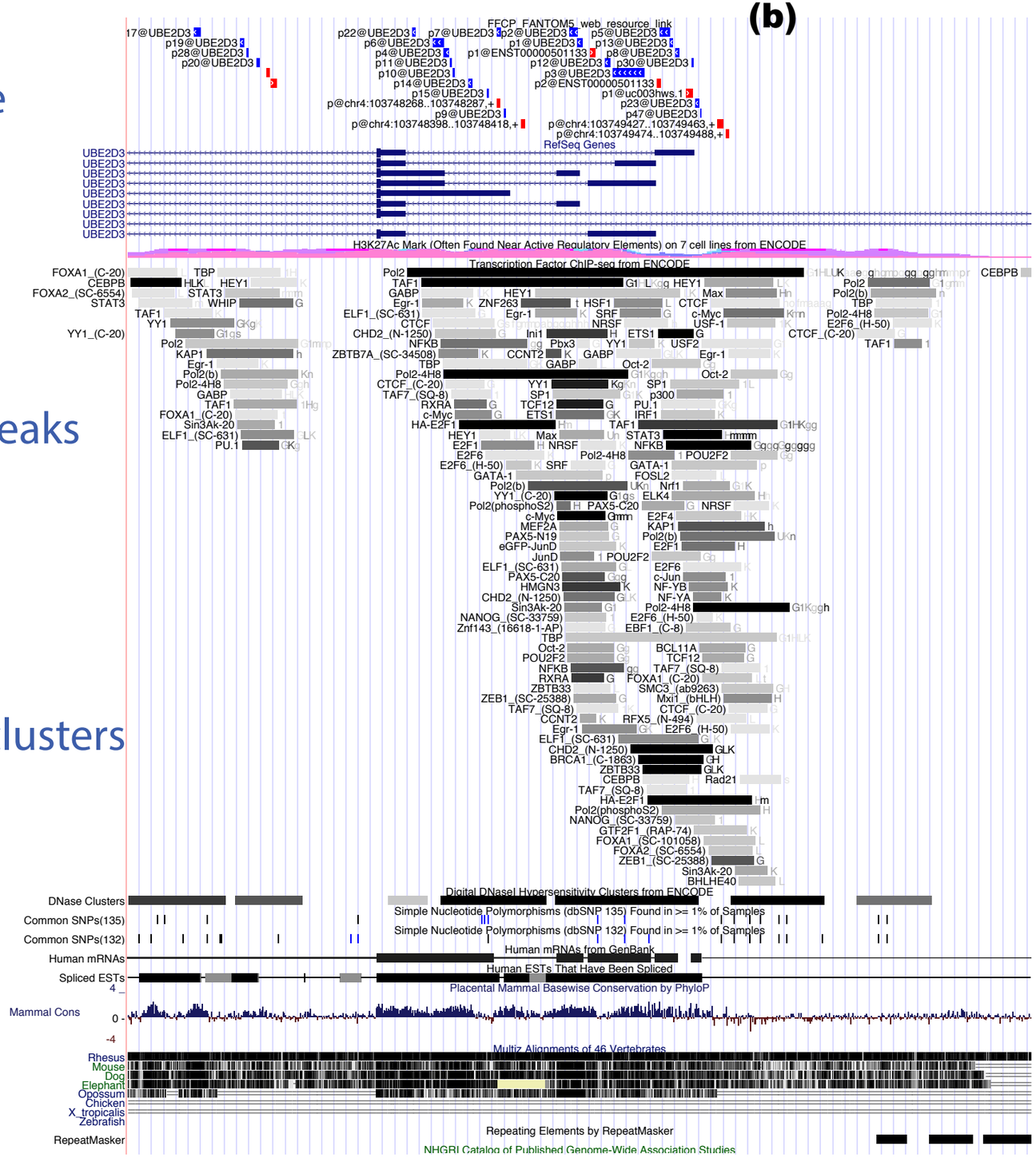
(I) Genome

(II) CAGE peaks

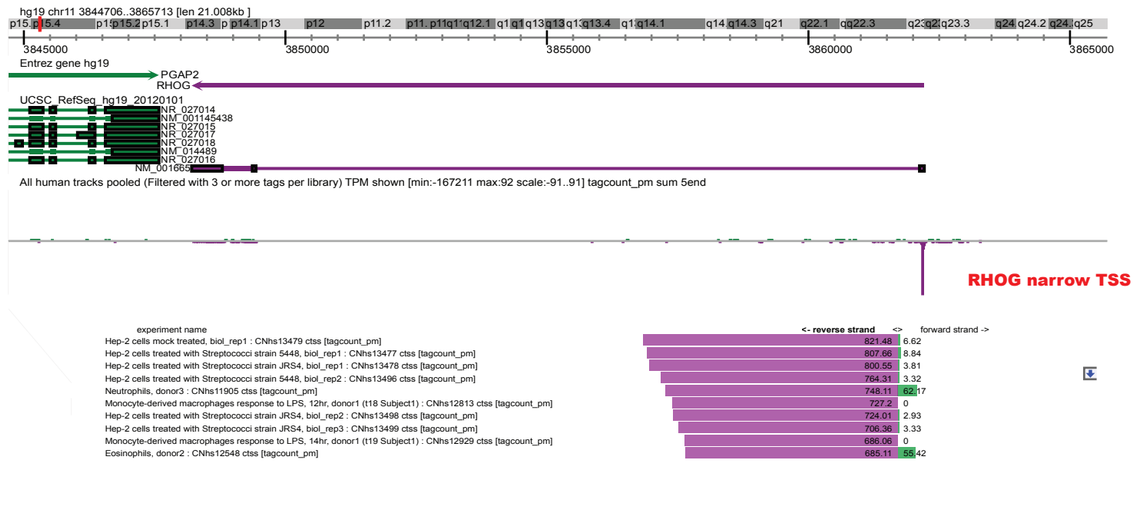
(III) CAGE clusters



(b)

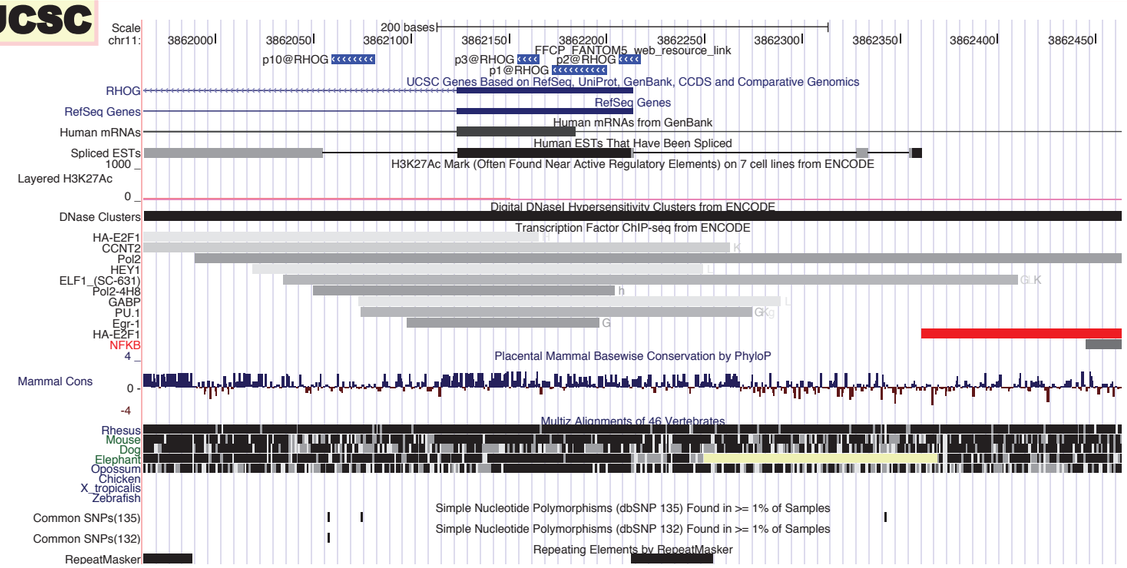


(I) RHOG

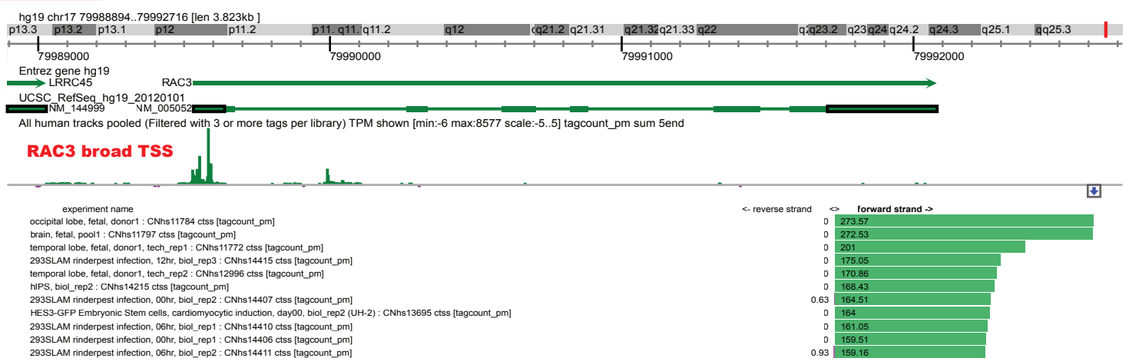


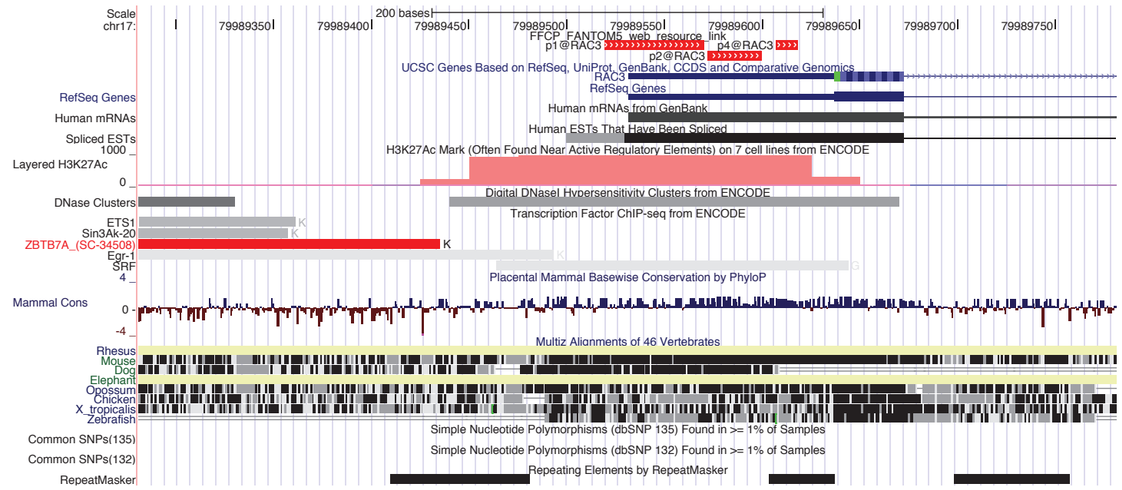
F5-ZENBU

UCSC

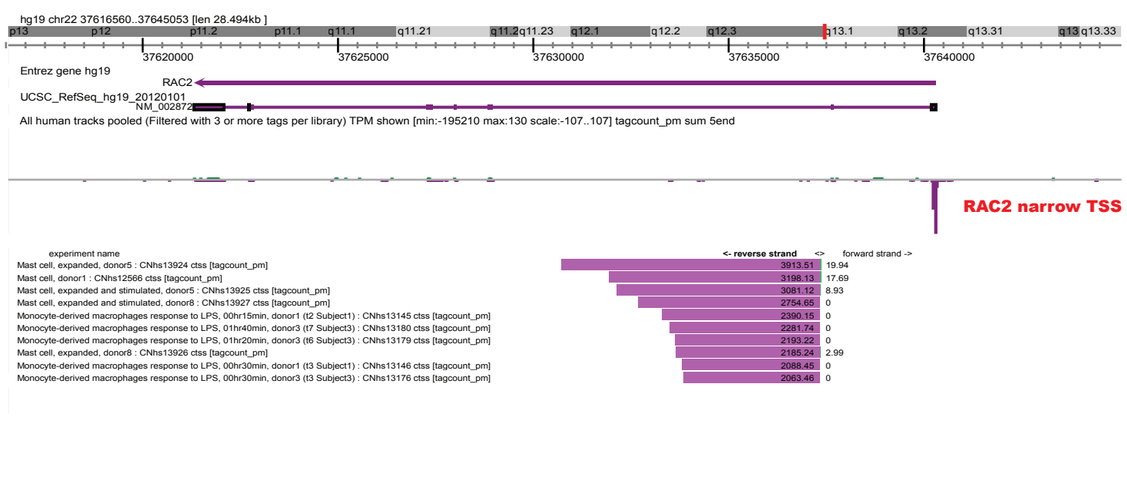


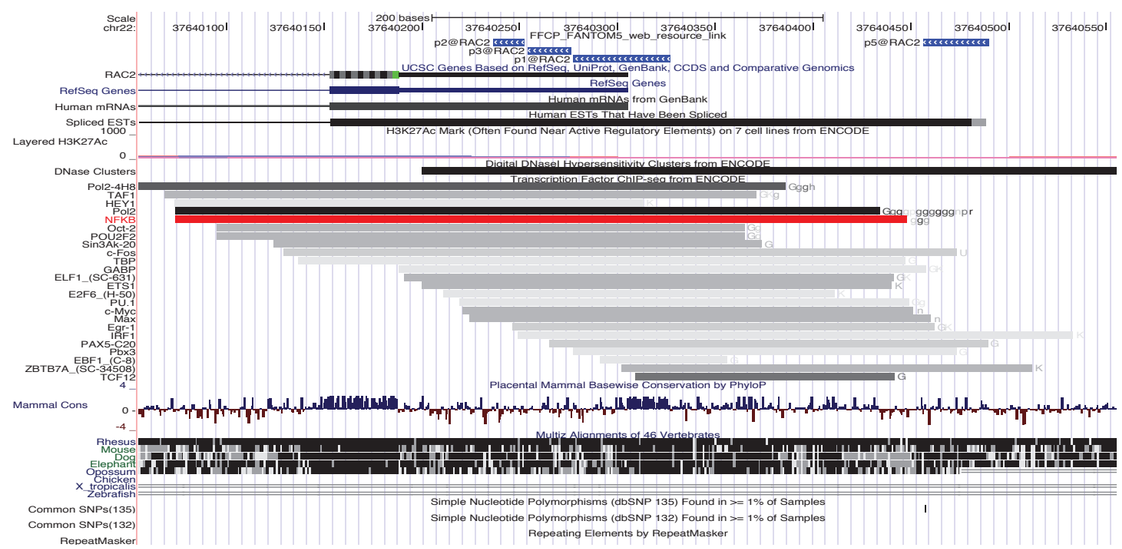
(II) RAC3

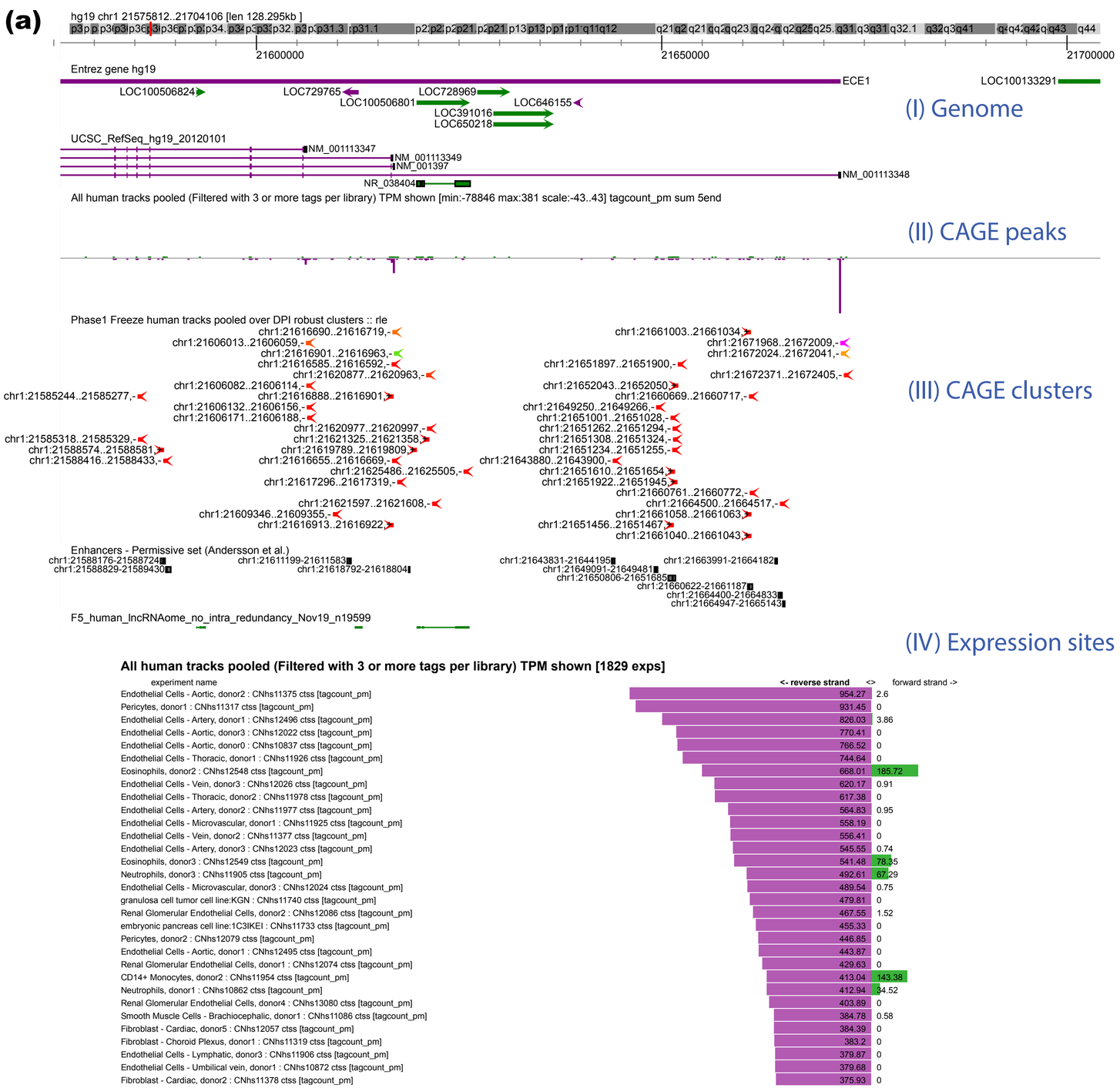


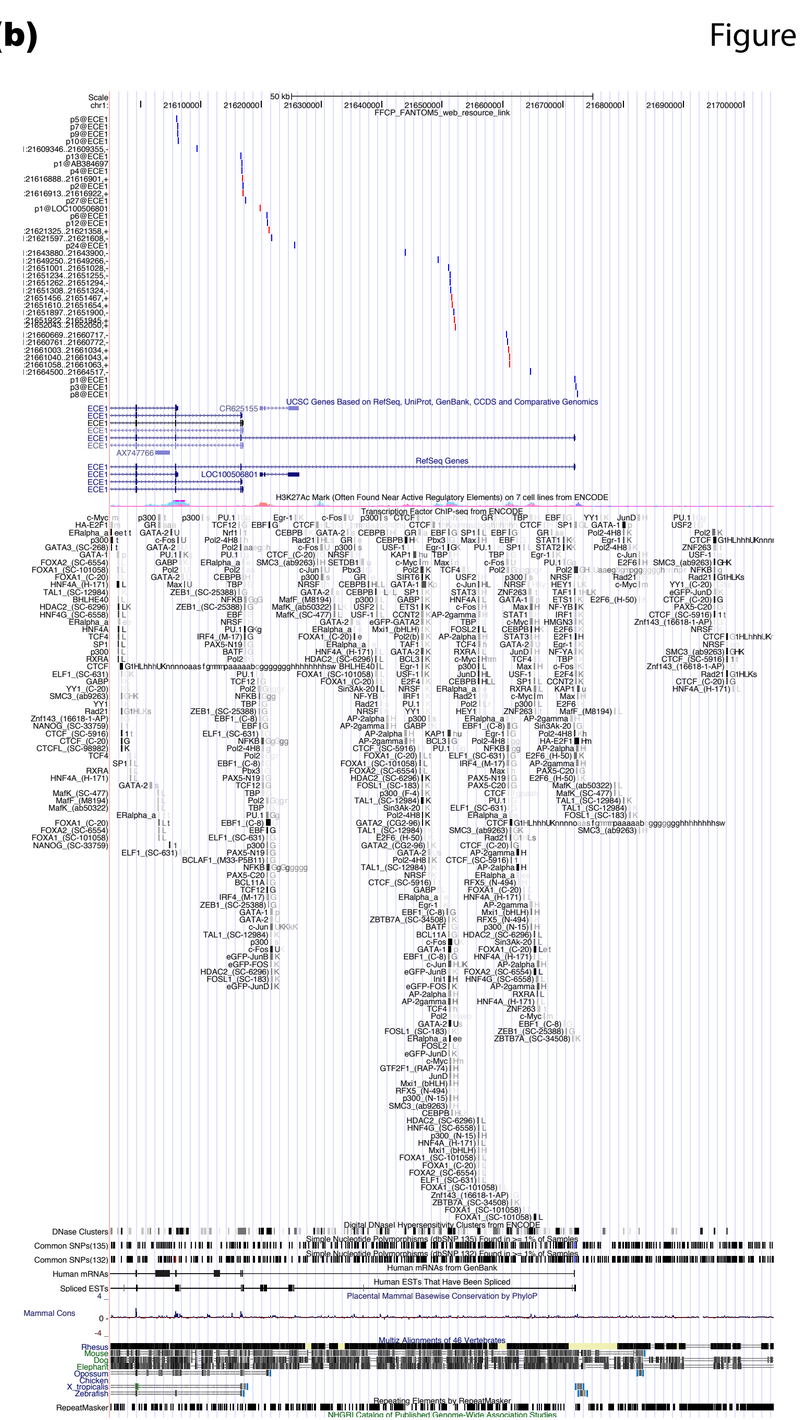


(III) RAC2






(a)
hg19 chr1 21575812..21704106 [len 128.295kb]
21600000
21650000
21700000
Entrez gene hg19
ECE1
LOC100133291
LOC100506824
LOC729765
LOC728969
LOC100506801
LOC646155
LOC391016
LOC650218
(I) Genome
UCSC_RefSeq_hg19_20120101
NM_001113347
NM_001113349
NM_001397
NM_001113348
NR_038404
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [min:-78846 max:381 scale:-43..43] tagcount_pm sum 5end
(II) CAGE peaks
Phase1 Freeze human tracks pooled over DPI robust clusters :: rle
chr1:21616690..21616719,-
chr1:21606013..21606059,-
chr1:21616901..21616963,-
chr1:21616585..21616592,-
chr1:21620877..21620963,-
chr1:21606082..21606114,-
chr1:21585244..21585277,-
chr1:21616888..21616901,+
chr1:21606132..21606156,-
chr1:21606171..21606188,-
chr1:21620977..21620997,-
chr1:21585318..21585329,-
chr1:21621325..21621358,+
chr1:21588574..21588581,+
chr1:21619789..21619809,+
chr1:21588416..21588433,-
chr1:21616655..21616669,-
chr1:21625486..21625505,-
chr1:21617296..21617319,-
chr1:21621597..21621608,-
chr1:21609346..21609355,-
chr1:21616913..21616922,+
chr1:21661003..21661034,+
chr1:21671968..21672009,-
chr1:21672024..21672041,-
chr1:21651897..21651900,-
chr1:21672371..21672405,-
chr1:21652043..21652050,+
chr1:21660669..21660717,-
chr1:21649250..21649266,-
chr1:21651001..21651028,-
chr1:21651262..21651294,-
chr1:21651308..21651324,-
chr1:21651234..21651255,-
chr1:21643880..21643900,-
chr1:21651610..21651654,+
chr1:21651922..21651945,+
chr1:21660761..21660772,-
chr1:21664500..21664517,-
chr1:21661058..21661063,+
chr1:21651456..21651467,+
chr1:21661040..21661043,+
(III) CAGE clusters
Enhancers - Permissive set (Andersson et al.)
chr1:21588176-21588724
chr1:21588829-21589430
chr1:21611199-21611583
chr1:21618792-21618804
chr1:21643831-21644195
chr1:21649091-21649481
chr1:21650806-21651685
chr1:21660622-21661187
chr1:21664400-21664833
chr1:21663991-21664182
chr1:21664947-21665143
F5_human_lncRNAome_no_intra_redundancy_Nov19_n19599
(IV) Expression sites
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [1829 exps]
experiment name
<- reverse strand
<>
forward strand ->
Endothelial Cells - Aortic, donor2 : CNhs11375 ctss [tagcount_pm] 954.27 2.6
Pericytes, donor1 : CNhs11317 ctss [tagcount_pm] 931.45 0
Endothelial Cells - Artery, donor1 : CNhs12496 ctss [tagcount_pm] 826.03 3.86
Endothelial Cells - Aortic, donor3 : CNhs12022 ctss [tagcount_pm] 770.41 0
Endothelial Cells - Aortic, donor0 : CNhs10837 ctss [tagcount_pm] 766.52 0
Endothelial Cells - Thoracic, donor1 : CNhs11926 ctss [tagcount_pm] 744.64 0
Eosinophils, donor2 : CNhs12548 ctss [tagcount_pm] 668.01 185.72
Endothelial Cells - Vein, donor3 : CNhs12026 ctss [tagcount_pm] 620.17 0.91
Endothelial Cells - Thoracic, donor2 : CNhs11978 ctss [tagcount_pm] 617.38 0
Endothelial Cells - Artery, donor2 : CNhs11977 ctss [tagcount_pm] 564.83 0.95
Endothelial Cells - Microvascular, donor1 : CNhs11925 ctss [tagcount_pm] 558.19 0
Endothelial Cells - Vein, donor2 : CNhs11377 ctss [tagcount_pm] 556.41 0
Endothelial Cells - Artery, donor3 : CNhs12023 ctss [tagcount_pm] 545.55 0.74
Eosinophils, donor3 : CNhs12549 ctss [tagcount_pm] 541.48 78.35
Neutrophils, donor3 : CNhs11905 ctss [tagcount_pm] 492.61 67.29
Endothelial Cells - Microvascular, donor3 : CNhs12024 ctss [tagcount_pm] 489.54 0.75
granulosa cell tumor cell line:KGN : CNhs11740 ctss [tagcount_pm] 479.81 0
Renal Glomerular Endothelial Cells, donor2 : CNhs12086 ctss [tagcount_pm] 467.55 1.52
embryonic pancreas cell line:1C3IKEI : CNhs11733 ctss [tagcount_pm] 455.33 0
Pericytes, donor2 : CNhs12079 ctss [tagcount_pm] 446.85 0
Endothelial Cells - Aortic, donor1 : CNhs12495 ctss [tagcount_pm] 443.87 0
Renal Glomerular Endothelial Cells, donor1 : CNhs12074 ctss [tagcount_pm] 429.63 0
CD14+ Monocytes, donor2 : CNhs11954 ctss [tagcount_pm] 413.04 143.38
Neutrophils, donor1 : CNhs10862 ctss [tagcount_pm] 412.94 34.52
Renal Glomerular Endothelial Cells, donor4 : CNhs13080 ctss [tagcount_pm] 403.89 0
Smooth Muscle Cells - Brachiocephalic, donor1 : CNhs11086 ctss [tagcount_pm] 384.78 0.58
Fibroblast - Cardiac, donor5 : CNhs12057 ctss [tagcount_pm] 384.39 0
Fibroblast - Choroid Plexus, donor1 : CNhs11319 ctss [tagcount_pm] 383.2 0
Endothelial Cells - Lymphatic, donor3 : CNhs11906 ctss [tagcount_pm] 379.87 0
Endothelial Cells - Umbilical vein, donor1 : CNhs10872 ctss [tagcount_pm] 379.68 0
Fibroblast - Cardiac, donor2 : CNhs11378 ctss [tagcount_pm] 375.93 0


(b)
Figure S15
50 kb
FFCP_FANTOM5_web_resource_link
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CR625155
AX747766
RefSeq Genes
LOC100506801
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Transcription Factor ChIP-seq from ENCODE
DNase Clusters
Common SNPs(135)
Common SNPs(132)
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Mammal Cons
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RepeatMasker
Repeating Elements by RepeatMasker
NHGRI Catalog of Published Genome-Wide Association Studies

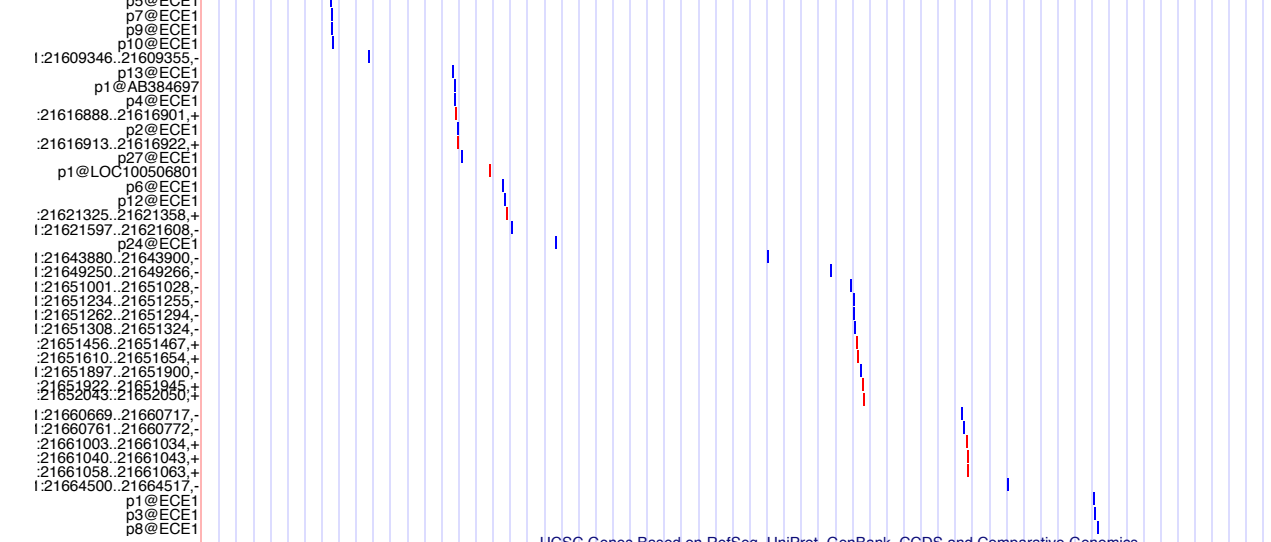

p5@ECE1
p7@ECE1
p9@ECE1
p10@ECE1
I:21609346..21609355,-
p13@ECE1
p1@AB384697
p4@ECE1
:21616888..21616901,+
p2@ECE1
:21616913..21616922,+
p27@ECE1
p1@LOC100506801
p6@ECE1
p12@ECE1
:21621325..21621358,+
I:21621597..21621608,-
p24@ECE1
I:21643880..21643900,-
I:21649250..21649266,-
I:21651001..21651028,-
I:21651234..21651255,-
I:21651262..21651294,-
I:21651308..21651324,-
:21651456..21651467,+
:21651610..21651654,+
I:21651897..21651900,-
:21651922..21651945,+
:21652043..21652050,+
I:21660669..21660717,-
I:21660761..21660772,-
:21661003..21661034,+
:21661040..21661043,+
:21661058..21661063,+
I:21664500..21664517,-
p1@ECE1
p3@ECE1
p8@ECE1
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics

Figure S16

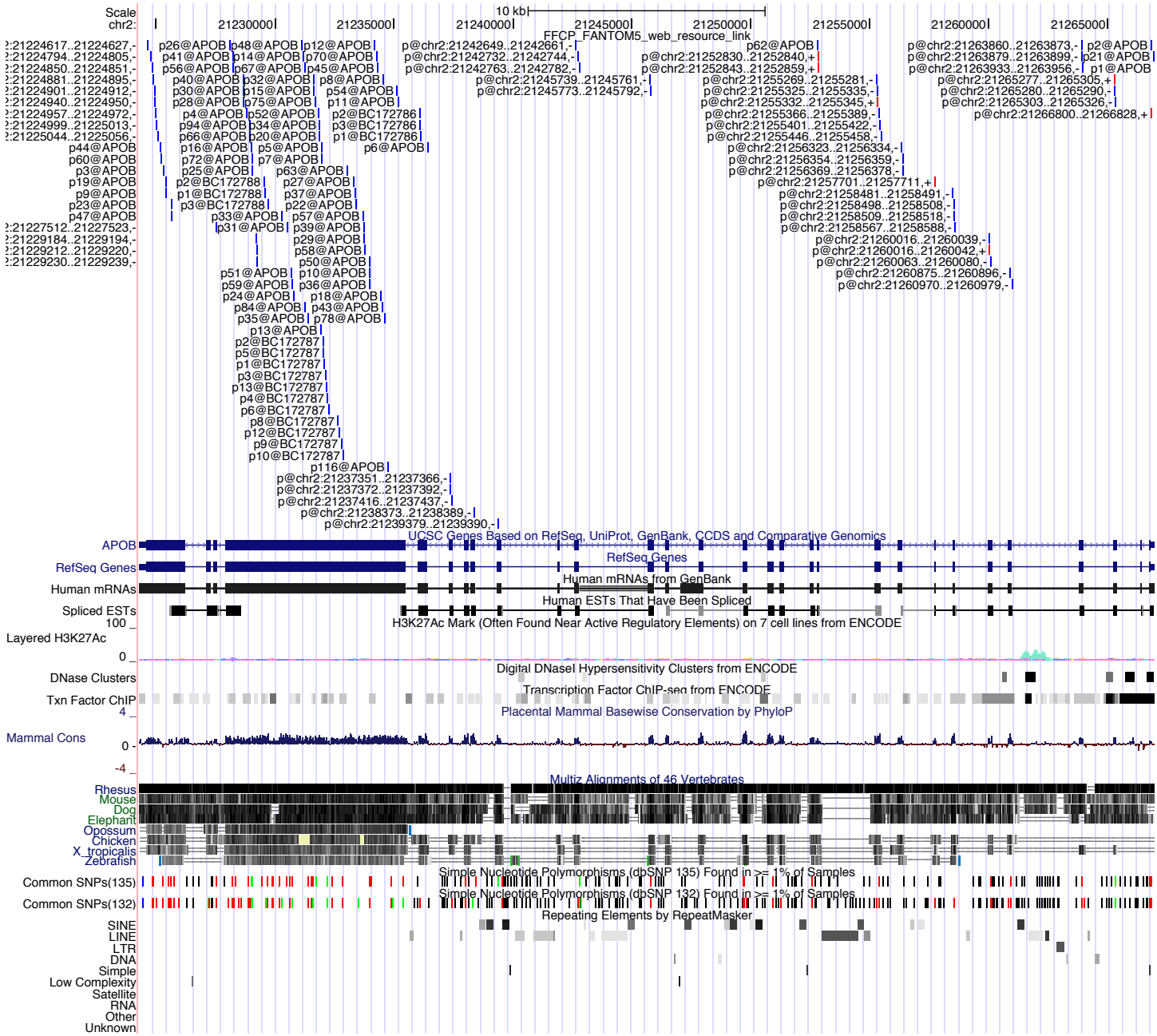
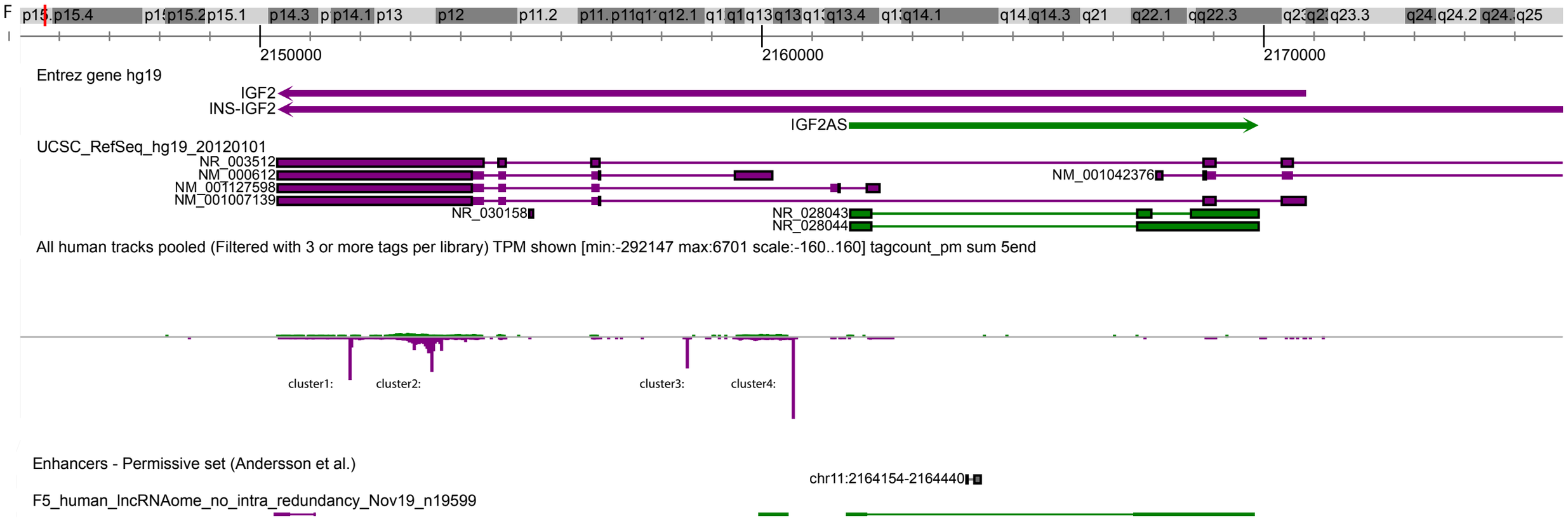

Scale
chr2:
10 kb
21230000
21235000
21240000
21245000
21250000
21255000
21260000
21265000
FFCP_FANTOM5_web_resource_link
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
APOB
RefSeq Genes
Human mRNAs
Spliced ESTs
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
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
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure S17


2150000
2160000
2170000
Entrez gene hg19
IGF2
INS-IGF2
IGF2AS
UCSC_RefSeq_hg19_20120101
NR_003512
NM_000612
NM_001127598
NM_001007139
NR_030158
NM_001042376
NR_028043
NR_028044
All human tracks pooled (Filtered with 3 or more tags per library) TPM shown [min:-292147 max:6701 scale:-160..160] tagcount_pm sum 5end
cluster1:
cluster2:
cluster3:
cluster4:
Enhancers - Permissive set (Andersson et al.)
chr11:2164154-2164440
F5_human_lncRNAome_no_intra_redundancy_Nov19_n19599


cluster1
Phase1 Freeze human tracks pooled over DPI robust clusters :: rle
chr11:2151798..2151813,-
chr11:2151829..2151841,+
experiment name
<- reverse strand
forward strand ->
Amniotic Epithelial Cells, donor2 : CNhs12098 ctss [tagcount_pm] 12708.93 0
amniotic membrane cells, donor1 : CNhs12502 ctss [tagcount_pm] 9368.13 0.69
amniotic membrane cells, donor3 : CNhs12379 ctss [tagcount_pm] 7731.45 0
amniotic membrane cells, donor2 : CNhs12503 ctss [tagcount_pm] 5668.58 1.06
Pericytes, donor2 : CNhs12079 ctss [tagcount_pm] 5658.56 0
tongue, fetal, donor1 : CNhs11768 ctss [tagcount_pm] 5089.61 7.83
Mesenchymal Stem Cells - umbilical, donor2 : CNhs12102 ctss [tagcount_pm] 5007.86 0
skeletal muscle, fetal, donor1 : CNhs11776 ctss [tagcount_pm] 4360.01 0
chorionic membrane cells, donor3 : CNhs12380 ctss [tagcount_pm] 4019.08 0
Amniotic Epithelial Cells, donor3 : CNhs12125 ctss [tagcount_pm] 3501.35 2.51


cluster2
Phase1 Freeze human tracks pooled over DPI robust clusters :: rle
chr11:2153332..2153354,-
chr11:2153356..2153374,-
chr11:2153253..2153337,+
chr11:2153389..2153416,-
experiment name
<- reverse strand
forward strand ->
HES3-GFP Embryonic Stem cells, cardiomyocytic induction, day08, biol_rep1 : CNhs13691 ctss [tagcount_pm] 9699.44 12.62
chorionic membrane cells, donor3 : CNhs12380 ctss [tagcount_pm] 9367.3 187.18
gastrointestinal carcinoma cell line:ECC12 : CNhs11738 ctss [tagcount_pm] 7368.15 118.81
amniotic membrane cells, donor3 : CNhs12379 ctss [tagcount_pm] 6549.92 69.73
carcinoid cell line:NCI-H727 : CNhs14244 ctss [tagcount_pm] 5889.99 70.71
rhabdomyosarcoma cell line:RMS-YM : CNhs11269 ctss [tagcount_pm] 5832.55 22
HES3-GFP Embryonic Stem cells, cardiomyocytic induction, day07, biol_rep1 : CNhs13658 ctss [tagcount_pm] 5628.56 32.64
HES3-GFP Embryonic Stem cells, cardiomyocytic induction, day08, biol_rep3 : CNhs13733 ctss [tagcount_pm] 5586.45 97.31
amniotic membrane cells, donor1 : CNhs12502 ctss [tagcount_pm] 5275.76 48.47
HES3-GFP Embryonic Stem cells, cardiomyocytic induction, day08, biol_rep1 : CNhs13658 ctss [tagcount_pm] 4626.69 85.66

Figure S18

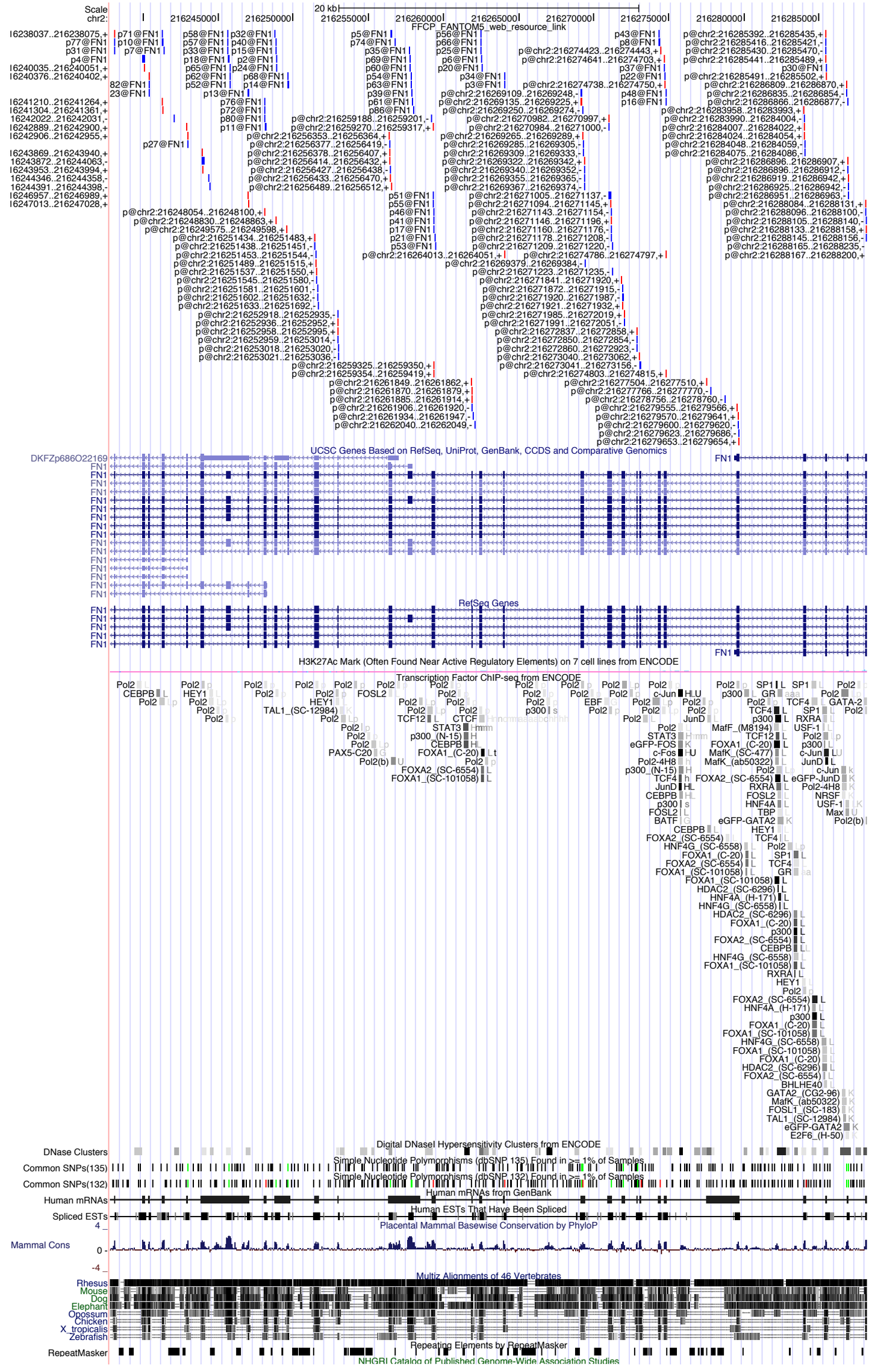

Scale
chr2:
20 kb
216245000
216250000
216255000
216260000
216265000
216270000
216275000
216280000
216285000
FFCP_FANTOM5_web_resource_link
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
DKFZp686O22169
FN1
RefSeq Genes
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Transcription Factor ChIP-seq from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
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
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RepeatMasker
Repeating Elements by RepeatMasker
NHGRI Catalog of Published Genome-Wide Association Studies

Figure 19 (a)

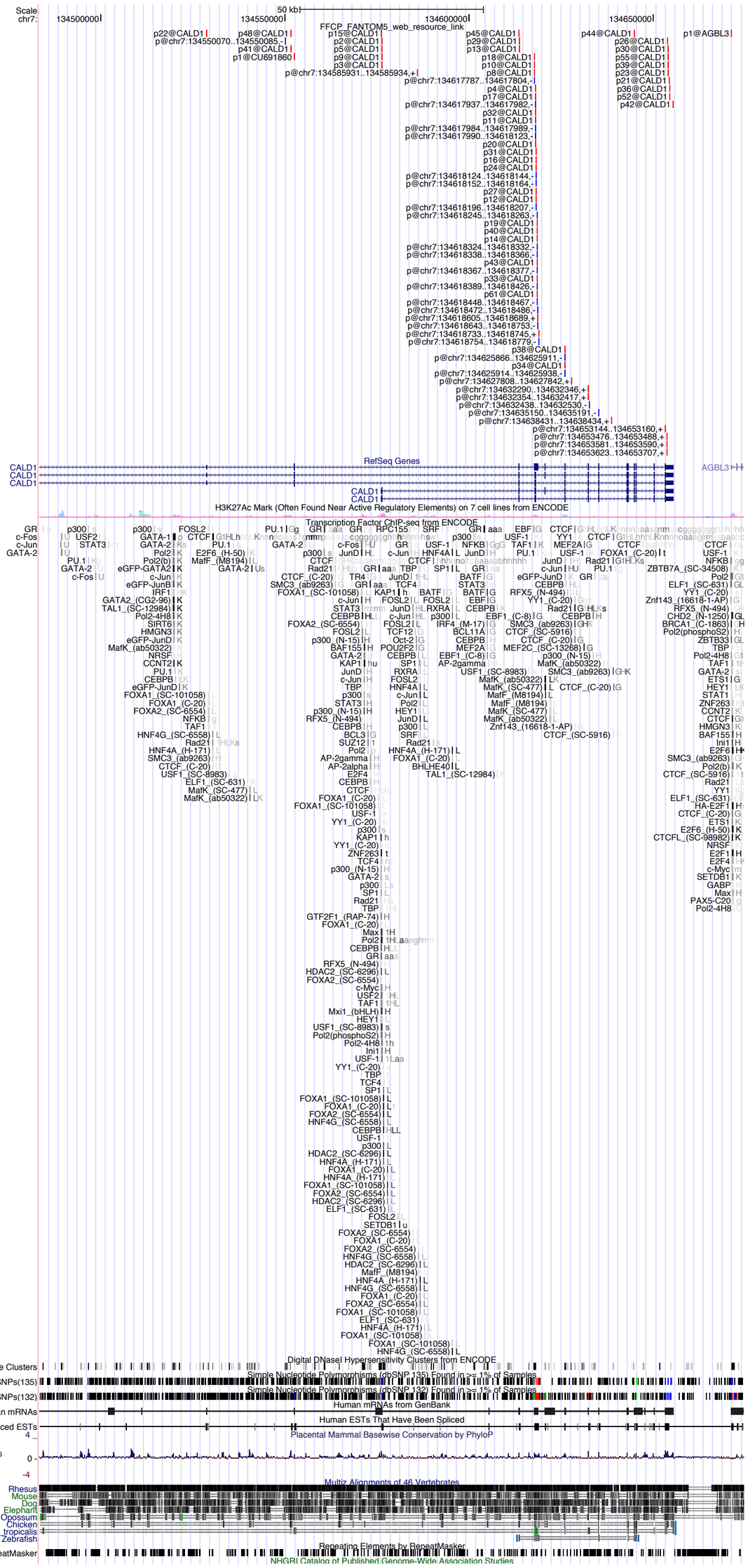


Figure 19 (b)

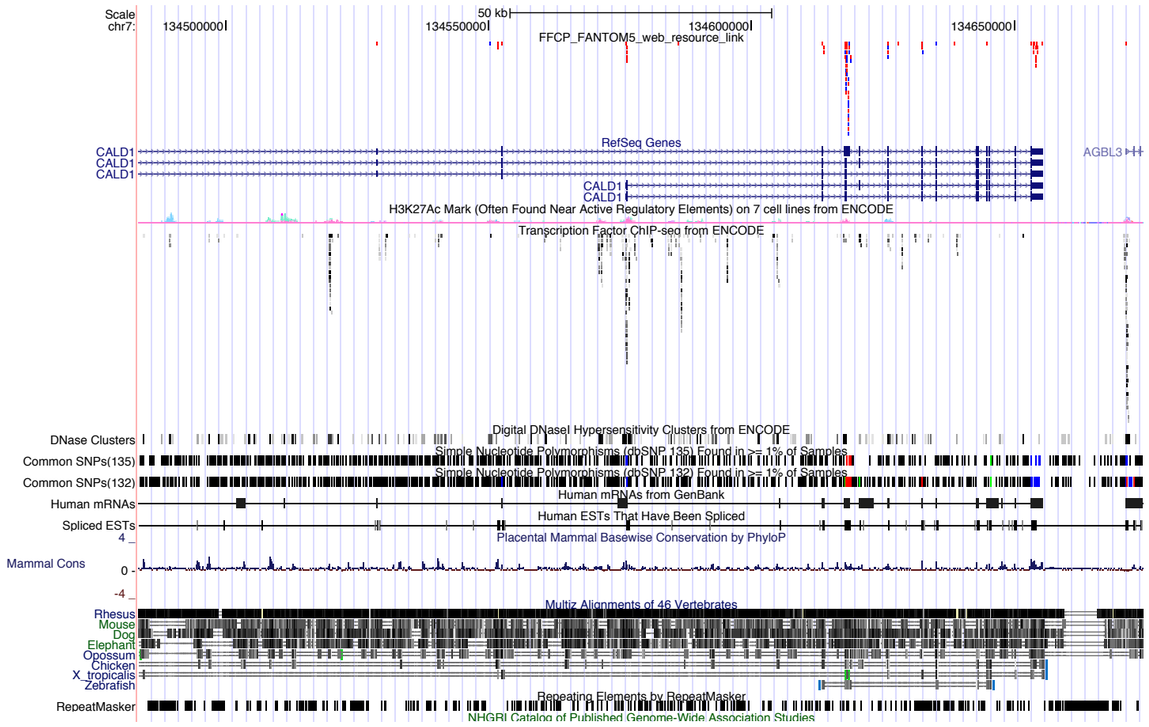


Table S2 (a)

Cluster name	Human	Mouse
lung	tpm.lung, adult, pool1.CNhs10625.10019-101D1.hg19 tpm.lung, fetal, donor1.CNhs11680.10068-101I5.hg19 tpm.lung, right lower lobe, adult, donor1.CNhs11786.10075-102A3.hg19	tpm.lung, embryo E14.CNhs10604.404-26F8.mm9 tpm.lung, embryo E15.CNhs11020.432-17C5.mm9 tpm.lung, embryo E16.CNhs10998.458-17B6.mm9 tpm.lung, embryo E17.CNhs10605.480-44A6.mm9 tpm.lung, embryo E18.CNhs10583.1287-20I6.mm9 tpm.lung, neonate N00.CNhs11224.640-42C6.mm9 tpm.lung, neonate N06.CNhs11212.683-20E5.mm9 tpm.lung, neonate N07.CNhs11111.715-19A3.mm9 tpm.lung, neonate N10.CNhs11219.750-23E6.mm9 tpm.lung, neonate N20.CNhs11109.822-18A4.mm9 tpm.lung, neonate N25.CNhs11119.1359-25C7.mm9 tpm.lung, neonate N30.CNhs11133.1394-42H2.mm9 tpm.lung, adult.CNhs10474.28-22B1.mm9 tpm.lung, embryo E12.CNhs10522.354-16G2.mm9
colon	tpm.colon, adult, donor1.CNhs11794.10082-102B1.hg19 tpm.colon, adult, pool1.CNhs10619.10014-101C5.hg19 tpm.colon, fetal, donor1.CNhs11780.10070-101I7.hg19	tpm.colon, adult.CNhs10468.36-18H7.mm9
diencephalon	tpm.diencephalon, adult.CNhs12610.10193-103E4.hg19	tpm.diencephalon, adult.CNhs10482.20-12F2.mm9 tpm.diencephalon, neonate N30.CNhs11201.1388-42B2.mm9
skin	tpm.skin, fetal, donor1.CNhs11774.10065-101I2.hg19 tpm.skin, adult, donor1.CNhs11785.10074-102A2.hg19 tpm.Skin - palm, donor1.CNhs13458.10286-104F7.hg19	tpm.skin, adult.CNhs10492.30-1C3.mm9 tpm.skin, neonate N00.CNhs11124.650-25G4.mm9 tpm.skin, neonate N03.CNhs11215.662-22H3.mm9 tpm.skin, neonate N06.CNhs11097.693-20I5.mm9 tpm.skin, neonate N10.CNhs11108.762-6C4.mm9
kidney	tpm.kidney, adult, pool1.CNhs10622.10017-101C8.hg19 tpm.kidney, fetal, pool1.CNhs10652.10045-101F9.hg19	tpm.kidney, neonate N25.CNhs11122.1353-25F3.mm9 tpm.kidney, neonate N30.CNhs11203.1385-42H1.mm9 tpm.kidney, embryo E14.CNhs10606.411-4I9.mm9 tpm.kidney, embryo E15.CNhs10997.434-16F8.mm9 tpm.kidney, embryo E16.CNhs10584.464-22A5.mm9 tpm.kidney, embryo E17.CNhs11028.483-18I2.mm9 tpm.kidney, embryo E18.CNhs11001.1288-20C7.mm9 tpm.kidney, neonate N00.CNhs11214.646-21G7.mm9 tpm.kidney, neonate N10.CNhs11206.758-6D5.mm9 tpm.kidney, neonate N20.CNhs11113.832-19I1.mm9
liver	tpm.liver, adult, pool1.CNhs10624.10018-101C9.hg19 tpm.liver, fetal, pool1.CNhs11798.10086-102B5.hg19	tpm.liver, embryo E15.CNhs10520.433-16D7.mm9 tpm.liver, embryo E16.CNhs10523.462-17F5.mm9 tpm.liver, embryo E17.CNhs10510.481-18A3.mm9 tpm.liver, embryo E18.CNhs10579.499-43G4.mm9 tpm.liver, neonate N00.CNhs11117.641-24F4.mm9 tpm.liver, neonate N03.CNhs11123.1345-25G2.mm9 tpm.liver, neonate N06.CNhs11101.684-20A8.mm9 tpm.liver, neonate N07.CNhs11103.716-20H2.mm9 tpm.liver, neonate N10.CNhs11115.751-24B9.mm9 tpm.liver, neonate N20.CNhs11220.823-25A3.mm9 tpm.liver, neonate N25.CNhs11198.1368-26H1.mm9 tpm.liver, neonate N30.CNhs11106.1382-42D1.mm9 tpm.liver, adult pregnant day01.CNhs10466.508-5B2.mm9 tpm.liver, embryo E12.CNhs10601.355-15F8.mm9 tpm.liver, embryo E13.CNhs10524.378-3H6.mm9 tpm.liver, embryo E14.CNhs10594.409-16E1.mm9
uterus	tpm.uterus, adult, pool1.CNhs11676.10100-102D1.hg19 tpm.uterus, fetal, donor1.CNhs11763.10055-101H1.hg19	tpm.uterus, adult pregnant day19.CNhs10497.590-15F5.mm9 tpm.uterus, adult.CNhs10509.92-27E5.mm9
tongue	tpm.tongue, fetal, donor1.CNhs11768.10059-101H5.hg19	tpm.tongue, adult.CNhs10499.32-1B4.mm9

	tpm.tongue, adult.CNhs12853.10203-103F5.hg19 tpm.tongue epidermis (fungiform papillae), donor1.CNhs13460.10288-104F9.hg19	
stomach	tpm.stomach, fetal, donor1.CNhs11771.10062-101H8.hg19	tpm.stomach, neonate N03.CNhs11193.1340-24B8.mm9 tpm.stomach, neonate N07.CNhs11210.718-18A9.mm9 tpm.stomach, neonate N25.CNhs11104.1355-25E4.mm9 tpm.stomach, neonate N30.CNhs11134.1386-42I1.mm9 tpm.stomach, adult.CNhs10503.33-1H6.mm9 tpm.stomach, embryo E12.CNhs10588.356-43E3.mm9 tpm.stomach, embryo E15.CNhs10603.1252-16G8.mm9 tpm.stomach, embryo E16.CNhs11022.459-17D4.mm9 tpm.stomach, embryo E17.CNhs11006.1035-18D5.mm9 tpm.stomach, embryo E18.CNhs10999.1286-20C6.mm9
ovary	tpm.ovary, adult, pool1.CNhs10626.10020-101D2.hg19	tpm.ovary, adult.CNhs10507.91-2I7.mm9 tpm.ovary, embryo E18.CNhs11040.505-44H7.mm9 tpm.ovary, neonate N00.CNhs11217.1299-22I4.mm9
aorta	tpm.aorta, adult, pool1.CNhs11760.10052-101G7.hg19	tpm.aorta, adult.CNhs10498.46-23H1.mm9
heart	tpm.heart, adult, diseased post-infarction, donor1.CNhs11757.10050-101G5.hg19 tpm.heart, adult, diseased, donor1.CNhs11758.10051-101G6.hg19 tpm.heart, adult, pool1.CNhs10621.10016-101C7.hg19 tpm.heart, fetal, pool1.CNhs10653.10046-101G1.hg19	tpm.heart, embryo E17.CNhs11025.479-18E5.mm9 tpm.heart, embryo E18.CNhs11030.1283-20G3.mm9 tpm.heart, neonate N00.CNhs11213.639-21E3.mm9 tpm.heart, neonate N03.CNhs11221.1349-25I2.mm9 tpm.heart, neonate N10.CNhs11118.749-24G1.mm9 tpm.heart, neonate N16.CNhs11209.782-15G1.mm9 tpm.heart, neonate N20.CNhs11127.821-26I6.mm9 tpm.heart, neonate N25.CNhs11196.1351-25D3.mm9 tpm.heart, neonate N30.CNhs11202.1390-42D2.mm9 tpm.heart, embryo E11.CNhs10586.331-24E9.mm9 tpm.heart, embryo E12.CNhs11015.353-12F5.mm9 tpm.heart, embryo E13.CNhs11013.376-3I9.mm9 tpm.heart, embryo E14.CNhs10597.403-26D4.mm9 tpm.heart, embryo E15.CNhs11017.431-16C8.mm9 tpm.heart, embryo E16.CNhs11021.457-17C6.mm9
prostate	tpm.prostate, adult, pool1.CNhs10628.10022-101D4.hg19	tpm.prostate, adult.CNhs10470.859-1F8.mm9
vagina	tpm.vagina, adult.CNhs12854.10204-103F6.hg19	tpm.vagina, adult.CNhs10502.89-27D5.mm9
spleen	tpm.spleen, adult, pool1.CNhs10631.10025-101D7.hg19 tpm.spleen, fetal, pool1.CNhs10651.10044-101F8.hg19	tpm.spleen, adult.CNhs10465.25-2G2.mm9 tpm.spleen, embryo E16.CNhs11035.461-43C1.mm9 tpm.spleen, embryo E18.CNhs11011.1271-21F2.mm9 tpm.spleen, neonate N10.CNhs11116.752-24D1.mm9 tpm.spleen, neonate N20.CNhs11112.824-19H1.mm9 tpm.spleen, neonate N25.CNhs11099.1354-25G3.mm9
placenta	tpm.placenta, adult, pool1.CNhs10627.10021-101D3.hg19	tpm.placenta, adult pregnant day10.CNhs10472.539-13I7.mm9 tpm.placenta, adult pregnant day17.CNhs10464.577-18G3.mm9
pancreas	tpm.pancreas, adult, donor1.CNhs11756.10049-101G4.hg19	tpm.pancreas, neonate N25.CNhs11094.1555-45E2.mm9 tpm.pancreas, neonate N30.CNhs11182.1548-44H5.mm9 tpm.pancreas, adult.CNhs10486.34-16E4.mm9

		tpm.pancreas, embryo E14.CNhs11012.405-44F4.mm9 tpm.pancreas, embryo E15.CNhs11042.1556-45F2.mm9 tpm.pancreas, embryo E16.CNhs11003.460-26E5.mm9 tpm.pancreas, embryo E17.CNhs10599.1558-45G3.mm9 tpm.pancreas, embryo E18.CNhs10580.1535-43I7.mm9 tpm.pancreas, neonate N00.CNhs11105.645-26G9.mm9 tpm.pancreas, neonate N01.CNhs11138.1531-43D7.mm9 tpm.pancreas, neonate N02.CNhs11139.1539-43D8.mm9 tpm.pancreas, neonate N16.CNhs11136.787-43B6.mm9
testis	tpm.testis, adult, pool1.CNhs10632.10026-101D8.hg19 tpm.testis, adult, pool2.CNhs12998.10096-102C6.hg19	tpm.testis, neonate N30.CNhs11130.1391-42E2.mm9 tpm.testis, adult.CNhs10504.57-7G5.mm9 tpm.testis, embryo E13.CNhs11031.389-23F5.mm9 tpm.testis, embryo E15.CNhs11034.443-27C5.mm9 tpm.testis, embryo E16.CNhs11033.470-26G5.mm9 tpm.testis, embryo E17.CNhs11029.486-18I5.mm9 tpm.testis, embryo E18.CNhs11027.503-18G5.mm9 tpm.testis, neonate N00.CNhs11189.1300-22B7.mm9 tpm.testis, neonate N07.CNhs11222.726-26I4.mm9 tpm.testis, neonate N10.CNhs11204.763-44D9.mm9 tpm.testis, neonate N20.CNhs11110.839-18H3.mm9
small intestine	tpm.small intestine, fetal, donor1.CNhs11773.10064-101I1.hg19 tpm.small intestine, adult, pool1.CNhs10630.10024-101D6.hg19	tpm.small intestine, neonate N16.CNhs11114.790-21I1.mm9
medulla oblongata	tpm.medulla oblongata, adult, donor10252.CNhs12315.10155-103A2.hg19 tpm.medulla oblongata, adult, pool1.CNhs10645.10038-101F2.hg19 tpm.medulla oblongata, newborn, donor10223.CNhs14079.10361-105F1.hg19	tpm.medulla oblongata, adult.CNhs10477.17-12C2.mm9 tpm.medulla oblongata, neonate N30.CNhs11200.1396-42A3.mm9
adrenal gland	tpm.adrenal gland, adult, pool1.CNhs11793.10081-102A9.hg19	tpm.adrenal gland, adult.CNhs10508.49-24D7.mm9 tpm.adrenal gland, embryo E14.CNhs11038.406-44C5.mm9 tpm.adrenal gland, embryo E16.CNhs11004.1254-43D6.mm9 tpm.adrenal gland, embryo E17.CNhs11043.1263-45I1.mm9 tpm.adrenal gland, embryo E18.CNhs11026.1262-18F5.mm9 tpm.adrenal gland, neonate N00.CNhs11191.1311-23C1.mm9 tpm.adrenal gland, neonate N25.CNhs11223.1377-27I3.mm9
spinal cord	tpm.spinal cord, adult, donor10258.CNhs14222.10369-105F9.hg19 tpm.spinal cord, fetal, donor1.CNhs11764.10056-101H2.hg19 tpm.spinal cord, newborn, donor10223.CNhs14077.10359-105E8.hg19 tpm.spinal cord, adult, donor10252.CNhs12227.10159-103A6.hg19	tpm.spinal cord, adult.CNhs10505.24-13C9.mm9
pituitary gland	tpm.pituitary gland, adult, donor10252.CNhs12229.10162-103A9.hg19	tpm.pituitary gland, embryo E13.CNhs11009.370-44C7.mm9 tpm.pituitary gland, embryo E14.CNhs11037.398-44A5.mm9 tpm.pituitary gland, embryo E15.CNhs10592.427-16B9.mm9 tpm.pituitary gland, embryo E16.CNhs11036.449-43F6.mm9 tpm.pituitary gland, embryo E17.CNhs11039.1265-

		44F5.mm9 tpm.pituitary gland, neonate N00.CNhs11190.1308-22E9.mm9 tpm.pituitary gland, adult.CNhs10493.21-1G8.mm9 tpm.pituitary gland, embryo E12.CNhs11018.346-16E6.mm9
thymus	tpm.thymus, adult, pool1.CNhs10633.10027-101D9.hg19 tpm.thymus, fetal, pool1.CNhs10650.10043-101F7.hg19	tpm.thymus, adult.CNhs10471.38-12B5.mm9 tpm.thymus, embryo E14.CNhs11041.402-44I4.mm9 tpm.thymus, embryo E15.CNhs11005.430-45I2.mm9 tpm.thymus, embryo E16.CNhs11002.456-26C5.mm9 tpm.thymus, embryo E17.CNhs10581.478-18C1.mm9 tpm.thymus, embryo E18.CNhs10595.1273-19G4.mm9 tpm.thymus, neonate N02.CNhs11181.1541-43F8.mm9 tpm.thymus, neonate N03.CNhs11137.2104-43C7.mm9 tpm.thymus, neonate N06.CNhs11197.681-26C4.mm9 tpm.thymus, neonate N07.CNhs11211.713-19D3.mm9 tpm.thymus, neonate N10.CNhs11194.748-24E4.mm9 tpm.thymus, neonate N20.CNhs11186.820-7A2.mm9 tpm.thymus, neonate N25.CNhs11125.1362-25G7.mm9 tpm.thymus, neonate N30.CNhs11132.1393-42G2.mm9
hippocampus	tpm.hippocampus, adult, donor10252.CNhs12312.10153-102I9.hg19 tpm.hippocampus, adult, donor10258.CNhs14227.10374-105G5.hg19 tpm.hippocampus, newborn, donor10223.CNhs14081.10363-105F3.hg19	tpm.hippocampus, adult.CNhs10478.13-16E8.mm9 tpm.hippocampus, neonate N00.CNhs11228.627-43G1.mm9
cerebellum	tpm.cerebellum, adult, donor10252.CNhs12323.10166-103B4.hg19 tpm.cerebellum, adult, pool1.CNhs11795.10083-102B2.hg19 tpm.cerebellum, newborn, donor10223.CNhs14075.10357-105E6.hg19	tpm.cerebellum, embryo E16, biol_rep3 (E16R3).CNhs13019.10143-102H8.mm9 tpm.cerebellum, embryo E17, biol_rep1 (E17R1).CNhs12818.10120-102F3.mm9 tpm.cerebellum, embryo E17, biol_rep2 (E17R2).CNhs13008.10132-102G6.mm9 tpm.cerebellum, embryo E17, biol_rep3 (E17R3).CNhs13020.10144-102H9.mm9 tpm.cerebellum, embryo E18, biol_rep1 (E18R1).CNhs12962.10121-102F4.mm9 tpm.cerebellum, embryo E18, biol_rep2 (E18R2).CNhs13009.10133-102G7.mm9 tpm.cerebellum, embryo E18, biol_rep3 (E18R3).CNhs13021.10145-102I1.mm9 tpm.cerebellum, neonate N00, biol_rep1 (P0R1).CNhs12963.10122-102F5.mm9 tpm.cerebellum, neonate N00, biol_rep2 (P0R2).CNhs13010.10134-102G8.mm9 tpm.cerebellum, neonate N00, biol_rep3 (P0R3).CNhs13022.10146-102I2.mm9 tpm.cerebellum, neonate N03, biol_rep1 (P3R1).CNhs13001.10123-102F6.mm9 tpm.cerebellum, neonate N03, biol_rep2 (P3R2).CNhs13011.10135-102G9.mm9 tpm.cerebellum, neonate N03, biol_rep3 (P3R3).CNhs13024.10147-102I3.mm9 tpm.cerebellum, neonate N06, biol_rep1 (P6R1).CNhs12819.10124-102F7.mm9 tpm.cerebellum, neonate N06, biol_rep2 (P6R2).CNhs13012.10136-102H1.mm9 tpm.cerebellum, neonate N06, biol_rep3 (P6R3).CNhs13025.10148-102I4.mm9 tpm.cerebellum, neonate N09, biol_rep1 (P9R1).CNhs12820.10125-102F8.mm9

		tpm.cerebellum, neonate N09, biol_rep2 (P9R2).CNhs13013.10137-102H2.mm9 tpm.cerebellum, neonate N09, biol_rep3 (P9R3).CNhs13026.10149-102I5.mm9 tpm.cerebellum, neonate N30.CNhs11135.1395-42I2.mm9 tpm.cerebellum, adult.CNhs10494.15-8B2.mm9 tpm.cerebellum, embryo E11, biol_rep1 (E11R1).CNhs12956.10114-102E6.mm9 tpm.cerebellum, embryo E11, biol_rep2 (E11R2).CNhs13002.10126-102F9.mm9 tpm.cerebellum, embryo E11, biol_rep3 (E11R3).CNhs13014.10138-102H3.mm9 tpm.cerebellum, embryo E12, biol_rep1 (E12R1).CNhs12957.10115-102E7.mm9 tpm.cerebellum, embryo E12, biol_rep2 (E12R2).CNhs13003.10127-102G1.mm9 tpm.cerebellum, embryo E12, biol_rep3 (E12R3).CNhs13015.10139-102H4.mm9 tpm.cerebellum, embryo E13, biol_rep1 (E13R1).CNhs12958.10116-102E8.mm9 tpm.cerebellum, embryo E13, biol_rep2 (E13R2).CNhs13004.10128-102G2.mm9 tpm.cerebellum, embryo E13, biol_rep3 (E13R3).CNhs13016.10140-102H5.mm9 tpm.cerebellum, embryo E14, biol_rep1 (E14R1).CNhs12960.10117-102E9.mm9 tpm.cerebellum, embryo E14, biol_rep2 (E14R2).CNhs13005.10129-102G3.mm9 tpm.cerebellum, embryo E14, biol_rep3 (E14R3).CNhs13017.10141-102H6.mm9 tpm.cerebellum, embryo E15, biol_rep1 (E15R1).CNhs12961.10118-102F1.mm9 tpm.cerebellum, embryo E15, biol_rep2 (E15R2).CNhs13006.10130-102G4.mm9 tpm.cerebellum, embryo E15, biol_rep3 (E15R3).CNhs13018.10142-102H7.mm9 tpm.cerebellum, embryo E16, biol_rep1 (E16R1).CNhs13000.10119-102F2.mm9 tpm.cerebellum, embryo E16, biol_rep2 (E16R2).CNhs13007.10131-102G5.mm9
RNA	tpm.Clontech Human Universal Reference Total RNA, pool1.CNhs10608.10000-101A1.hg19	tpm.Clontech Mouse Universal Reference Total RNA, pool1.CNhs10609.10001-101A3.mm9
eye	tpm.eye, fetal, donor1.CNhs11762.10054-101G9.hg19 tpm.eye - muscle inferior rectus, donor1.CNhs13444.10272-104E2.hg19 tpm.eye - muscle lateral, donor2.CNhs13442.10298-104H1.hg19 tpm.eye - muscle medial, donor2.CNhs13443.10299-104H2.hg19 tpm.eye - muscle superior, donor2.CNhs13441.10297-104G9.hg19 tpm.eye - vitreous humor, donor1.CNhs13440.10268-104D7.hg19	tpm.eyeball, embryo E14.CNhs10521.399-16E2.mm9 tpm.eyeball, embryo E15.CNhs10593.426-16C9.mm9 tpm.eyeball, embryo E17.CNhs11023.1261-18D4.mm9 tpm.eyeball, neonate N00.CNhs11207.633-15C6.mm9 tpm.eyeball, neonate N01.CNhs11140.1532-43E7.mm9 tpm.eyeball, neonate N02.CNhs11205.1551-44G8.mm9 tpm.eyeball, neonate N16.CNhs11188.777-19A2.mm9 tpm.eyeball, adult.CNhs10484.31-12G4.mm9 tpm.eyeball, embryo E12.CNhs11016.345-16C6.mm9
RNA	tpm.SABiosciences XpressRef Human Universal Total RNA, pool1.CNhs10610.10002-101A5.hg19	tpm.SABiosciences XpressRef Mouse Universal Total RNA, pool1.CNhs10611.10003-101A7.mm9
epididymis	tpm.epididymis, adult.CNhs12847.10197-103E8.hg19	tpm.epididymis, adult.CNhs10490.58-23B2.mm9 tpm.epididymis and seminiferous tubule, neonate N30.CNhs11199.1387-42A2.mm9 tpm.epididymis and seminiferous tubule, neonate N00.CNhs11218.1310-23B1.mm9

Table S2 (b)

Cluster name	Human	Mouse
Cluster 1, subcluster various mouse tissues - nervous, reproductive and digestive system	-	tpm stomach− embryo E12−mm9 tpm intestine− embryo E12−mm9 tpm forelimb− embryo E11−mm9 tpm forelimb− embryo E13−mm9 tpm forelimb− embryo E12−mm9 tpm lung− embryo E12−mm9 tpm lung− embryo E14−mm9 tpm intestine− embryo E13−mm9 tpm gonad− embryo E13−mm9 tpm testis− embryo E13−mm9 tpm pituitary gland− embryo E12−mm9 tpm pituitary gland− embryo E13−mm9 tpm pituitary gland− embryo E14−mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm whole body− embryo E11−mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm eyeball− embryo E12−mm9 tpm whole body− embryo E12−mm9 tpm kidney− embryo E17−mm9 tpm kidney− embryo E18−mm9 tpm kidney− embryo E16−mm9 tpm kidney− neonate N00−mm9 tpm intestine− embryo E15−mm9 tpm stomach− embryo E15−mm9 tpm stomach− embryo E16−mm9 tpm intestine− embryo E16−mm9 tpm forelimb− embryo E14−mm9 tpm forelimb− embryo E15−mm9 tpm ovary− embryo E18−mm9 tpm ovary− neonate N00−mm9 tpm epididymis and seminiferous tubule− neonate N00−mm9 tpm testis− embryo E17−mm9 tpm testis− embryo E16−mm9 tpm testis− embryo E18−mm9 tpm testis− embryo E15−mm9 tpm testis− neonate N00−mm9 tpm testis− neonate N07−mm9 tpm testis− neonate N10−mm9 tpm pancreas− embryo E15−mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm corpus striatum− neonateN00−mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep2 −mm9

		tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep2 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm cerebellum− biol_rep1 −mm9 tpm cerebellum− biol_rep3 −mm9 tpm hippocampus− neonate N00−mm9 tpm pituitary gland− embryo E16−mm9 tpm pituitary gland− embryo E15−mm9 tpm adrenal gland− neonate N00−mm9 tpm adrenal gland− embryo E18−mm9 tpm adrenal gland− embryo E16−mm9 tpm adrenal gland− embryo E17−mm9 tpm adrenal gland− neonate N25−mm9 tpm adrenal gland− adult−mm9 tpm visual cortex-mm9 (various samples) tpm cortex− neonate N30−mm9 tpm cortex− adult−mm9 tpm eyeball− adult−mm9 pm hippocampus− adult−mm9 tpm diencephalon− adult−mm9 tpm corpora quadrigemina− adult−mm9 tpm diencephalon− neonate N30−mm9 tpm corpus striatum− adult−mm9 tpm medulla oblongata− adult−mm9 tpm spinal cord− adult−mm9 tpm medulla oblongata− neonate N30−mm9 tpm oviduct− adult pregnant day01−mm9 tpm uterus− adult−mm9 tpm ovary− adult−mm9 tpm urinary bladder− adult−mm9 tpm stomach− embryo E17−mm9 tpm stomach− embryo E18−mm9 tpm stomach− neonate N03−mm9 tpm stomach− neonate N07−mm9 pm kidney− neonate N25−mm9 tpm kidney− neonate N20−mm9 tpm kidney− neonate N30−mm9 tpm uterus− adult pregnant day19−mm9
Cluster 1, subcluster human newborn and fetal brain, reproductive tissues, lung and some outliers.	pm medial temporal gyrus− newborn−hg19 tpm occipital cortex− newborn−hg19 tpm hippocampus− newborn−hg19 tpm amygdala− newborn−hg19 tpm caudate nucleus− newborn−hg19 tpm thalamus− newborn−hg19 tpm medial frontal gyrus− newborn−hg19 tpm temporal lobe− fetal−hg19 tpm occipital lobe− fetal−hg19 tpm parietal lobe− fetal−hg19 tpm brain− fetal−hg19 tpm duodenum− fetal−hg19 tpm adipose− donor3.CNhs13974.10186.103D6.hg19−hg19 tpm adipose− donor2.CNhs13973.10185.103D5.hg19−hg19 tpm adipose− donor4.CNhs13975.10187.103D7.hg19−hg19 tpm adipose− donor1.CNhs13972.10184.103D4.hg19−hg19 tpm appendix− adult.CNhs12842.10189.103D9.hg19−hg19	

	tpm vein− adult.CNhs12844.10191.103E2.hg19−hg19 tpm lymph node− adult−hg19 tpm spleen− fetal−hg19 tpm spleen− adult−hg19 tpm dura mater− adult−hg19 tpm cruciate ligament− donor2.CNhs13439.10295.104G7.hg19−hg19 tpm achilles tendon− donor2.CNhs13435.10292.104G4.hg19−hg19 tpm eye vitreous humor− donor1.CNhs13440.10268.104D7.hg19−hg19 tpm kidney− fetal−hg19 tpm Clontech Human Universal Reference Total RNA− pool1.CNhs10608.10000.101A tpm adipose tissue− adult−hg19 tpm breast− adult−hg19 tpm bone marrow− adult.CNhs12845.10192.103E3.hg19−hg19 tpm lung− adult−hg19 tpm lung− right lower lobe−hg19 tpm lung− fetal−hg19 tpm prostate− adult−hg19 tpm uterus− adult−hg19 tpm ovary− adult−hg19 tpm cervix− adult−hg19 tpm epididymis− adult.CNhs12847.10197.103E8.hg19−hg19 tpm aorta− adult−hg19 tpm Urethra− donor2.CNhs13464.10319.105A4.hg19−hg19 tpm gall bladder− adult.CNhs12848.10198.103E9.hg19−hg19 tpm bladder− adult−hg19 tpm smooth muscle− adult−hg19 tpm colon− adult−hg19 tpm colon− fetal−hg19 tpm vagina− adult.CNhs12854.10204.103F6.hg19−hg19 tpm placenta− adult−hg19 tpm adrenal gland− adult−hg19 tpm trachea− adult−hg19 tpm tongue epidermis 28fungiform papillae 29− donor1.CNhs13460.10288.104F9.hg1 tpm tonsil− adult−hg19 tpm esophagus− adult−hg19 tpm Skin palm− donor1.CNhs13458.10286.104F7.hg19−hg19 tpm ductus deferens− adult.CNhs12846.10196.103E7.hg19−hg19 tpm seminal vesicle− adult.CNhs12851.10201.103F3.hg19−hg19	
Cluster1, mixed subcluster thymus, skin and some human fetal samples	tpm thymus− fetal−hg19 tpm thymus− adult−hg19 tpm rectum− fetal−hg19 tpm uterus− fetal−hg19 tpm umbilical cord− fetal−hg19 tpm skin− fetal−hg19 tpm skin− adult−hg19 tpm throat− fetal−hg19 tpm trachea− fetal−hg19 tpm eye− fetal−hg19 tpm stomach− fetal−hg19 t	ttpm lung− embryo E15−mm9 tpm lung− embryo E16−mm9 tpm kidney− embryo E15−mm9 tpm pancreas− embryo E14−mm9 tpm kidney− embryo E14−mm9 tpm thymus− embryo E14−mm9 tpm thymus− embryo E15−mm9 tpm thymus− embryo E16−mm9 tpm thymus− embryo E17−mm9 tpm thymus− neonate N03−mm9 tpm thymus− neonate N07−mm9 tpm thymus− neonate N02−mm9 tpm thymus− neonate N06−mm9 tpm thymus− neonate N10−mm9 tpm thymus− embryo E18−mm9 tpm thymus− neonate N20−mm9 tpm thymus− neonate N25−mm9

		tpm thymus− neonate N30−mm9 tpm thymus− adult−mm9 tpm accessory axillary lymph node− adult−mm9 tpm skin− neonate N03−mm9 tpm skin− neonate N00−mm9 tpm forelimb− embryo E17−mm9 tpm forelimb− embryo E18−mm9 tpm skin− neonate N06−mm9 tpm skin− neonate N10−mm9 tpm skin− adult−mm9
Cluster 1, mixed subcluster - heart	tpm heart− adult−hg19 tpm heart− adult−hg19 tpm left ventricle− adult−hg19 tpm heart pulmonic valve− adult.CNhs12856.10206.103F8.hg19−hg19 tpm heart− adult−hg19 tpm heart− fetal−hg19 tpm heart tricuspid valve− adult.CNhs12857.10207.103F9.hg19−hg19 tpm heart mitral valve− adult.CNhs12855.10205.103F7.hg19−hg19 tpm left atrium− adult−hg19	tpm heart− embryo E16−mm9 tpm heart− embryo E17−mm9 tpm heart− embryo E18−mm9 tpm heart− neonate N00−mm9 tpm heart− embryo E14−mm9 tpm heart− embryo E15−mm9 tpm heart− neonate N10−mm9 tpm heart− neonate N16−mm9 tpm heart− neonate N03−mm9 tpm heart− embryo E12−mm9 tpm heart− embryo E11−mm9 tpm heart− embryo E13−mm9
Cluster 1, mixed subcluster tongue, intestine + outliers	tpm colon− adult−hg19 tpm small intestine− adult−hg19 tpm skeletal muscle− fetal−hg19 tpm diaphragm− fetal−hg19 tpm tongue− fetal−hg19 tpm tongue− adult.CNhs12853.10203.103F5.hg19−hg19 tpm skeletal muscle− adult−hg19 tpm artery−adult.CNhs12843.10190.103E1.hg19−hg19 tpm cerebrospinal fluid− donor2.CNhs13437.10294.104G6.hg19−hg19 tpm penis− adult.CNhs12850.10200.103F2.hg19−hg19 tpm skeletal muscle soleus muscle− donor1.CNhs13454.10282.104F3.hg19−hg19 tpm throat− adult.CNhs12858.10209.103G2.hg19−hg19	tpm Follicle Associated Epithelium− pool3−mm9 tpm intestine− neonate N01−mm9 tpm intestine− adult−mm9 tpm intestine− neovnate N06−mm9 tpm intestine− neonate N10−mm9 tpm intestine− neonate N07−mm9 tpm intestine− embryo E17−mm9 tpm intestine− embryo E18−mm9 tpm intestine− neonate N25−mm9 tpm Ileum epithelium− pool1−mm9 tpm cecum− adult−mm9 tpm stomach− neonate N30−mm9 tpm whole body− neonate N06−mm9 tpm whole body− neonate N10−mm9 tpm bone 28os femoris 29− neonate N02−mm9 tpm whole body− neonate N01−mm9 tpm muscle 28biceps femoris 29− neonate N30−mm9 tpm tongue− adult−mm9 tpm eyeball− embryo E14−mm9 tpm eyeball− embryo E15−mm9 tpm eyeball− embryo E17−mm9 tpm eyeball− neonate N01−mm9 tpm eyeball− neonate N02−mm9 tpm eyeball− neonate N00−mm9 tpm eyeball− neonate N16−mm9 tpm testis− neonate N20−mm9 tpm testis− neonate N30−mm9 tpm testis− adult−mm9
Cluster 2, subcluster human brain	tpm paracentral gyrus− adult−hg19 tpm parietal lobe− adult−hg19 tpm nucleus accumbens− adult−hg19 tpm insula− adult−hg19 tpm frontal lobe− adult−hg19 tpm postcentral gyrus− adult−hg19 tpm occipital pole− adult−hg19 tpm temporal lobe− adult−hg19 tpm brain− adult−hg19 tpm occipital lobe− adult−hg19 tpm brain− adult−hg19 tpm medial temporal gyrus− adult−hg19 tpm parietal lobe− adult−hg19 tpm middle temporal gyrus− donor10252.CNhs12316.10156.103A3.hg19−hg19 tpm parietal lobe adult− donor10196.CNhs13797.10171.103B9.hg19−hg19	-

	tpm hippocampus− adult−hg19 tpm substantia nigra− adult−hg19 tpm thalamus− adult−hg19 tpm spinal cord− adult−hg19 tpm thalamus adult− donor10196.CNhs13794.10168.103B6.hg19−hg19 tpm globus pallidus adult− donor10196.CNhs13801.10175.103C4.hg19−hg19 tpm spinal cord adult− donor10196.CNhs13807.10181.103D1.hg19−hg19 tpm locus coeruleus adult− donor10196.CNhs13808.10182.103D2.hg19−hg19 tpm occipital cortex− adult−hg19 tpm diencephalon− adult.CNhs12610.10193.103E4.hg19−hg19 tpm pons− adult−hg19 tpm globus pallidus− newborn−hg19 tpm thalamus− adult−hg19 tpm substantia nigra− adult−hg19 tpm globus pallidus− adult−hg19 tpm locus coeruleus− adult−hg19 tpm spinal cord− adult−hg19 tpm medulla oblongata− adult−hg19 tpm corpus callosum− adult−hg19 tpm optic nerve− donor1.CNhs13449.10277.104E7.hg19−hg19 tpm substantia nigra adult− donor10196.CNhs13803.10178.103C7.hg19−hg19 tpm spinal cord− fetal−hg19 tpm medulla oblongata− newborn−hg19 tpm locus coeruleus− newborn−hg19 tpm substantia nigra− newborn−hg19 tpm spinal cord− newborn−hg19 tpm putamen− newborn−hg19 tpm kidney− adult−hg19 tpm cerebral meninges− adult.CNhs12840.10188.103D8.hg19−hg19 tpm retina− adult−hg19 tpm pineal gland− adult−hg19 tpm pineal gland adult− donor10196.CNhs13804.10179.103C8.hg19−hg19 tpm medulla oblongata− adult−hg19 tpm medulla oblongata adult− donor10196.CNhs13800.10174.103C3.hg19−hg1	
Cluster 2, mixed subcluster RNA and liver	tpm SABiosciences XpressRef Human Universal Total RNA− pool1.CNhs10610.10002 tpm Universal RNA Human Normal Tissues Biochain− pool1.CNhs10612.10007.101B tpm liver− adult−hg19 tpm liver− fetal−hg19	tpm liver− neonate N25−mm9 tpm liver− neonate N30−mm9 tpm liver− neonate N20−mm9 tpm liver− adult pregnant day01−mm9 tpm Universal RNA Mouse Normal Tissues Biochain− pool1−mm9 tpm liver− neonate N07−mm9 tpm liver− neonate N10−mm9 tpm liver− neonate N06−mm9 tpm liver− neonate N00−mm9 tpm liver− embryo E18−mm9 tpm liver− neonate N03−mm9 tpm whole body− embryo E18−mm9 tpm whole body− neonate N00−mm9 tpm SABiosciences XpressRef Mouse Universal Total RNA− pool1−mm9 tpm amnion− adult pregnant day17−mm9
Cluster 2, subcluster mouse embryo liver, spleen, lung, bone etc.	tpm blood− adult−hg19	spleen− neonate N20−mm9 tpm spleen− neonate N10−mm9 tpm spleen− adult−mm9 tpm spleen− neonate N25−mm9 tpm spleen− embryo E18−mm9 tpm liver− embryo E13−mm9 tpm liver− embryo E14−mm9

		tpm liver− embryo E15−mm9 tpm liver− embryo E12−mm9 tpm bone 28os femoris 29− adult−mm9 tpm epididymis and seminiferous tubule− neonate N30−mm9 tpm heart− neonate N30−mm9 tpm heart− neonate N25−mm9 tpm heart− neonate N20−mm9 tpm adrenal gland− embryo E14−mm9 tpm whole body− embryo E13−mm9 tpm whole body− embryo E14−mm9 tpm spleen− embryo E16−mm9 tpm lung− embryo E17−mm9 tpm lung− embryo E18−mm9 tpm lung− neonate N00−mm9 tpm bone 28os femoris 29− neonate N20−mm9 tpm bone 28os femoris 29− neonate N16−mm9 tpm whole body− embryo E17−mm9 tpm whole body− embryo E16−mm9 tpm whole body− embryo E14−mm9 tpm whole body− embryo E17−mm9 tpm liver− embryo E16−mm9 tpm liver− embryo E17−mm9 tpm liver− embryo E17−mm9 tpm lung− neonate N07−mm9 tpm lung− neonate N06−mm9 tpm lung− neonate N10−mm9 tpm lung− neonate N20−mm9 tpm lung− neonate N25−mm9 tpm lung− neonate N30−mm9 tpm Clontech Mouse Universal Reference Total RNA−pool1−mm9
Cluster 3, mixed endocrine glands, pancreas, digestive system of mouse	tpm duodenum− fetal−hg19 tpm temporal lobe− fetal−hg19 tpm pancreas− adult−hg19 tpm pituitary gland− adult−hg19 tpm pituitary gland adult−donor10196.CNhs13805.10180.103C9.hg19−hg19 tpm submaxillary gland−adult.CNhs12852.10202.103F4.hg19−hg19 tpm salivary gland− adult−hg19	tpm intestine− neonate N00−mm9 tpm small intestine− neonate N16−mm9 tpm intestine− neonate N30−mm9 tpm intestinal mucosa− adult−mm9 tpm colon− adult−mm9 tpm pancreas− adult−mm9 tpm pancreas− neonate N30−mm9 tpm pancreas− neonate N25−mm9 tpm pancreas− neonate N16−mm9 tpm pancreas− embryo E16−mm9 tpm pancreas− embryo E17−mm9 tpm pancreas− embryo E18−mm9 tpm pancreas− neonate N02−mm9 tpm pancreas− neonate N01−mm9 tpm pancreas− neonate N00−mm9 tpm stomach− neonate N25−mm9 tpm stomach− adult−mm9 tpm placenta− adult pregnant day10−mm9 tpm placenta− adult pregnant day17−mm9 tpm pituitary gland− embryo E17−mm9 tpm pituitary gland− neonate N00−mm9 tpm pituitary gland− adult−mm9 tpm submandibular gland− adult−mm9

Table S8

esophagus, adult, pool1.CNhs10620.10015.101C6.hg19 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, donor1.CNhs13972.10184.103D4.hg19 HUMAN		
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.CNhs12849.10199.103F1.hg19 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

Table S8

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, pool1.CNhs11677.10093.102C3.hg19 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.CNhs12849.10199.103F1.hg19 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

Table S8

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, pool1.CNhs11677.10093.102C3.hg19 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.CNhs12854.10204.103F6.hg19 HUMAN		
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, pool1.CNhs10627.10021.101D3.hg19 Homo/Pan/Gorilla		

Table S8

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.CNhs12851.10201.103F3.hg19 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, pool2.CNhs12998.10096.102C6.hg19 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

Table S8

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, donor1.CNhs11763.10055.101H1.hg19 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, donor10258.CNhs14225.10372.105G3.hg19 HUMAN		
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, donor1.CNhs11782.10072.101I9.hg19 HUMAN		

Table S8

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, donor10258.CNhs14225.10372.105G3.hg19 HUMAN		
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, donor1.CNhs11782.10072.101I9.hg19 HUMAN		
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

Table S8

dura mater, adult, donor1.CNhs10648.10041.101F5.hg19 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