

**Silencing of transposable elements may not be a major driver of regulatory  
evolution in primate iPSCs**

Michelle C. Ward<sup>1,2\*</sup>, Siming Zhao<sup>1</sup>, Kaixuan Luo<sup>1</sup>, Bryan J. Pavlovic<sup>1</sup>, Mohammad M. Karimi<sup>3</sup>,  
Matthew Stephens<sup>1,4</sup>, Yoav Gilad<sup>1,2\*</sup>

<sup>1</sup>Department of Human Genetics, University of Chicago, Chicago, IL, USA

<sup>2</sup>Department of Medicine, University of Chicago, Chicago, IL, USA

<sup>3</sup>MRC London Institute of Medical Sciences, Imperial College, London, UK

<sup>4</sup>Department of Statistics, University of Chicago, Chicago, IL, USA

\*Correspondence should be addressed to M.C.W ([mcward@uchicago.edu](mailto:mcward@uchicago.edu)), and Y.G  
([gilad@uchicago.edu](mailto:gilad@uchicago.edu)).

**Contact:**

Cummings Life Sciences Center, 920 E. 58th Street, CLSC 317

Chicago, IL, 60637, USA

Telephone: (773) 702-8507

Fax: (773) 834-8470

**Short title:** Conservation of TE silencing in primate iPSCs

**Keywords:** Transposable elements, gene regulation, evolution

## Abstract

Transposable elements (TEs) comprise almost half of primate genomes and their aberrant regulation can result in deleterious effects. In pluripotent stem cells, rapidly-evolving KRAB-ZNF genes target TEs for silencing by H3K9me3. To investigate the evolution of TE silencing, we performed H3K9me3 ChIP-seq experiments in induced pluripotent stem cells from ten human and seven chimpanzee individuals. We identified four million orthologous TEs and found the SVA and ERV families to be marked most frequently by H3K9me3. We found little evidence of inter-species differences in TE silencing, with as many as 82% of putatively silenced TEs marked at similar levels in humans and chimpanzees. TEs that are preferentially silenced in one species are a similar age to those silenced in both species, and are not more likely to be associated with expression divergence of nearby orthologous genes. Our data suggest limited species-specificity of TE silencing across six million years of primate evolution.



## Introduction

Over half of primate genomes are annotated as transposable elements (TEs) (1, 2). Primate TE sequences reflect their evolutionary history; some TEs are conserved over multiple phylogenetic clades and orders, whereas others are restricted to particular lineages. Considering the degree of sequence similarity between the human and chimpanzee genomes, it is not surprising that many of the same TEs are present in both species (3, 4). In general, TE activity, especially from endogenous retroviruses (ERVs), has declined in the hominid lineage relative to other mammalian lineages (5). However, a fraction of evolutionarily recent TEs are active in the human genome, including HERV-K ERVs (6-10), members of the Alu (11, 12), L1 (13, 14), and SVA (15, 16) families. Most studies on mammalian TEs have focused on humans and mice, with a handful that describe the ERV and L1 TE families in chimpanzees (17-19).

Because of their transposition competence, TEs are considered a potential source of regulatory innovation (20, 21). Indeed, TEs can serve as primate-specific regulatory sequence (22, 23), act as enhancer elements (24-26), carry transcription factor binding sites (27-30), and contribute themselves to the transcriptome (31-33). The influence of TEs on transcriptional programs has been studied in many broad contexts from the immune system (34) to pregnancy (35), to pluripotency (36). Yet, ultimately, the regulatory impact of TEs has not been extensively studied at a genome-wide scale in a comparative framework.

Earlier work using a Down syndrome mouse model carrying a copy of human chromosome 21 showed that in general, genes on human chromosome 21 are similarly regulated regardless of whether the chromosome is in a human or a mouse cellular environment (37). However, human-specific TEs become aberrantly activated (marked by H3K4me3) on human chromosome 21 in the

mouse cellular environment (38). It was therefore speculated that host repressive factors needed to properly silence these human TEs may be absent in the mouse nuclear environment, which is ~60 million years diverged from human. This suggested a relatively rapid evolution of TE silencing mechanisms, in line with the evolutionary arms race hypothesis.

Indeed, due to the potential deleterious effects of new TE insertions, especially in the context of developmental regulatory programs, TE activity is thought to be tightly controlled by host repressive machinery. In particular, it has been shown that in embryonic stem cells (ESCs), KRAB-containing Zinc Finger genes (KRAB-ZNFs) can recognize TEs in a sequence-specific manner, recruit the co-factor TRIM28 (also known as KAP1), and histone methyltransferase SETDB1 (also known as ESET), to effect silencing through the deposition of the repressive histone three lysine nine trimethylation (H3K9me3) modification (39-45). After the establishment of TE silencing by H3K9me3, TEs are subjected to *de novo* DNA methylation, which persists in somatic and germ cells (46-53). Some TEs can evade detection by host KRAB-ZNFs by acquiring mutations. It was thus hypothesized that the host repressor mechanisms need to evolve rapidly in order to maintain genome integrity (54-56).

TEs are silenced by KRAB-ZNFs particularly in early embryonic development (57), their most likely niche for retrotransposition (58); however it is difficult to obtain biological material from humans and other primates to determine how this process evolves. Induced pluripotent stem cells (iPSCs) provide a model of the cells of the embryonic inner cell mass to sidestep this challenge. This cell type is of further interest as we know of examples of primate-specific TEs that are not silenced, but actively transcribed and functionally relevant in this cell type in human (notably HERVH elements) (9, 59-64). This might imply specificity of TE regulatory mechanisms, and hints at the fact that

evolutionarily recent TE sequence could play a role in differentiating gene regulatory networks between species.

The establishment of a panel of non-retroviral reprogrammed iPSCs from multiple humans and chimpanzees has allowed us to comprehensively characterize the extent of TE silencing in both species (65-67). In order to gain insight into how TE regulation, specifically TE silencing, evolves between humans and chimpanzees, we globally profiled the predominant output of KRAB-ZNF-targeted silencing mechanisms in pluripotent cells: the histone modification H3K9me3. Our goals were to identify inter-species differences in TE silencing patterns, characterize gene regulatory divergence that can potentially be explained by differences in TE silencing between species, and develop hypotheses regarding the mechanisms that underlie such inter-species differences in TE regulation. Our observations led us to conclude that, at least in our comparative iPSC system, differences in TE silencing do not drive substantial gene regulatory divergence between these two closely related species.

## Results

To perform a comparative study of TE silencing, we characterized genome-wide distributions of the repressive histone modification H3K9me3 in iPSCs from seven chimpanzees and ten humans (Figure 1A, Table 1 in Supplementary File 1). The iPSC lines we used were deeply characterized in this study and previously (see Methods; and (65-67)). Briefly, all cell lines were confirmed to have a normal karyotype (Figure 1-Figure Supplement 1A), have the ability to form embryoid bodies that consist of cells from the three main germ layers (Figure 1-Figure Supplement 1B), display global gene expression signatures consistent with pluripotent cells (Figure 1-Figure Supplement 2), and

express key pluripotency factors detectable by immunocytochemistry (Figure 1-Figure Supplement 3). All samples were processed in species-balanced batches at each experimental step (Table 2 in Supplementary File 1).

We characterized the distribution of H3K9me3 in the 17 iPSC lines by performing chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). In order to increase genome-mapping confidence, paired-end sequencing libraries were prepared, and only high quality, properly paired reads retained (see Methods and Figure 1-Figure Supplement 4 and Table 3 in Supplementary File 1). As a first measure of quality of our data, consistent with previous reports, we observed an enrichment of H3K9me3 ChIP-seq reads at *ZNF* genes in both species (Figure 1-Figure Supplement 5)(68).

We identified regions of broad H3K9me3 ChIP enrichment in each individual independently (at an FDR of 10%; see Methods and Figure 1-Figure Supplement 6). A read sub-sampling analysis indicates that we sequenced our samples to a depth that allows us to detect most enriched regions (Figure 1-Figure Supplement 7). In order to compare H3K9me3 enrichment across species, we focused exclusively on reciprocally best match orthologous regions in human and chimpanzee (see Methods). We excluded from our data peaks that mapped outside those orthologous genomic regions (approximately 10% of the ChIP-seq data were excluded; Figure 1-Figure Supplement 8 and Table 3 in Supplementary File 1). We further filtered the data based on overall genome mappability, to exclude regions in which sequencing reads can be mapped to one species at a substantially higher probability than the other (using a cutoff of 0.8 in both species; see Methods). We defined orthologous H3K9me3 ChIP-seq regions as those where a ChIP-seq peak, contained within orthologous human-chimpanzee genomic regions, was identified in at least one individual, regardless of species (see Methods). We refer to this combined set as 160,113 orthologous ChIP-

seq regions (rather than ChIP-seq peaks), because these regions are defined across all individuals, while the ChIP-seq peak may have been identified in only a subset of individuals.

The definition of ChIP-seq regions allows a quantitative comparison of H3K9me3 enrichment across species. We therefore calculated the number of sequencing reads that were mapped to each orthologous ChIP-seq region in each individual. By using this approach, we sidestep the difficult challenge of accounting for incomplete power to detect 'significant' ChIP-seq peaks in each individual or species independently. To avoid considering regions where we clearly have insufficient power to compare read counts across species, we included only the 150,390 regions with at least one read count in more than half of the individuals, regardless of species.

We considered overall properties of the read count data in the ChIP-seq orthologous regions. As expected, data from different individuals cluster by species (with the exception of one human individual, which therefore was removed from subsequent analysis; Figure 1-Figure Supplement 9-10), and inter-species variation in read counts is somewhat greater than intra-species variation (Spearman's correlation in human vs. human = 0.90, chimpanzee vs. chimpanzee = 0.89, and human vs. chimpanzee = 0.80, Figure 1-Figure Supplement 11). We then used the framework of a linear model with a fixed effect for species to identify orthologous ChIP-seq regions with differences in H3K9me3 enrichment between humans and chimpanzees (Figure 1B, Supplementary File 2, and Methods). At  $FDR < 0.01$ , we classified 16,288 (11%) regions as differentially enriched between the two species (Figure 1C-D, Figure 1-Figure Supplement 12 and see examples in Figure 1E). The proportion of differentially enriched regions is robust with respect to our data filtering choices (Figure 1-Figure Supplement 13 and Tables 4 in Supplementary File 1), and the choice of approach to analyze the data (Table 5 in Supplementary File 1).

## **Orthologous TEs tend to be silenced more often than species-specific TEs**

To compare TE silencing by H3K9me3 across species, we identified a comprehensive set of 4,036,865 autosomal orthologous TEs in the human and chimpanzee genomes (80.2% of all annotated human TEs, see Methods and Figure 2A, Figure 2-Figure Supplement 1). We considered the ChIP-seq data in the context of the orthologous TEs, and observed that 12% of orthologous TEs (436,049 instances) overlap an orthologous H3K9me3 region with at least 50% of their length (denoted 'overlapping', Figure 2B and Supplementary File 3; we note that our subsequent conclusions are robust with respect to the specific cutoff used to classify a TE as overlapping with H3K9me3 regions). The TEs that do not overlap with an H3K9me3 region ('non-overlapping'; 3,600,816 instances), and are putatively not silenced by this repressive mechanism, are an equally interesting category of elements and are therefore included in subsequent analyses. Consistent with previous reports (42, 50), we found enrichment of LTRs (Pearson's Chi-squared test;  $P < 0.001$ ), and SVA elements ( $P < 0.001$ ), among TEs that overlap H3K9me3 regions (Figure 2C).

Given the proposed evolutionary arms race between TEs and the host repressive machinery (69), we sought to investigate the within-species patterns of silencing at both orthologous and non-orthologous TEs. We further categorized non-orthologous TEs into those that are only annotated in that species (defined as 'species-specific TEs'; for example, SVA-E & SVA-F, which are only annotated in human; Table 6 in Supplementary File 1). Generally, we found that the same TE classes that most frequently overlap with H3K9me3 regions in the orthologous TE set, also most frequently overlap with H3K9me3 in the non-orthologous TE set (Figure 3). However, within each species, the SVA and LINE TE classes show a reduction in the proportion of TEs overlapping H3K9me3 regions in the non-orthologous, and species-specific categories ( $P$  (SVA)  $< 0.001$ ;  $P$

(LINE)  $< 0.001$ ); Figure 3-Figure Supplement 1). The magnitude of the effect differs between these two TE classes: For example, 85% of orthologous SVA elements overlap H3K9me3, but only 41% of non-orthologous SVA elements overlap the regions marked by H3K9me3. In contrast, 17% and 14% orthologous and non-orthologous LINE elements overlap H3K9me3 regions, respectively. In this case, the difference in effect size could reflect heterogeneity in silencing amongst the diverse elements within the large LINE class. We also found a reduction of TE and H3K9me3 overlap in human LTRs ( $P < 0.001$ ), but an increase in H3K9me3 overlap with LTRs that are chimpanzee-specific ( $P < 0.001$ ; Figure 3). In general, these observations are consistent with the notion that more recent TEs are generally less likely to be silenced, in line with the evolutionary arms race theory (54-56).

### **The majority of orthologous TEs are similarly silenced in human and chimpanzee**

To determine the extent to which orthologous TEs are similarly silenced in both species, we performed a comparative quantitative analysis of the ChIP-seq read counts in the orthologous TEs that overlap H3K9me3 regions. As mentioned previously, our comparative analysis is not dependent on the identification of 'significant' ChIP-seq peaks in both species. Instead, we tested for evidence of differences in H3K9me3 read counts between humans and chimpanzees in all ChIP-seq regions. When we could not reject the null hypothesis (at FDR  $< 0.01$ ) that similar numbers of H3K9me3 ChIP-seq reads are mapped to orthologous TEs, we referred to the silencing status of these TEs as 'shared'. We thus classified TEs into the following categories (Supplementary File 1 - Table 2): (i) TEs that do not overlap with an orthologous H3K9me3 region (non-overlapping); (ii) TEs that are associated with similar enrichment of H3K9me3 in both species (shared); (iii) TEs that are associated with significantly higher enrichment of H3K9me3 in humans or (iv) in chimpanzees.

We observed that the proportion of TEs that overlap with H3K9me3 regions differs substantially across different TE classes (for example, 60% of SVA elements overlap H3K9me3 regions compared to only 10% of DNA transposons). Yet, the proportion of TEs for which we have evidence for inter-species differences in silencing is consistent, regardless of TE class (Figure 4A). Indeed, while ultimately most TEs (88%) do not overlap H3K9me3 regions, when TEs do overlap H3K9me3 regions, they are generally (82% of TEs) silenced to the same extent in the two species. Our analysis includes hundreds of thousands of TEs that overlap H3K9me3 regions; thus even just 18% of TEs that overlap H3K9me3 and are silenced differently across species still comprise of tens of thousands of instances (53,201 TEs enriched for H3K9me3 in human, and 28,110 TEs in chimpanzee).

Thousands of instances of differential silencing notwithstanding, the overall emerging pattern is that most TEs are silenced similarly between species. Yet, it is challenging to draw conclusions based on the inability to reject a null hypothesis of no differences between the species. We therefore used an alternative approach, based on an assumption of a unimodal distribution of effects (ashr) (70) to obtain a more sensitive estimate of the proportion of inter-species differences at different effect sizes. This unimodal assumption in an Empirical Bayes framework produces smaller (more conservative) estimates of the proportion of tests under the null through adaptive shrinkage. This test is therefore more sensitive to smaller differences in effect size across species. Using this approach, we again find a high degree of conservation in TE silencing, with practically all inter-species difference in silencing associated with small effect sizes (smaller than 2 fold; Figure 1-Figure Supplement 14).

Given the diversity of TEs within a class, we proceeded by considering a higher resolution classification of TEs to the eleven most predominant TE families representing each of the five TE



classes, and then to TE types within specific families (Figure 2). There is some degree of heterogeneity in inter-species silencing differences ('effect size') between different TE families (75-84% depending on the family; Figure 4-Figure Supplement 1). We particularly focused on the LTR and LINE classes, given the previously reported association of these TEs with TRIM28-mediated silencing mechanisms (42, 71). Human LTRs are comprised of HERV elements which can be alternatively classified into Class I (HERV-F,H,I,E,R,W families), Class II (HERV-K), and Class III (HERV-L), based on the original retrovirus that integrated into the genome (72). Using a subset of these HERV types curated by Turelli *et al.*, where the annotation is likely to be correct, we observed considerable heterogeneity in the frequency with which different HERV types overlap H3K9me3 regions (Figure 4-Figure Supplement 2A). Notably, less than 20% of 2,079 orthologous HERVH elements, which are generally known to be expressed in pluripotent cells, overlap with H3K9me3 regions (compared to 80% of 3,484 THE1B elements, which are not known to be expressed in pluripotent cells, for example). This heterogeneity is also reflected in the proportion of TEs that are similarly silenced in the two species (55-95% sharing, Figure 4-Figure Supplement 2B). However, the total number of elements within the more divergent TE types is small (fewer than 100 TE instances per category with less than 75% sharing). In contrast, when turning our attention to a subset of mammalian-specific L1 LINE elements (71), we found less variability in overlap with H3K9me3 across types (5-25% are overlapping; Figure 4-Figure Supplement 3A). However, of those L1 elements that overlap H3K9me3, the majority (82% of 21,710 elements) are similarly silenced in both species (Figure 4-Figure Supplement 3B). These results suggest that there is some degree of sequence specificity in the silencing mechanisms across TE types.

## **Silenced TEs tend to have similar functional properties across species**

We next sought to identify properties that might distinguish TEs based on our silencing classifications (shared, species-specific, or not silenced). We first considered the length of a TE, which may relate to its transposition competency and hence need for silencing. On average, TEs silenced in at least one species are longer than those that do not overlap orthologous H3K9me3 regions (Wilcoxon-rank sum test;  $P < 0.001$ ; Figure 5A). To confirm that longer TEs do not overlap H3K9me3 more often by chance alone, we artificially increased the length of the non-overlapping TEs to match the median length of TEs marked by H3K9me3 in both species. The artificially longer TEs do not overlap with H3K9me3 regions more often, suggesting that TE elements are silenced in a non-random manner, which cannot be simply explained by length in itself. We found the same pattern when we analyzed TEs by class, which in itself is associated with different lengths ( $P < 0.001$ ; Figure 5-Figure Supplement 1A). Interestingly, across TE families there is a strong correlation between the proportion of TEs that overlap H3K9me3 regions and the median length of the TEs in the family (Pearson's correlation = 0.93; Figure 5-Figure Supplement 1B-C). However, within families the median length of TEs that do not overlap H3K9me3 is highly similar to that of TEs that do (Pearson correlation for median TE lengths across families = 0.99; Figure 5-Figure Supplement 1D), again suggesting that this is not a feature of length itself. When we stratified TEs by the magnitude of inter-species silencing divergence (absolute effect size), we found that shorter TEs are associated with higher silencing divergence ( $P < 0.001$ ; Figure 5B). Put together, these observations suggest that longer TEs may be silenced more often because they are more likely to have unwanted regulatory potential, and are therefore more likely to contain sequences that recruit silencing machinery. This is consistent with the finding that more KRAB-ZNF proteins (sequence-specific mediators of TE silencing) bind longer TEs (73).

To gain insight into the relationship between TE silencing and TE transposition and mobility, we asked whether TE copy number is correlated with silencing. Across TE families, we found only a

weak correlation between TE copy number and the proportion that overlap with H3K9me3 regions (Pearson's correlation = -0.47; permutation  $P = 0.15$ ; Figure 5-Figure Supplement 2). Moreover, in contrast to the general trend, TEs in the two smallest families (780 SVAs and 7,983 ERVKs) overlap H3K9me3 regions most often (Figure 5-Figure Supplement 2).

Given that TEs can acquire mutations and transpose over time, we compared the sequence divergence between each individual TE instance and the corresponding consensus TE sequence (see Methods). We found that TEs that are silenced similarly in both species are evolutionarily less diverged from the consensus sequence than TEs that do not overlap H3K9me3 regions ( $P < 0.001$ ; Figure 5C). This pattern is observed within the LTR, LINE, SINE and DNA TE classes ( $P < 0.001$ ; Figure 5-Figure Supplement 3A), but this pattern does not hold within all TE families, as might be expected considering the diversity of TE types and families within a class (Figure 5-Figure Supplement 3B). Across all TEs, we did not find a difference in sequence divergence between TEs that are silenced in just one or both species (Figure 5C) but stratification of TEs by the magnitude of the inter-species silencing effect reveals that when inter-species differences in silencing are larger, the sequences of the TEs are more diverged from the consensus sequence ( $P < 0.001$ ; Figure 5D).

We next considered proximity of TEs to genes by determining the distance between each TE and the closest orthologous annotated Transcription Start Site (TSS). TEs that are silenced in both species are significantly further away from the TSS than TEs that are preferentially silenced in one species, or those that do not overlap with H3K9me3 regions ( $P < 0.001$ ; Figure 5E). Again, this pattern is observed within the LTR, LINE, SINE and DNA classes, but not within all TE families ( $P < 0.001$ ; Figure 5-Figure Supplement 4). TEs that show greater divergence in silencing across species are also more likely to be closer to the TSS than TEs with lower divergence ( $P < 0.001$ ;

Figure 5F). These results suggest that, across families, there are some differences in the functional properties of silenced TEs.

In order to determine whether silenced TEs are also associated with other genomic features, we overlaid orthologous TEs with chromatin state annotations defined in human ESCs (74). The majority of TEs, regardless of silencing status, are associated with chromatin regions annotated as 'heterochromatin and low signal'. We found that TEs that are preferentially silenced in human overlap annotated regulatory regions less frequently compared with TEs that are preferentially silenced in chimpanzee ( $P < 0.001$ ; Figure 5G and Figure 5-Figure Supplement 5). We also considered published ChIP-seq and DNase hypersensitivity (DHS) data in H1 hESCs. We found that a relatively small fraction of TEs overlap regions of open chromatin or ChIP-seq peaks: Less than 20% of TEs overlap active histone marks and less than 1% of TEs overlap DHS. We found that TEs that are preferentially silenced in human overlap regions of active gene regulation less often than TEs preferentially silenced in chimpanzee. The active chromatin annotations we considered include marking by H3K27ac, DHS, enhancer elements (marked by H3K4me1, p300) and regions marked by RNA polymerase II occupancy (for all comparisons  $P < 0.01$ ; Figure 5-Figure Supplement 6). Notably, there is no significant difference between TEs preferentially silenced in either species with respect to overlap with the active promoter-associated mark, H3K4me3.

### **TE silencing is not a major driver of inter-species gene expression divergence**

To determine the potential impact of TE silencing on the regulation of neighboring genes, we characterized gene expression levels in the same iPSC lines from both species (using RNA-sequencing; see Methods, Figure 6A, Figure 6-Figure Supplement 1A and Tables 2-3 in Supplementary File 1). After restricting the data to a set of human-chimpanzee orthologous exons

(75) the RNA-seq data clusters primarily by species, and then by population (i.e. Caucasian or Yoruba / cell-type of origin; Figure 6-Figure Supplement 1B). Indeed, after filtering lowly expressed genes, species emerges as the primary driver of variation in the RNA-seq data, as expected (Figure 6-Figure Supplement 1C).

To characterize the level of TE silencing associated with each orthologous gene, we defined a *cis*-regulatory region 10 kb upstream of the TSS of each gene that includes at least one annotated TE (Methods and Figure 6A). As expected, we observed that gene expression levels decrease when TEs in the *cis* regulatory window overlap with H3K9me3 regions ( $P < 0.01$ ; Figure 6B). This observation is consistent with the general reported role of the H3K9me3 histone modification in gene regulation (76, 77). The effect of TE silencing on gene expression is recapitulated when extending the window size to 20 kb, while TEs silenced within 1 kb upstream of the TSS have a stronger effect on gene expression (Figure 6-Figure Supplement 2). We then asked whether inter-species differences in TE silencing are associated with expression divergence of nearby genes.

To obtain a quantitative measure of overall TE silencing in each species, we summed H3K9me3 ChIP-seq read counts from orthologous H3K9me3 regions overlapping TEs within 10 kb of annotated genes. We used the same linear model framework described above to estimate inter-species difference in H3K9me3 counts, within a 10 kb window upstream of the TSS. Of the windows that contain at least one orthologous H3K9me3 region overlapping a TE, 79% have similar levels of H3K9me3 enrichment in human and chimpanzee (3,523), with 11% of regions enriched in human (499), and 10% of regions enriched in chimpanzee (463), at FDR of 1% (Figure 6C). We then considered inter-species differences in expression levels for genes associated with similar or species-specific H3K9me3 enrichment at TEs.

We found no significant difference in the expression divergence effect size between genes associated with species-specific or shared silenced TE, using 10, 20 or 40 kb windows upstream of the TSS (Figure 6C and Figure 6-Figure Supplement 3). We did find a weak, negative association between gene expression and silencing divergence when we considered silenced TEs that are located within 1 kb upstream of the TSS (only 165 genes in that set; Figure 6-Figure Supplement 3 and 4; Table 7 in Supplementary File 1). It is tempting to conclude that TEs (or silenced TEs) near the TSS can have a marked contribution to expression divergence; however, H3K9me3 enrichment in regions that overlap a 1 kb region around the TSS is similarly associated with expression divergence, irrespective of the presence of a TE (Figure 6-Figure Supplement 4). Our observations are robust with respect to a wide range of FDR cutoffs used to classify inter-species differences in silencing, suggesting that incomplete power is not a likely explanation for the overall lack of association between expression divergence and TE silencing (Table 8 in Supplementary File 1).

To further investigate the overall lack of association between silencing and gene expression divergence, we performed an analogous analysis, this time using gene expression divergence as the anchor. We found no association between silencing and inter-species expression divergence: genes that are classified as differentially expressed between humans and chimpanzees are not more likely to be associated with inter-species differences in H3K9me3 at TEs in their putative regulatory regions (Figure 6D and Figure 6-Figure Supplement 5A). Again, we could only find a relationship between silencing and expression divergence when considering H3K9me3 overlapping TEs within 1 kb upstream of the TSS. The observation of no association of inter-species expression divergence and TE silencing is robust with respect to a wide range of statistical cutoffs used to classify either differential expression or differential silencing (Figure 6-Figure Supplement 5B and Table 9 in Supplementary File 1). Indeed, when we considered the entire distribution, the

correlation between inter-species effect sizes for gene expression divergence and H3K9me3 enrichment is weak (Pearson's correlation = -0.0005 in a 10 kb window; the correlation in divergence increases somewhat to -0.14 when considering a 1 kb window upstream of the TSS; Figure 6-Figure Supplement 3). Moreover, we observed nearly as many genes whose expression divergence is consistent with the observed patterns of TE silencing divergence (namely, TE silencing associated with lower expression level) as those that show the opposite, inconsistent pattern (Figure 6E).

We looked at the relationship between expression divergence and TE silencing more closely by considering orthologous TEs that are most likely to regulate neighboring gene expression. Specifically, we focused on the 532 genes closest to TEs that overlap either enhancer (marked by H3K4me1) or promoter (marked by H3K4me3) regions in human, and yet are silenced in chimpanzee (Figure 5-Figure Supplement 6). The results are consistent with our observations for the entire set. We found a relationship between expression divergence and silencing of TEs that overlap the promoter-associated mark but not of TEs that overlap enhancers (typically further from the TSS; Figure 6F and Figure 6-Figure Supplement 6).

We next reasoned that perhaps TEs that are actively expressed in this cell type, in at least one species, are more likely to affect gene expression divergence. Indeed, TEs can be transcribed and contribute to lncRNA and neighboring gene expression (4, 50). Using our RNA-seq data, we identified eleven poly-adenylated TE types that are expressed in chimpanzee, and seven that are expressed in human (Methods; Table 10A in Supplementary File 1; note that this approach does not allow us to capture non-polyadenylated transcripts such as eRNAs). In both species, we identified the expression of HERVH elements, which are known to be expressed in human pluripotent cells. Of the nine TE types that are annotated in both species and expressed in at least

one species, 1,025 instances are contained within our orthologous TE set (Table 10B in Supplementary File 1 and 123 instances overlap an H3K9me3 region. Only 14 of these TE instances are enriched for H3K9me3 in human, and 22 in chimpanzee ( $FDR < 0.01$ ). These TEs are proximal (within 10 kb) of only 19 orthologous genes. Thus, due to the small numbers we are unable to properly test the association between TE silencing/expression and gene expression divergence. In any case, even if there is such an association in these cases, the small number of instances suggests that gene regulation by an actively expressed poly-adenylated TE in one species is not a major mechanism of gene expression divergence.

Further attempting to find a relationship between TE silencing and expression divergence, we hypothesized that perhaps imperfect annotation of orthologous TEs may affect our observations. A clear association between silencing of non-orthologous TEs and gene expression divergence would provide some measure of support for this possible explanation. We thus performed a similar analysis focused only on the non-orthologous TEs. In this analysis we considered non-orthologous TEs that are located 10 kb upstream of annotated orthologous genes. We classified the non-orthologous TEs as 'silenced' or 'not silenced' based on all H3K9me3 regions identified in that species (we did not require H3K9me3 regions to be orthologous in this analysis). We found, once again, that there is no association between TE silencing and expression divergence (Figure 6-Figure Supplement 7). A significant effect is only consistently observed when considering a 1 kb window immediately upstream of the TSS (a set that includes 289 genes; Figure 6-Figure Supplement 7). In other words, TE silencing by H3K9me3, regardless of orthology, has a modest effect on gene expression divergence in human and chimpanzee iPSC lines.

We next focused our attention on known mediators of TE silencing, the KRAB-ZNF genes, in an attempt to identify species-specific regulatory effects of TE silencing. KRAB-ZNF genes, in addition



to directing the repression of TEs, are themselves marked by H3K9me3 (68, 78). We intersected a curated list of 256 KRAB-ZNF genes (54), filtered for orthology, with our list of orthologous H3K9me3 regions. We found that 104 (41%) of these genes overlap with orthologous H3K9me3 regions (Figure 6-Figure Supplement 8A). The proportion of KRAB-ZNFs with similar H3K9me3 enrichment in both species (78% of overlapping genes or 81 instances) is comparable to that observed for TEs (Figure 6-Figure Supplement 8B). 17 KRAB-ZNFs are enriched for H3K9me3 in human and 6 are enriched for H3K9me3 in chimpanzee. We found that the 256 KRAB-ZNF genes are no more likely to be differentially expressed between human and chimpanzee than all 17,354 genes (52% of KRAB-ZNFs are classified as differentially expressed and 49% of all genes; FDR of 1%; Figure 6-Figure Supplement 9A and Table 10 in Supplementary File 1). There are 14 KRAB-ZNF genes whose TSS overlaps an H3K9me3 region that is enriched in human (8) or chimpanzee (6). The association between silencing and gene expression divergence is consistent with the pattern we observed for all genes whose TSS overlaps an H3K9me3 region (Figure 6-Figure Supplement 4) but due to the small numbers in each category, we cannot reject the null hypothesis of no association (Figure 6-Figure Supplement 9B).

Finally, as H3K9me3 is not the only repressive histone modification in the genome, it is possible that we do not observe a strong association between silencing and expression divergence because there is redundancy in histone-based TE silencing mechanisms. To address this possibility, we considered genome-wide data from the Polycomb-repressed associated histone modification H3K27me3 in human hESCs. We found that TEs preferentially silenced by H3K9me3 in chimpanzee are not associated more often with increased levels of H3K27me3 compared to TEs preferentially silenced by H3K9me3 in human (Figure 6-Figure Supplement 10). This suggests that H3K27me3 and H3K9me3 are not redundant in silencing TEs across species.

## Discussion

The potential for TEs to participate in the regulation of gene expression has received increasing attention as genomic technologies to study these elements advance (79-82). However, the role of TEs can be paradoxical. They can serve as regulatory sequence when they are bound by transcription factors, or localize in regions of open chromatin. Conversely, TEs can be actively silenced to prevent unwanted activity in a cell-type dependent manner. One cell type where TE activity can be particularly deleterious to genome integrity is embryonic stem cells. Indeed, it was found that TE activity is generally restricted through repressive histone modifications in mouse embryonic stem cells (83, 84). TE activity and regulation has been studied in many contexts, but there are no published comparative studies of TE silencing in primates.

In order to gain insight into how TE silencing mediated by the TRIM28/SETDB1 pathway evolves in pluripotent cells from the hominid lineage, we profiled the distribution of the repressive histone modification H3K9me3 in human and chimpanzee iPSCs. The rationale behind our approach is the expectation that TEs that are marked by H3K9me3 are silent and are therefore not mobile, nor able to act as regulatory elements in this cell type. In contrast, TEs that are unmarked and not silenced, could be expressed, or capable of transposition, or participation in gene regulation. Alternatively, these elements may have acquired sufficient mutations to lose the ability to transpose, or influence gene regulatory programs, and are therefore not silenced because they do not 'threaten' genome integrity. We were particularly interested in finding differences in silencing of the same TEs

between species, as this could imply that these TEs act as regulatory elements in only one species, and could contribute to phenotypic differences between species.

We found that there are differences in the TE families that are preferentially silenced in both species. This suggests specificity of silencing and targeting mechanisms, which are shared in humans and chimpanzees. In our study, up to 60% of SVA, 50% of ERV, and 15% of L1 families overlap with H3K9me3 in both species. It is likely that TE silencing in our human and chimpanzee iPSC lines is mediated by the TRIM28 co-repressor as it is known to be involved in silencing ERVs in mouse ESCs (40), and SVAs, HERVs and L1 elements in human ESCs (42, 71). This specificity is likely achieved through sequence-specific KRAB-ZNFs that recruit TRIM28. Indeed, specific KRAB-ZNFs recognize specific TE families (69, 73, 85).

While TEs silenced by H3K9me3 in both species tend to be less diverged from the corresponding TE consensus sequence than those that are not silenced, TEs that have large inter-species differences in silencing (more than 2 fold) tend to resemble TEs that are not silenced. This could imply that TEs whose silencing differs across species are less likely to require silencing, or that they are more likely to have enhancer function in just one species. Indeed, amongst TE-derived enhancers defined by CAGE in human, there is a depletion of younger, primate-specific TEs (86). Assuming an evolutionary arms race model, one expects that TEs and their KRAB-ZNF repressors would co-evolve; indeed there is a correlation between the age of KRAB-ZNF genes and their target TEs (41), and related KRAB-ZNFs bind related TEs (85). However, the recent observation that many KRAB-ZNF gene – TE pairs arose at the same time during evolutionary history, and that this relationship is maintained to this day despite the fact that these TEs are no longer able to transpose, challenges the idea of a broad evolutionary arms race (73). In our study, we found silencing by H3K9me3 in TEs that are no longer transposition competent. It could be that these

TEs are not silenced to prevent their mobility, but rather to prevent their potential regulatory activity.

TEs that are silenced in both species are located further away from the TSS than TEs that are silenced preferentially in one species. This difference between the 'shared' and 'species-specific' silenced TEs indicates that we do not have an overwhelming proportion of false negatively classified TEs in the 'shared' group. This observation also raises the possibility that TEs that are silenced only in one species are more likely to contribute to gene expression differences between species. However, we did not find this to be the case. Within species, genes with silenced TEs that overlap promoters have lower expression levels, on average, than genes with proximal TEs that are not silenced. Across species, inter-species differences in TE silencing are only associated with divergence in gene expression when the TEs are located in promoter regions, within 1 kb of the TSS. Further from genes, we focused on TEs that overlap enhancer regions, as Imbeault *et al.*, have shown that KRAB-ZNF gene-bound TEs only affect neighboring gene expression if the TE overlaps marks of enhancers (73). In our data, however, we did not find such an effect.

Our observation of overall lack of association between TE silencing and expression divergence should be considered in the following context: First, while it is intuitive to relate a promoter to a gene, it is more challenging to relate a distant enhancer to the gene it regulates. Second, it has recently been shown that while many TEs contribute to primate-specific *cis*-regulatory elements (CREs), many of these TEs act as transcriptional repressors and hence silencing of these elements may in fact lead to increased expression of neighboring genes (23). Third, we did not investigate the functional consequences of sequence changes to orthologous TEs, so our results do not account for the fact that mutations in the sequence of TEs in one species may inactivate them, eliminating the need for histone-based silencing. Fourth, we have restricted our analysis to

orthologous genes; it is possible that regulatory novelty co-occurs with gene novelty, though admittedly this is rare when the species considered are human and chimpanzee. Fifth, we were unable to interrogate silencing of recently active TEs, due to our stringent H3K9me3 ChIP-seq read mapping strategy, which excludes reads that do not map uniquely.

Overall, we identified tens of thousands of inter-species differences in silencing at individual TE instances; however considering that there are roughly four million human-chimpanzee orthologous TE sequences in the genome, these differences account for a minority of elements (2%). Of the TEs that are silenced, the observed level of conservation of TE silencing in human and chimpanzee is in line with what is observed for all H3K9me3 regions in the genome irrespective of the presence of a TE. Conservation of active regulatory regions, determined by H3K27ac, in human and chimpanzee is also thought to be similar (87). It may be that even divergence at a small fraction of regulatory elements between human and chimpanzee – reflecting tens of thousands of instances – is sufficient to explain much of the observed divergence in gene expression levels. Yet, though we performed multiple analyses to uncover such a relationship, we concluded that, at most, inter-species differences in TE silencing have a modest effect on gene expression divergence in iPSCs.

We acknowledge that proving the absence of a species-specific effect is challenging. In general, it is difficult to draw strong conclusions based on the inability to reject a null hypothesis. Nevertheless, our observations are robust with respect to a wide range of statistical thresholds used to identify TE silencing or gene expression differences. Our observations should also be interpreted in the context of our initial assumption about what can be learned from characterizing TE silencing by H3K9me3 in iPSCs. First, while H3K9me3-mediated TE silencing is a well-established repressive mechanism in pluripotent cells, additional mechanisms of TE silencing exist;

including RNA interference, CpG methylation and other repressive histone modifications such as H3K27me3 (88-91), and thus our findings may not extend to these other silencing pathways. For example, Marchetto *et al.*, showed that there is differential regulation of L1 elements by piRNA-associated PIWIL2 and APOBEC3B in human and chimpanzee iPSCs (19). Pluripotent stem cells can be considered to be an usual cell type in many respects, including TE regulation. In these cells, TE activity is known to be regulated by a histone-based silencing mechanism following the global resetting of DNA methylation. While there are examples of this TE repression pathway in differentiated cell types (92, 93), it is largely believed that DNA methylation replaces histone-mediated repression following the DNA de-methylation phase in embryogenesis. Therefore alternative TE silencing mechanisms predominant in other cell types, such as DNA methylation, may be more dynamic across species. Similarly, one should also consider that the ability of TEs to act as enhancer elements has been shown to be cell-type specific (26). Our results may therefore not extend to other somatic cell types, which are not under the same level of regulatory constraint as iPSCs.

Our results complement several previous studies, which explored the role for transposable elements in primate gene regulation. The primate-specific HERVH LTR element is a human pluripotency marker (62), and has been shown to have similar levels of expression in chimpanzee iPSCs (4). Indeed, we find HERVH to be expressed in human and chimpanzee iPSCs, and to be depleted for H3K9me3, suggesting that this element is essential for the pluripotency network in both humans and chimpanzees. Many open chromatin regions in human are derived from primate-specific TEs (22), and newly evolved CREs are enriched in young SVA and LTR elements (23). Trizzino *et al.* found that 2% of all LTRs are active (marked by H3K27ac) in liver but a quarter of genes with an LTR insertion affect gene expression divergence between multiple primate species, and that less than 1% of CREs are differentially active between human and chimpanzee. This

suggests that while many TEs, specifically young TEs, are important for influencing inter-species differences in expression, the majority of TEs in the genome do not act as regulatory sequences. When considering the broader context for the contribution of mammalian TEs to mediating inter-species differences in gene regulation, previous elegant work has dissected the role of specific TEs in inter-species differences between mice and rats (25, 94). TE activity is substantial in the rodent lineage compared to hominids (5), which suggests that TEs can more quickly rewire regulatory networks in these species. Consistent with the reduction in TE activity in primates compared to other mammals, ChIP-seq assays for the highly conserved transcriptional insulator CTCF, have shown reduced TE-mediated propagation of CTCF binding sites in primates compared to other mammalian lineages (30, 95).

In summary, to date, there have been few genome-wide studies demonstrating the contribution of TEs to gene regulation in primates. TEs have been shown to comprise most primate-specific regulatory sequence (22), to contribute to primate-conserved non-coding transcripts (4), and lineage-specific gene regulation (23). More broadly, while gene expression differences between primates have been shown to often be associated with inter-species differences in histone modifications (96, 97), the chromatin landscape is generally highly conserved in primates. For example, as many as 84% of genomic regions enriched for the active histone modification, H3K27ac, are similar between humans and chimpanzees in iPSC-derived neural crest cells (87). Our results provide further support for the notion that chromatin state is generally highly conserved between closely related primate species. We have not found evidence that TEs, whether orthologous or not, have substantially rewired pluripotent stem cell gene regulatory networks between humans and chimpanzees.

## Materials and methods

### Key Resources Table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
cell line ( <i>H.sapiens</i> , Female)	H20682 iPSC	This paper		Age 23, Caucasian, fibroblast origin
cell line ( <i>H.sapiens</i> , Male)	H20961 iPSC	10.1371/journal.pgen.1005793		
cell line ( <i>H.sapiens</i> , Female)	H21194 iPSC	10.1371/journal.pgen.1005793		
cell line ( <i>H.sapiens</i> , Female)	H21792 iPSC	This paper		Age 20, Caucasian, fibroblast origin
cell line ( <i>H.sapiens</i> , Male)	H28126 iPSC	10.1371/journal.pgen.1005793		
cell line ( <i>H.sapiens</i> , Male)	H28815 iPSC	This paper		Age 19, Caucasian, fibroblast origin
cell line ( <i>H.sapiens</i> , Female)	H18489 iPSC	10.1101/gr.224436.117		
cell line ( <i>H.sapiens</i> , Female)	H18511 iPSC	10.1101/gr.224436.117		
cell line ( <i>H.sapiens</i> , Male)	H19098 iPSC	10.1101/gr.224436.117		
cell line ( <i>H.sapiens</i> , Male)	H19101 iPSC	10.1101/gr.224436.117		
cell line ( <i>P.troglodytes</i> , Female)	C3647 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Male)	C3649 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Female)	C3651 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Female)	C40210 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Female)	C40280 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Male)	C4955 iPSC	10.7554/eLife.07103		
cell line ( <i>P.troglodytes</i> , Male)	C8861 iPSC	10.7554/eLife.07103		
antibody	anti-Histone H3K9me3 (rabbit polyclonal)	Abcam	abcam:8898	5 ug

### Samples

We used seven biological replicates (individuals) for chimpanzee and ten biological replicates for human (encompassing two different populations). This number of replicates has been shown to be sufficient to capture intra-species variation allowing for robust identification of inter-species differences in gene expression (65, 96, 97). Each biological replicate (i.e. different individual within a species) was assayed once. Technical replication was not performed. Experimental processing was designed such that technical variables (such as RNA extraction batch) were not confounded with the variable of interest (species).



## **iPSC culture**

All chimpanzee and Caucasian (CAU) iPSCs were reprogrammed from fibroblasts with episomal vectors (65, 67), while the Yoruba (YRI) iPSCs were similarly reprogrammed from lymphoblastoid cell lines (LCLs) (66). It has been shown, however, that cell-type of origin does not affect iPSC DNA methylation or gene expression patterns (67). After reprogramming, iPSC colonies were cultured on a Mouse Embryonic Fibroblast feeder layer for 12-15 passages prior to conversion to feeder-free growth for 6-26 passages. Three new iPSC lines (H20682, H21792, H28815) were generated as previously described for H20961 (referred to as Ind1 F-iPSC in Burrows *et al.*), H28126 (Ind3 F-iPSC) and H21194 (Ind4 F-iPSC) (67).

Human and chimpanzee feeder-independent stem cell lines were maintained at 70% confluence on Matrigel hESC-qualified Matrix (354277, Corning, Bedford, MA, USA) at a 1:00 dilution. Cells were cultured in Essential 8 Medium (A1517001, ThermoFisher Scientific, Waltham, MA, USA) at 37°C with 5% (vol/vol) CO<sub>2</sub> with daily media changes. Cells were passaged by enzyme-free dissociation (0.5 mM EDTA, 300 mM NaCl in PBS), and seeded with ROCK inhibitor Y-27632 (ab120129, Abcam, Cambridge, MA, USA). All cell lines tested negative for mycoplasma contamination.

## **iPSC quality control**

All iPSCs were assessed for the expression of pluripotency markers, differentiation ability, and genomic instability: these cells express pluripotency factors, can spontaneously differentiate into all

three germ layers following embryoid body formation, and display normal karyotypes. Gene expression from all samples was measured using the HumanHT12 Illumina Gene Expression Array, and data analyzed using the PluriTest bioinformatic assay to determine whether their global gene expression signature matches that of known pluripotent cells (98). We also assayed for the presence of episomal reprogramming vector sequence by qPCR. All three new iPSC lines (H20682, H21792, H28815), and previously described iPSC lines passed these quality control metrics (65-67). We identified one human (H28815) that tested positive for episomal reprogramming vector sequence. However, because this sample was not an obvious outlier in our data we chose to include it in our study.

### **Immunocytochemistry**

iPSCs were fixed in 4% paraformaldehyde in PBS, permeabilized in PBS-T (0.25% Triton X-100 in PBS), and blocked for 30 min in 2.5% BSA in PBS-T. Cells were incubated with primary antibodies in 2.5% BSA in PBS-T overnight at a 1:100 dilution: SSEA-4 mouse IgG (sc-21704, Santa Cruz Biotechnology, Santa Cruz, CA, USA), and Oct-3/4 rabbit IgG (sc-9081). Secondary antibodies were diluted to 1:500 in 2.5% BSA-PBS-T: donkey anti-mouse IgG, Alexa 488 (A21202, ThermoFisher Scientific), donkey anti-rabbit IgG Alexa 594 (A21207) and incubated for 1 hour at room temperature. Nuclei were counter-stained with 1 µg/ml Hoechst 33342 (H3570, ThermoFisher Scientific).

### **ChIP-seq**

ChIP-seq experiments were largely performed according to a published protocol (99). Briefly, 30 million cells were cross-linked with 1% formaldehyde for 10 min at room temperature followed by

quenching with 2.5 M glycine. Following cell lysis and sonication (Covaris S2: 4 min, duty cycle 10%, 5 intensity, 200 cycles per burst in 4x 6 × 16 mm tubes per individual), lysates were incubated with 5 µg H3K9me3 antibody (8898, Abcam) overnight. ChIP and 50 ng input DNA from each individual was end-repaired, A-tailed and ligated to paired-end Illumina TruSeq ChIP Sample Preparation kit sequencing adapters before 18 cycles of PCR amplification. 200–300 bp DNA fragments were selected for sequencing. ChIP enrichment at known target genes (*ZNF333* and *ZNF554*) was quantified and confirmed by qPCR. ChIP-seq experiments were performed in two species-balanced batches. Input and ChIP libraries were multiplexed separately and 50 base pairs sequenced paired-end on the HiSeq2500 rapid run mode according to the manufacturer's instructions.

Sequencing data quality was verified using FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Paired-end ChIP-seq reads from each species were mapped to their respective genome (hg19 or panTro3) using BWA (version 0.7.12) (100) and a mapping quality filter of MAQ >10. PCR duplicates were removed by samtools (version 0.1.19)(101). Properly-paired reads only were selected for further analysis.

ChIP and Input reads from each individual were used to call peaks in each individual using MACS2 (102) with the *broadpeak* option at various *qvalue* cut-offs. A lenient threshold of *qvalue*=0.1 was used in subsequent analysis to maximize the number of regions identified in each species.

To determine the minimum number of sequencing reads required to saturate the number of identified peaks, a read sub-sampling analysis was performed in two individuals from each species. This analysis revealed that the number of peaks called approaches saturation at a median read depth of 6.9 million paired ChIP reads across the four individuals tested (Figure 1-Figure

Supplement 7A). Sub-sampling Input reads, while maintaining the number of ChIP reads constant, saturates the number of peaks called at a median of 5.7 million paired Input reads across the four samples (Figure 1-Figure Supplement 7B). The read depth of the vast majority of samples (31/34) fall within this range.

Peak co-ordinates in each individual in each species were then converted to the alternative genome using liftOver (103) and the best reciprocal chain (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro3/reciprocalBest/>) between hg19 and panTro3 requiring 70% sequence match. Peak regions in each individual that could be reciprocally and uniquely mapped between the two genomes were kept to generate a list of orthologous peak regions. Orthologous peak regions in each individual were concatenated to generate a final orthologous H3K9me3 region set across both species. 50-mer mappability for every base in the hg19 and panTro3 genome was determined based on the GEM algorithm (104). Mappability of all orthologous regions was determined, and only those regions with an average mapping quality score >0.8 in each species were retained.

The number of H3K9me3 ChIP-seq and Input reads falling into these orthologous peak regions in each individual from each species was determined using HTSeq (105). Filtered, autosomal H3K9me3 ChIP-seq and Input read counts in each orthologous region (peak) from each individual in each species were compared using spearman correlation analysis. Having used Input samples to identify regions of H3K9me3 enrichment and identify ChIP-seq samples with low signal, only H3K9me3 ChIP-seq samples were taken forward in the analysis. As sample H20961 was an outlier it was removed from further analysis. H3K9me3 counts were standardized and log transformed to generate log<sub>2</sub>cpm values for PCA analysis.

### Identifying differentially enriched H3K9me3-enriched regions

Orthologous H3K9me3 regions with 0 H3K9me3 counts in more than half of the individuals (8) were removed. The R/Bioconductor package DESeq2 (106) was used to identify which of the orthologous H3K9me3 regions have differential enrichment in H3K9me3 read counts between humans and chimpanzees. Each individual within a species was considered to be a replicate sample. Regions with low mean read counts were filtered prior to multiple-testing correction. We controlled significance at  $FDR < 0.01$  by the Benjamini-Hochberg method. Regions with a  $\log_2$  fold change  $>0$  and an adjusted p-value of  $<0.01$  were considered to be 'Human-enriched', and those with  $\log_2$  fold change  $<0$  and an adjusted p-value of  $<0.01$  as 'Chimpanzee-enriched'. Those regions not significantly differentially enriched are defined as 'Shared'.

Differential H3K9me3 enrichment across species was also assessed by transforming read counts by voom and identifying differentially enriched regions using limma (107). Significance was controlled at  $FDR < 0.01$ .

The proportion of regions differentially enriched at a range of  $\log_2$  fold changes was calculated both using DESeq2 output, and using an Empirical Bayes approach with adaptive shrinkage (ashr)(70).

### Orthologous TE generation

The RepeatMasker track (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*.1996-2010 <http://www.repeatmasker.org>)(1) from the hg19 genome assembly (103) was converted to panTro3

genome coordinates using liftOver reciprocal chain files, requiring a 70% sequence match. These panTro3 coordinates were intersected with the panTro3 RepeatMasker track, requiring that 50% of base pairs overlap, and that the hg19-derived TE name matches that of the panTro3 RepeatMasker annotation. This set was subsequently lifted back over to hg19 again requiring a 70% match. The same procedure was followed starting with the panTro3 RepeatMasker track. Once RepeatMasker tracks were confidently obtained on the same genome, they were intersected (requiring 50% overlap). The distribution of orthologous TEs within the five main TE classes (LINE, SINE, LTR, DNA and SVA) is similar to the overall distribution of TE classes in each species (namely, when we do not require TEs to be orthologous; (Figure 2-Figure Supplement 1)).

### **Identifying silenced TEs**

To obtain a high confidence set of TEs that are silenced, TEs that overlap H3K9me3-enriched regions by at least 50% of their length are considered to be 'Overlapping'. We acknowledge that our stringent H3K9me3 region and orthologous TE definitions likely mean that young, active TEs will be excluded from the majority of our analyses.

### **Non-orthologous TE analysis**

The non-orthologous TE set for each species was generated after excluding the orthologous TEs from all RepeatMasker annotated TEs in that species. The list of species-specific TEs was generated by selecting non-orthologous TEs, where the TE name is not present in the RepeatMasker track of the other species. When determining the overlap between H3K9me3 regions and TEs with differing orthology status (orthologous, non-orthologous, and species-

797 specific), we considered all H3K9me3 regions identified in each species independently (regardless  
798 of orthology).

799

#### 800 **TE sequence divergence from the consensus TE sequence**

801

802 The number of substitutions between the sequence of each TE instance and the consensus TE  
803 sequence (milliDiv score), was obtained from the RepeatMasker track in human.

804

#### 805 **Orthologous TSS annotation**

806

807 A file containing annotated human transcription start sites (TSS) from hg19 was downloaded from  
808 the Table Browser of the UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgTables>) by  
809 filtering for the 'txStart' from Ensembl genes (108). A single TSS was assigned per gene using the  
810 TSS of the 5' most transcript from genes oriented on the sense strand, and the 3' most transcript  
811 from genes on the anti-sense strand. Only TSSs corresponding to 30,030 genes contained in a  
812 primate orthologous exon file were retained (75).

813

#### 814 **Chromatin states of orthologous TEs**

815

816 Orthologous TEs categorized by H3K9me3 silencing category (Shared, Human-enriched,  
817 Chimpanzee-enriched and Non-overlapping) were intersected with hg19 chromatin states defined  
818 by ChromHMM (74) from data produced and analyzed by the ENCODE Consortium (109) and  
819 obtained from the UCSC Genome Browser hg19 database  
820 (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/>) (110). Chromatin states from H1  
821 hESCs (wgEncodeBroadHmmH1heschHMM.bed), HepG2 (wgEncodeBroadHmmHepg2HMM.txt)

822 and GM12878 (wgEncodeBroadHmmGm12878HMM.txt) were used. An intersection is defined as  
823 a 1 bp overlap.

824

825 **Associating orthologous TEs with functional genomics annotations**

826

827 Orthologous TEs categorized by H3K9me3 silencing category (Shared, Human-enriched,  
828 Chimpanzee-enriched and Non-overlapping) were intersected with ChIP-seq and DNase I  
829 hypersensitivity data obtained from the H1 hESC line  
830 (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/>) as described above. An intersection  
831 is called when 50% of the length of the TE overlaps with the genomic feature.

832

833 The following ENCODE datasets from the UCSC Genome Browser hg19 database were used:

834 wgEncodeBroadHistoneH1hescH3k4me1StdPk.txt

835 wgEncodeBroadHistoneH1hescH3k4me3StdPk.txt

836 wgEncodeBroadHistoneH1hescH3k27acStdPk.txt

837 wgEncodeBroadHistoneH1hescH3k27me3StdPk.txt

838 wgEncodeAwgDnaseUwdukeH1hescUniPk.txt

839 wgEncodeAwgTfbsHaibH1hescP300V0416102UniPk.txt

840 wgEncodeAwgTfbsHaibH1hescPol2V0416102UniPk.txt

841 wgEncodeAwgTfbsHaibH1hescCtcfsc5916V0416102UniPk.txt

842 wgEncodeAwgTfbsHaibH1hescYy1sc281V0416102UniPk.txt

843 wgEncodeAwgTfbsHaibH1hescNanogsc33759V0416102UniPk.txt

844 wgEncodeAwgTfbsHaibH1hescPou5f1sc9081V0416102UniPk.txt

845



TEs have been found to mediate the differential wiring of pluripotency networks in humans and mice by carrying species-specific transcription factor binding sites for the core pluripotency factors Nanog and Oct4 (29). To determine whether inter-species TE silencing differences at these TE-associated transcription factor binding sites could mediate inter-species differences in the regulation of pluripotency, we overlaid our orthologous TEs with binding locations for these transcription factors and two ubiquitously-expressed factors (CTCF and YY1). We found less than 0.5% of TEs overlapping these transcription factor binding locations (<100 instances per category) making it difficult to make robust conclusions (Figure 5-Figure Supplement 6). However, the small number of instances indicates that differences in H3K9me3-mediated TE silencing do not contribute to rewiring pluripotency networks in these closely related species.

#### **RNA-seq**

RNA was extracted from ~3 million cells using the ZR-Duet DNA/RNA extraction kit (D7001, Zymo, Irvine, CA, USA) in two species-balanced batches. All RNA samples were of good quality: RIN scores of 10 were obtained for all samples except C40280 (9.7). All 17 samples were pooled together for RNA-seq library generation using the Illumina Truseq kit. The library pool was sequenced 50 base pairs, paired-end on the Hiseq2500.

Paired-end RNA-seq reads from each species were trimmed and mapped to each genome using Tophat2 (version 2.0.13) (111). Only properly-paired reads were taken for further analysis. The number of sequencing reads is similar across individuals from both species (median human: 42,557,599, median chimpanzee: 33,987,431)(Figure 6-Figure Supplement 1 and and Table 3 in Supplementary File 1). The number of reads falling into orthologous meta-exons across 30,030 Ensembl genes from hg19, panTro3 and rheMac3 (75) was determined using featureCounts within

subread (version 1.5.0) (112). Autosomal genes with >0 counts in >10/17 individuals were retained. Counts were quantile normalized across all samples, standardized, and log-transformed to generate gene expression levels expressed as normalized  $\log_2\text{cpm}$ .

To identify genes differentially expressed between the two species, counts were transformed by voom (107) prior to differential expression analysis using limma (113). We controlled significance at  $\text{FDR} < 0.01$ .

### **Association between H3K9me3-mediated orthologous TE silencing and neighboring gene expression**

We selected the annotated TSSs described above that correspond to orthologous genes for which we have gene expression data. We used the genomic coordinates of these annotations to define 1, 10, 20 and 40 kb windows upstream of the TSS. We then identified all TEs contained within each window, retaining the previous annotation of whether the TE overlaps an H3K9me3 region or not. Those TEs overlapping an orthologous H3K9me3 region had additional information in the form of H3K9me3 counts in each individual of each species. H3K9me3 counts in each orthologous H3K9me3 region overlapping a TE were summed within a window to yield a final H3K9me3 TE silencing count per gene per species. If multiple TEs fell within the same orthologous H3K9me3 region, this region was only counted once. H3K9me3 count values overlapping all TEs within a window was used as input for DESeq2 to determine inter-species differences in TE silencing per gene as previously described.

To determine gene expression divergence of genes overlapping H3K9me3 regions at the TSS, regardless of the presence of a TE, a 1 kb region around the TSS was created. All 141,642

H3K9me3 orthologous regions (irrespective of the whether the regions overlaps a TE or not) were overlapped with the defined TSS corresponding to orthologous genes. Genes where at least 50% of the length of the 1 kb TSS region overlap with an H3K9me3 region are considered to be overlapping the TSS.

### **Effect of TEs within human H3K4me1 and H3K4me3 regions on inter-species gene expression divergence**

To identify TEs with potential to act as regulatory regions important in mediating inter-species gene expression differences, a subset of orthologous TEs that also overlap a human H3K4me1/H3K4me3 region in the H1 hESC line were used. These TEs were assigned their closest gene based on the location of the TSS, and only unique orthologous genes were retained. Inter-species gene expression divergence was determined for these genes and compared to the divergence of unique orthologous genes closest to all H3K4me1/H3K4me3 regions in H1 hESCs.

### **TE expression**

The expression of TE types was determined as previously described (50). Briefly, multi-mapping, paired-end RNA-seq reads were retained for both humans and chimpanzees. RepeatMasker coordinates for all annotated instances of each TE type were obtained in each species. RPKM values for each TE type, in each species, were calculated by normalizing agglomerated read coverage of all annotated genomic copies of that TE type in the reference genome, to the agglomerated length of the TE type, as well as the total number of exonic reads. TE types were considered to be expressed if they had RPKM values >1.

## **Association between H3K9me3-mediated non-orthologous TE silencing and neighboring gene expression**

In each species independently, we used all previously determined non-orthologous TEs annotated as overlapping or not overlapping any H3K9me3 region identified in that species. Non-orthologous TEs were overlapped with the 1, 10 and 20 kb windows identified upstream of each orthologous gene's TSS as described above. To identify TSS positions in the chimpanzee genome, the 1 bp TSS annotations of orthologous genes determined in human were lifted over to the chimpanzee genome as previously described in the generation of orthologous TEs. Similarly, a 1 kb region around each human TSS was lifted over to the chimpanzee genome, and only orthologous single bp TSSs where the broader 1 kb region lifts over to the chimpanzee genome were retained.

Each gene was categorized as 'silent' if all non-orthologous TEs in the window upstream of the TSS overlap H3K9me3, or 'non-silent' if at least one non-orthologous TE in the upstream window did not overlap with a H3K9me3 region. Results are consistent if we further stratify categories of genes where upstream TEs are fully silenced (all TEs overlap H3K9me3), or non-silenced (no TEs overlap H3K9me3).

## **KRAB-ZNF gene expression and overlap with H3K9me3 regions**

To identify those KRAB-ZNF genes marked by H3K9me3, a curated list of 346 genes (54) was filtered to include only those contained within our set of orthologous genes. These 256 genes were overlapped with orthologous H3K9me3 regions. KRAB-ZNF genes where 50% of their length overlaps with H3K9me3 are considered to be overlapping.

KRAB-ZNF genes where 50% of the length of a 1 kb region around the TSS overlaps with an orthologous H3K9me3 region were considered to be overlapping H3K9me3 at the TSS, and used in downstream gene expression divergence analysis.

## **Supporting Information**

**Supplementary File 1:** Supplemental Tables 1-11.

**Supplementary File 2:** Table with 141,642 orthologous H3K9me3 regions, their ChIP-seq read counts, DESeq2 results (baseMean, log<sub>2</sub>FC, lfcSE, stat, pval, padj), and classifications as Shared, Human-enriched and Chimpanzee-enriched.

**Supplementary File 3:** Table with 4,242,188 orthologous TEs, their overlap with orthologous H3K9me3 regions, and classification as Shared, Human-enriched and Chimpanzee-enriched.

## **Data Access**

All data have been deposited in the Gene Expression Omnibus ([www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)) under accession number GSE96712.

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=sloxgksifvinnwr&acc=GSE96710>

## **Acknowledgements**

We thank Matthew Lorincz as well as all members of the Gilad lab, especially Sebastian Pott, for helpful and insightful discussions, and Amy Mitrano for performing RNA extractions and preparing sequencing libraries. We thank the ENCODE Consortium for making their data freely available to use. M.C.W is supported by an EMBO Long-Term Fellowship (ALTF 751-2014) and the European Commission Marie Curie Actions.

## Competing Interests

The authors declare no financial and non-financial competing interests.

## References

1. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* 2000;16(9):418-20.
2. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 2011;7(12):e1002384.
3. Chimpanzee S, Analysis C. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature.* 2005;437(7055):69-87.
4. Ramsay L, Marchetto MC, Caron M, Chen SH, Busche S, Kwan T, et al. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. *BMC Genomics.* 2017;18(1):214.
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.
6. Tonjes RR, Lower R, Boller K, Denner J, Hasenmaier B, Kirsch H, et al. HERV-K: the biologically most active human endogenous retrovirus family. *J Acquir Immune Defic Syndr Hum Retrovirol.* 1996;13 Suppl 1:S261-7.
7. Medstrand P, Mager DL. Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol.* 1998;72(12):9782-7.
8. Fuchs NV, Loewer S, Daley GQ, Izsvak Z, Lower J, Lower R. Human endogenous retrovirus K (HML-2) RNA and protein expression is a marker for human embryonic and induced pluripotent stem cells. *Retrovirology.* 2013;10:115.
9. Klawitter S, Fuchs NV, Upton KR, Munoz-Lopez M, Shukla R, Wang J, et al. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat Commun.* 2016;7:10286.

- 1003 10. Wildschutte JH, Williams ZH, Montesion M, Subramanian RP, Kidd JM, Coffin JM.  
1004 Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc*  
1005 *Natl Acad Sci U S A*. 2016;113(16):E2326-34.
- 1006 11. Batzer MA, Deininger PL. A human-specific subfamily of Alu sequences. *Genomics*.  
1007 1991;9(3):481-7.
- 1008 12. Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ, Deininger PL. Amplification  
1009 dynamics of human-specific (HS) Alu family members. *Nucleic Acids Res*. 1991;19(13):3619-  
1010 23.
- 1011 13. Kazazian HH, Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE.  
1012 Haemophilia A resulting from de novo insertion of L1 sequences represents a novel  
1013 mechanism for mutation in man. *Nature*. 1988;332(6160):164-6.
- 1014 14. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. Hot L1s  
1015 account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*.  
1016 2003;100(9):5280-5.
- 1017 15. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, et al. SVA elements: a hominid-  
1018 specific retroposon family. *J Mol Biol*. 2005;354(4):994-1007.
- 1019 16. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH, Jr. SVA elements are nonautonomous  
1020 retrotransposons that cause disease in humans. *Am J Hum Genet*. 2003;73(6):1444-51.
- 1021 17. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, et al. Lineage-  
1022 specific expansions of retroviral insertions within the genomes of African great apes but not  
1023 humans and orangutans. *PLoS Biol*. 2005;3(4):e110.
- 1024 18. Mun S, Lee J, Kim YJ, Kim HS, Han K. Chimpanzee-specific endogenous retrovirus  
1025 generates genomic variations in the chimpanzee genome. *PLoS One*. 2014;9(7):e101195.
- 1026 19. Marchetto MC, Narvaiza I, Denli AM, Benner C, Lazzarini TA, Nathanson JL, et al.  
1027 Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature*.  
1028 2013;503(7477):525-9.
- 1029 20. Kazazian HH, Jr. Mobile elements: drivers of genome evolution. *Science*.  
1030 2004;303(5664):1626-32.
- 1031 21. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution.  
1032 *Nat Rev Genet*. 2009;10(10):691-703.
- 1033 22. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory  
1034 sequences are derived from transposable elements. *PLoS Genet*. 2013;9(5):e1003504.
- 1035 23. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, et al.  
1036 Transposable elements are the primary source of novelty in primate gene regulation. *Genome*  
1037 *Res*. 2017.
- 1038 24. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene  
1039 regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet*.  
1040 2011;43(11):1154-9.
- 1041 25. Chuong EB, Rumi MA, Soares MJ, Baker JC. Endogenous retroviruses function as  
1042 species-specific enhancer elements in the placenta. *Nat Genet*. 2013;45(3):325-9.
- 1043 26. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within  
1044 specific transposable element families associates with tissue-specific enhancer landscape. *Nat*  
1045 *Genet*. 2013;45(7):836-41.

1046 27. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, et al. Species-specific  
1047 endogenous retroviruses shape the transcriptional network of the human tumor suppressor  
1048 protein p53. *Proc Natl Acad Sci U S A*. 2007;104(47):18613-8.

1049 28. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the  
1050 mammalian transcription factor binding repertoire via transposable elements. *Genome Res*.  
1051 2008;18(11):1752-62.

1052 29. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements  
1053 have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*.  
1054 2010;42(7):631-4.

1055 30. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of  
1056 retrotransposon expansion remodel genome organization and CTCF binding in multiple  
1057 mammalian lineages. *Cell*. 2012;148(1-2):335-48.

1058 31. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated  
1059 retrotransposon transcriptome of mammalian cells. *Nat Genet*. 2009;41(5):563-71.

1060 32. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long  
1061 noncoding RNAs. *Genome Biol*. 2012;13(11):R107.

1062 33. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable  
1063 elements are major contributors to the origin, diversification, and regulation of vertebrate  
1064 long noncoding RNAs. *PLoS Genet*. 2013;9(4):e1003470.

1065 34. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-  
1066 option of endogenous retroviruses. *Science*. 2016;351(6277):1083-7.

1067 35. Lynch VJ, Nnamani MC, Kapusta A, Brayer K, Plaza SL, Mazur EC, et al. Ancient  
1068 transposable elements transformed the uterine regulatory landscape and transcriptome  
1069 during the evolution of mammalian pregnancy. *Cell Rep*. 2015;10(4):551-61.

1070 36. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al.  
1071 Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*.  
1072 2012;487(7405):57-63.

1073 37. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, et al.  
1074 Species-specific transcription in mice carrying human chromosome 21. *Science*.  
1075 2008;322(5900):434-8.

1076 38. Ward MC, Wilson MD, Barbosa-Morais NL, Schmidt D, Stark R, Pan Q, et al. Latent  
1077 regulatory potential of human-specific repetitive elements. *Mol Cell*. 2013;49(2):262-72.

1078 39. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, et al. Proviral  
1079 silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature*.  
1080 2010;464(7290):927-31.

1081 40. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1  
1082 controls endogenous retroviruses in embryonic stem cells. *Nature*. 2010;463(7278):237-40.

1083 41. Najafabadi HS, Mnaimneh S, Schmitges FW, Garton M, Lam KN, Yang A, et al. C2H2 zinc  
1084 finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol*. 2015;33(5):555-  
1085 62.

1086 42. Turelli P, Castro-Diaz N, Marzetta F, Kapopoulou A, Raclot C, Duc J, et al. Interplay of  
1087 TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res*.  
1088 2014;24(8):1260-70.



1089 43. Friedli M, Turelli P, Kapopoulou A, Rauwel B, Castro-Diaz N, Rowe HM, et al. Loss of  
1090 transcriptional control over endogenous retroelements during reprogramming to  
1091 pluripotency. *Genome Res.* 2014;24(8):1251-9.

1092 44. Wolf D, Goff SP. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature.*  
1093 2009;458(7242):1201-4.

1094 45. Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ, 3rd. SETDB1: a novel KAP-  
1095 1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated  
1096 silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* 2002;16(8):919-32.

1097 46. Rowe HM, Trono D. Dynamic control of endogenous retroviruses during development.  
1098 *Virology.* 2011;411(2):273-87.

1099 47. Quenneville S, Turelli P, Bojkowska K, Raclot C, Offner S, Kapopoulou A, et al. The  
1100 KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific  
1101 DNA methylation patterns maintained during development. *Cell Rep.* 2012;2(4):766-73.

1102 48. Rowe HM, Friedli M, Offner S, Verp S, Mesnard D, Marquis J, et al. De novo DNA  
1103 methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET.  
1104 *Development.* 2013;140(3):519-29.

1105 49. Smith ZD, Chan MM, Humm KC, Karnik R, Mekhoubad S, Regev A, et al. DNA  
1106 methylation dynamics of the human preimplantation embryo. *Nature.* 2014;511(7511):611-5.

1107 50. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, et al. DNA methylation  
1108 and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and  
1109 chimeric transcripts in mESCs. *Cell Stem Cell.* 2011;8(6):676-87.

1110 51. Walter M, Teissandier A, Perez-Palacios R, Bourc'his D. An epigenetic switch ensures  
1111 transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife.*  
1112 2016;5.

1113 52. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male  
1114 germ cells lacking Dnmt3L. *Nature.* 2004;431(7004):96-9.

1115 53. Barau J, Teissandier A, Zamudio N, Roy S, Nalesso V, Herault Y, et al. The DNA  
1116 methyltransferase DNMT3C protects male germ cells from transposon activity. *Science.*  
1117 2016;354(6314):909-12.

1118 54. Kapopoulou A, Mathew L, Wong A, Trono D, Jensen JD. The evolution of gene  
1119 expression and binding specificity of the largest transcription factor family in primates.  
1120 *Evolution.* 2016;70(1):167-80.

1121 55. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, Kim J, et al. A  
1122 comprehensive catalog of human KRAB-associated zinc finger genes: insights into the  
1123 evolutionary history of a large family of transcriptional repressors. *Genome Res.*  
1124 2006;16(5):669-77.

1125 56. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes.  
1126 *Genome Res.* 2011;21(11):1800-12.

1127 57. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from  
1128 conflicts to benefits. *Nat Rev Genet.* 2017;18(2):71-86.

1129 58. Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea GO, Munoz-Lopez M, et  
1130 al. Heritable L1 retrotransposition in the mouse primordial germline and early embryo.  
1131 *Genome Res.* 2017;27(8):1395-405.

1132 59. Santoni FA, Guerra J, Luban J. HERV-H RNA is abundant in human embryonic stem cells  
1133 and a precise marker for pluripotency. *Retrovirology.* 2012;9:111.

1134 60. Fort A, Hashimoto K, Yamada D, Salimullah M, Keya CA, Saxena A, et al. Deep  
1135 transcriptome profiling of mammalian stem cells supports a regulatory role for  
1136 retrotransposons in pluripotency maintenance. *Nat Genet.* 2014;46(6):558-66.

1137 61. Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, et al. The retrovirus HERVH is a  
1138 long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol.*  
1139 2014;21(4):423-5.

1140 62. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, et al. Primate-specific  
1141 endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature.*  
1142 2014;516(7531):405-9.

1143 63. Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, et al. Intrinsic  
1144 retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature.*  
1145 2015;522(7555):221-5.

1146 64. Wissing S, Munoz-Lopez M, Macia A, Yang Z, Montano M, Collins W, et al.  
1147 Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. *Hum Mol*  
1148 *Genet.* 2012;21(1):208-18.

1149 65. Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, et  
1150 al. A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative  
1151 functional genomics. *Elife.* 2015;4:e07103.

1152 66. Banovich NE, Li YI, Raj A, Ward MC, Greenside P, Calderon D, et al. Impact of regulatory  
1153 variation across human iPSCs and differenti- ated cells. *bioRxiv.* 2016.

1154 67. Burrows CK, Banovich NE, Pavlovic BJ, Patterson K, Gallego Romero I, Pritchard JK, et  
1155 al. Genetic Variation, Not Cell Type of Origin, Underlies the Majority of Identifiable Regulatory  
1156 Differences in iPSCs. *PLoS Genet.* 2016;12(1):e1005793.

1157 68. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al.  
1158 Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-30.

1159 69. Jacobs FM, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An  
1160 evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1  
1161 retrotransposons. *Nature.* 2014;516(7530):242-5.

1162 70. Stephens M. False discovery rates: a new deal. *Biostatistics.* 2017;18(2):275-94.

1163 71. Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, et al.  
1164 Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev.* 2014;28(13):1397-  
1165 409.

1166 72. van der Kuyl AC. HIV infection and HERV expression: a review. *Retrovirology.*  
1167 2012;9:6.

1168 73. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the  
1169 evolution of gene regulatory networks. *Nature.* 2017.

1170 74. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic  
1171 annotation of the human genome. *Nat Biotechnol.* 2010;28(8):817-25.

1172 75. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y. Sex-specific and lineage-specific  
1173 alternative splicing in primates. *Genome Res.* 2010;20(2):180-9.

1174 76. Bilodeau S, Kagey MH, Frampton GM, Rahl PB, Young RA. SetDB1 contributes to  
1175 repression of genes encoding developmental regulators and maintenance of ES cell state.  
1176 *Genes Dev.* 2009;23(21):2484-9.

1177 77. Mozzetta C, Boyarchuk E, Pontis J, Ait-Si-Ali S. Sound of silence: the properties and  
1178 functions of repressive Lys methyltransferases. *Nat Rev Mol Cell Biol.* 2015;16(8):499-513.

1179 78. O'Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, Chang HY, et al. Genome-wide  
1180 analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* 2007;3(6):e89.  
1181 79. Davidson EH, Britten RJ. Regulation of gene expression: possible role of repetitive  
1182 sequences. *Science.* 1979;204(4397):1052-9.  
1183 80. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human  
1184 regulatory sequences from transposable elements. *Trends Genet.* 2003;19(2):68-72.  
1185 81. Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev*  
1186 *Genet.* 2008;9(5):397-405.  
1187 82. Goke J, Ng HH. CTRL+INSERT: retrotransposons and their contribution to regulation  
1188 and innovation of the transcriptome. *EMBO Rep.* 2016;17(8):1131-44.  
1189 83. Schlesinger S, Goff SP. Retroviral transcriptional regulation and embryonic stem cells:  
1190 war and peace. *Mol Cell Biol.* 2015;35(5):770-7.  
1191 84. Robbez-Masson L, Rowe HM. Retrotransposons shape species-specific embryonic stem  
1192 cell gene expression. *Retrovirology.* 2015;12:45.  
1193 85. Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, et al.  
1194 Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res.*  
1195 2016;26(12):1742-52.  
1196 86. Simonti CN, Pavlicev M, J.A. C. Corinne N. Simonti, Mihaela Pavličev, John A. Capra;  
1197 Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary  
1198 Age, and Subject to Pleiotropic Constraints. *Molecular Biology and Evolution.* 2017;msx219.  
1199 87. Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, et al. Enhancer  
1200 divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell.*  
1201 2015;163(1):68-83.  
1202 88. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the  
1203 genome. *Nat Rev Genet.* 2007;8(4):272-85.  
1204 89. Huda A, Marino-Ramirez L, Jordan IK. Epigenetic histone modifications of human  
1205 transposable elements: genome defense versus exaptation. *Mob DNA.* 2010;1(1):2.  
1206 90. Day DS, Luquette LJ, Park PJ, Kharchenko PV. Estimating enrichment of repetitive  
1207 elements from high-throughput sequence data. *Genome Biol.* 2010;11(6):R69.  
1208 91. Leeb M, Pasini D, Novatchkova M, Jaritz M, Helin K, Wutz A. Polycomb complexes act  
1209 redundantly to repress genomic repeats and genes. *Genes Dev.* 2010;24(3):265-76.  
1210 92. Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, et al. Transposable Elements  
1211 and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Dev Cell.*  
1212 2016;36(6):611-23.  
1213 93. Brattas PL, Jonsson ME, Fasching L, Nelander Wahlestedt J, Shahsavani M, Falk R, et al.  
1214 TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human  
1215 Neural Progenitor Cells. *Cell Rep.* 2017;18(1):1-11.  
1216 94. Sundaram V, Choudhary MN, Pehrsson E, Xing X, Fiore C, Pandey M, et al. Functional  
1217 cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nat Commun.*  
1218 2017;8:14550.  
1219 95. Schwalie PC, Ward MC, Cain CE, Faure AJ, Gilad Y, Odom DT, et al. Co-binding by YY1  
1220 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate  
1221 genomes. *Genome Biol.* 2013;14(12):R148.

96. Cain CE, Blekhman R, Marioni JC, Gilad Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics*. 2011;187(4):1225-34.
97. Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol*. 2014;15(12):547.
98. Muller FJ, Schuldt BM, Williams R, Mason D, Altun G, Papapetrou EP, et al. A bioinformatic assay for pluripotency in human cells. *Nat Methods*. 2011;8(4):315-7.
99. Schmidt D, Wilson MD, Spyrou C, Brown GD, Hadfield J, Odom DT. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*. 2009;48(3):240-8.
100. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-60.
101. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-9.
102. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.
103. Speir ML, Zweig AS, Rosenbloom KR, Raney BJ, Paten B, Nejad P, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res*. 2016;44(D1):D717-25.
104. Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, et al. Fast computation and applications of genome mappability. *PLoS One*. 2012;7(1):e30377.
105. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-9.
106. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
107. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014;15(2):R29.
108. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(Database issue):D493-6.
109. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
110. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res*. 2013;41(Database issue):D56-63.
111. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
112. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-30.
113. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.

## Figure Legends

**Figure 1: Majority of H3K9me3 regions are similarly enriched in human and chimpanzee iPSCs. (A)** Experimental design of the study. **(B)** A heuristic description of the computational pipeline to identify regions of differential H3K9me3 enrichment between humans and chimpanzees. Human genome, ChIP-seq peaks and ChIP-seq reads (blue); chimpanzee genome, ChIP-seq peaks and ChIP-seq reads (black), orthologous H3K9me3 regions (purple). **(C)** Volcano plot representing the regions similarly enriched in both species (grey: Shared), enriched in human (blue), or enriched in chimpanzee (brown) at 1% FDR. **(D)** Pie-chart representing the proportion of regions within the Shared, Human-enriched and Chimpanzee-enriched categories. **(E)** Examples of individual loci representing each of the three defined categories. The top track represents sequencing reads from H3K9me3 ChIP-seq experiments, and the bottom track gene expression levels of the *MALRD1*, *CACNA1B* and *FANK1* genes determined by RNA-seq. Also see Figure 1- Figure Supplements 1-14.

**Figure 2: Transposable element silencing is associated with particular transposable element classes. (A)** Bioinformatic pipeline used to identify TEs that are orthologous between humans and chimpanzees using RepeatMasker (rmsk) tracks in each species. **(B)** The proportion of all TEs that overlap an orthologous H3K9me3 region (H3K9me3 peak identified in at least one individual of one species) (purple) compared to those that do not overlap an orthologous H3K9me3 region (magenta). **(C)** Each orthologous TE instance is further hierarchically classified into Class, Family and Type. The proportion of all orthologous TEs belonging to the five main TE classes (SINE: yellow, LINE: orange, DNA: peach, SVA: brown, LTR: red) that do not overlap an H3K9me3 region in either species, and the proportion of TEs that do overlap a H3K9me3 region. The small

number of SVA elements means that this class is not visible in the pie chart. Also see Figure 2-  
Figure Supplement 1.

**Figure 3: Patterns of preferential TE class-based silencing are maintained as orthology decreases.** Within-species analysis of the proportion of orthologous TEs (left panel), TEs that are not orthologous between the two species (middle panel), and TEs that are not orthologous and whose type is annotated in a single species (right panel), that overlap any H3K9me3-enriched region identified in that species. Magenta represents those TEs that do not overlap with any H3K9me3 region identified in that species, blue represents TEs that overlap any H3K9me3 regions in human, and black those TEs that overlap with H3K9me3 regions in chimpanzee. The absence of data in the species-specific category reflects that there are no species-specific annotations in these TE classes. Asterisk denotes when there is a significant difference between orthology categories (Pearson's Chi-squared test;  $P < 0.001$ ). Also see Figure 3-Figure Supplement 1.

**Figure 4: Majority of orthologous TEs may be similarly silenced in humans and chimpanzees. (A)** The proportion of orthologous TEs in each class that do not overlap an orthologous H3K9me3 region (magenta), show similar H3K9me3 enrichment in both species (grey: Shared), are enriched in human (blue), or are enriched in chimpanzee (brown) are shown. Of those orthologous TEs that overlap an orthologous H3K9me3 region, the distribution between those that are Shared, Human-enriched, or Chimpanzee-enriched is shown in the right panel. **(B)** The proportion of orthologous TE elements belonging to each class within each of the four H3K9me3 silencing categories (LINE: orange, SINE: yellow, LTR: red, SVA: brown, DNA: peach). The small number of SVA elements means that this class is not visible in the pie chart. Also see Figure 4-Figure Supplements 1-3.

1315

1316 **Figure 5: Majority of silenced orthologous TEs have similar properties in humans and**  
1317 **chimpanzees. (A)** The length of all orthologous TE instances occurring within Shared, Human-  
1318 enriched and Chimpanzee-enriched H3K9me3 silencing categories. **(B)** The length of all  
1319 orthologous TEs stratified by inter-species divergence in effect size. **(C)** The sequence divergence  
1320 from the consensus TE sequence of TEs categorized as in (A). **(D)** The sequence divergence of  
1321 TEs categorized as in (B). **(E)** The distance between each TE and the closest orthologous TSS  
1322 categorized as in (A). **(F)** The distance between each TE and the closest orthologous TSS  
1323 categorized as in (B). Grey: Shared, blue: Human-enriched, brown: Chimpanzee-enriched,  
1324 magenta: does not overlap an orthologous H3K9me3 region, light teal:  $\log_2$  fold change less than 2  
1325 between species, dark teal:  $\log_2$  fold change greater than 2 between species. **(G)** Orthologous TEs  
1326 categorized by silencing status that overlap chromatin states annotated in the H1 hESC line.  
1327 Asterisk denotes when there is a significant difference between the Shared and Non-overlapping,  
1328 and high and low divergence categories (Wilcoxon-rank sum test;  $P < 0.001$ ). Also see Figure 5-  
1329 Figure Supplements 1-6.

1330

1331

1332 **Figure 6: TE silencing divergence between species is not correlated with gene expression**  
1333 **divergence. (A)** Schematic representation of the experiment to determine how TE silencing affects  
1334 nearby gene expression; orthologous TEs (green rectangles), orthologous H3K9me3 regions  
1335 (purple triangles), human read counts (blue), and chimpanzee read counts (black). **(B)** Human and  
1336 chimpanzee gene expression levels of orthologous genes with at least one TE present in a 10 kb  
1337 window upstream of the TSS. Genes are categorized by the presence or absence of H3K9me3  
1338 regions overlapping TEs. **(C)** Gene expression divergence between human and chimpanzee at  
1339 genes where TE silencing in the upstream 10 kb region is classified as Shared, Human-enriched or

Chimpanzee-enriched. The proportion of genes associated with each TE silencing category is represented by the pie chart. **(D)** Stratification of genes based on differential expression status: genes that are similarly expressed in human and chimpanzee (dark grey), those that are up-regulated in human (turquoise), and those that are up-regulated in chimpanzee (dark brown) (FDR < 1%). The proportion of TEs marked by H3K9me3 in a 10 kb window upstream of the TSS in each gene category that are Shared, enriched in human, and enriched in chimpanzee are represented as pie charts. Box plots represent human-chimpanzee divergence of H3K9me3 enrichment overlapping TEs for each gene category. **(E)** Divergence in H3K9me3-mediated TE silencing between species versus neighboring gene expression divergence. Genes which have a H3K9me3 log fold divergence of >1 or <-1, and a gene expression divergence log fold change of >1 or <-1 are shown in red. **(F)** Inter-species gene expression divergence of orthologous genes closest to TEs overlapping H3K9me3 regions enriched in chimpanzee, and overlapping human H3K4me3/H3K4me1 regions identified in H1 hESCs. The closest orthologous genes to all human H3K4me3/H3K4me1 regions in H1 cells are used for comparison. Asterisk denotes a significant difference between categories (Wilcoxon-rank sum test;  $P < 0.01$ ). Also see Figure 6-Figure Supplements 1-10.

## Figure Supplement Legends

**Figure 1-Figure Supplement 1: Newly generated human fibroblast-derived iPSCs have normal karyotypes and are able to differentiate into each of the three germ layers. (A)** Chromosome integrity analysis by metaphase spread. **(B)** Immunocytochemistry for markers from each of the three germ layers in spontaneously formed embryoid bodies. Evidence for endoderm differentiation is shown by positive staining for AFP and HNF3b, ectoderm by MAP2 staining, and mesoderm by SMA staining. Nuclei are stained with Hoechst. **(C)** PCR products of the episomal



reprogramming vectors used to reprogram the fibroblasts. Pos-1: 1 pg mixture of reprogramming plasmids. Neg-1: 100 ng mixture of human and chimpanzee fibroblast DNA without added vector, Pos-2: 100 ng mixture of human and chimpanzee fibroblast DNA with a 1 pg mixture of reprogramming plasmids. H22815 tested positive for episomal vector but as it was not an outlier in any other analysis we chose to include it in our study.

**Figure 1-Figure Supplement 2: Human fibroblast-derived iPSCs exhibit gene expression profiles consistent with pluripotent cells. (A)** Expression of the core pluripotency genes *OCT3/4*, *NANOG* and *SOX2* relative to *GAPDH* determined by qPCR. Error bars represent standard deviation from three technical replicates **(Bi)** PluriTest-derived pluripotency scores (Muller *et al.*) as a measure of the overall pluripotency gene expression signature of each sample. The empirical range in which samples' global gene expression profiles are consistent with known pluripotent cells is shown in red. Non-iPSCs fall in the score range indicated by the blue lines. **(Bii)** PluriTest-derived novelty score as a measure of contamination from differentiated cells. The empirical threshold (1.67) below which samples' gene expression profiles are consistent with known pluripotent stem cells is shown in red. Fibroblast-derived iPSCs previously described in Burrows *et al.* are shown for comparison.

**Figure 1-Figure Supplement 3: Human and chimpanzee iPSCs express key pluripotency factors.** Immunocytochemistry for nuclear OCT3/4 and cytoplasmic SSEA-4 in a subset of iPSCs from male and female human and chimpanzee individuals. Nuclei are stained with Hoechst. iPSC-derived cardiomyocytes (iPSC-CM) are shown as a negative control.

**Figure 1-Figure Supplement 4: Number of ChIP and Input sequencing reads are similar across species.** Total number of high quality, aligned read pairs for each individual from human and chimpanzee.

**Figure 1-Figure Supplement 5: H3K9me3 ChIP-seq reads are enriched at *ZNF* genes in both species.** H3K9me3 ChIP-seq reads in 10 human iPSCs (blue), the H1 hESC line from ENCODE (purple), and 7 chimpanzees (black) at the *ZNF554* and *ZNF555* genes.

**Figure 1-Figure Supplement 6: The number of ChIP-seq peaks is equivalent in human and chimpanzee.** Number of H3K9me3 ChIP-seq peaks identified in each individual at five different FDR thresholds.

**Figure 1-Figure Supplement 7: ChIP-seq peak numbers approach saturation with increasing read depth. (A)** Number of H3K9me3 ChIP-seq peaks called at a FDR of 10% while sub-sampling the number of ChIP-seq reads from C3651 (red) and H19098 (blue) (~15 million total reads), and C3649 (black) and H20682 (purple) (~9 million total reads). **(B)** Number of H3K9me3 ChIP-seq peaks called at a FDR of 10% while sub-sampling the number of Input reads from C3651 and H19098 (~35 million total reads), and C3649 and H20682 (~16 million total reads).

**Figure 1-Figure Supplement 8: Reciprocal peak mapping retains ~90% of ChIP-seq peaks.** Number of H3K9me3 ChIP-seq peaks obtained in human (H) at a FDR of 10% that map to the chimpanzee genome (H-c), that map back to the human genome (H-c-h) and back to the chimpanzee genome (H-c-h-c). The same analysis was performed from the chimpanzee perspective (C).

**Figure 1-Figure Supplement 9: ChIP-seq data clusters by species. (A)** Spearman correlation of ChIP-seq read counts for H3K9me3 ChIP and Input samples from human (H) and chimpanzee (C) in 160,113 autosomal regions. **(B)** Spearman correlation of ChIP-seq read counts for H3K9me3 ChIP samples only in human and chimpanzee.

**Figure 1-Figure Supplement 10: ChIP-seq data separates by species.** PCA analysis of H3K9me3 ChIP-seq read counts in 160,113 orthologous regions in humans from a Yoruba population (YRI: light blue), humans from a Caucasian population (CAU: dark blue), and chimpanzees (black). This plot was generated after removing the outlier, H20961, identified in Figure 1-Figure Supplement 9.

**Figure 1-Figure Supplement 11: Inter-species variation is greater than intra-species variation.** H3K9me3 ChIP-seq normalized counts (divided by size factor) in 150,390 orthologous regions (counts >0 in >8 individuals). **(A)** Mean H3K9me3 counts in 5 humans versus mean counts in 4 humans. **(B)** Mean counts in 3 chimpanzees versus mean counts in 4 chimpanzees. **(C)** Mean human count (9 individuals) versus mean chimpanzee count (7 individuals).

**Figure 1-Figure Supplement 12: Differential enrichment analysis identifies H3K9me3 regions enriched in each species. (A)** Distribution of *P* values from the *DESeq2* differential enrichment analysis. **(B)** Example loci from each of the three categories determined by differential enrichment analysis. Normalized ChIP-seq counts in each individual from each species in each of the three identified categories: higher H3K9me3 enrichment in chimpanzee (Chimpanzee-enriched), similar enrichment in both species (Shared), and higher enrichment in human (Human-enriched).

**Figure 1-Figure Supplement 13: Proportion of regions differentially enriched between species is robust with respect to the filtering approach used.** Instead of requiring >8 individuals across both species to have >0 counts per region, a species-specific filtering step was used such that only regions with >0 counts in >4 chimpanzees and >5 humans were included (greater than half of the total individuals in each species). MA plots at four different FDR thresholds (10, 5, 1, 0.1 % FDR) with the proportion of differentially enriched regions are shown. At 1% FDR, 11% of genes are differentially enriched as in the species-independent filtering approach.

**Figure 1-Figure Supplement 14: Majority of H3K9me3 regions differentially enriched between species have a small effect size.** The proportion of regions at each absolute  $\log_2$  fold change identified between species estimated following adaptive shrinkage. 87% of H3K9me3 regions have an absolute  $\log_2$  fold change between species less than 1.

**Figure 2-Figure Supplement 1: Orthologous TEs reflect the distribution of TE classes in the human and chimpanzee genome. (A)** Bioinformatic pipeline to generate orthologous TEs. **(B)** The proportion of each TE class in the human genome (hg19), chimpanzee genome (panTro3), and our orthologous TE set.

**Figure 3-Figure Supplement 1: Species-specific SVA elements overlap H3K9me3 regions less frequently than orthologous SVAs.** Left panel: overlap of orthologous SVA elements by orthologous H3K9me3 regions. Middle panel: all SVA elements in the chimpanzee genome and the proportion that overlap all H3K9me3 regions identified in chimpanzee. Right panel: all SVA elements in the human genome and the proportion that overlap all H3K9me3 regions identified in human. The estimated age of each of the six SVA types is shown in millions of years (Myr)(Wang *et al.*, 2005).

**Figure 4-Figure Supplement 1: TEs across families have similar levels of H3K9me3 enrichment between species. (A)** The proportion of TEs in each family that do not overlap an orthologous H3K9me3 region (magenta), overlap an orthologous region and show similar H3K9me3 enrichment in both species (grey: Shared), are enriched in human (blue: Human-enriched), or are enriched in chimpanzee (brown: Chimpanzee-enriched) are shown. **(B)** The proportion of each TE family that overlaps an orthologous H3K9me3 region and is either Shared, Human-enriched or Chimpanzee-enriched. The total number of orthologous elements within each TE family is shown on the left.

**Figure 4-Figure Supplement 2: HERV TE types have variable overlap with orthologous H3K9me3 regions and inter-species enrichment. (A)** The proportion of HERV Class I, II and III types that do not overlap an orthologous H3K9me3 region (magenta), overlap an orthologous region and show similar H3K9me3 enrichment in both species (grey: Shared), are enriched in human (blue: Human-enriched), or are enriched in chimpanzee (brown: Chimpanzee-enriched) are shown. **(B)** The proportion of each HERV family member that overlaps an orthologous H3K9me3 region and is either Shared, Human-enriched or Chimpanzee-enriched. Curated list of HERV elements from Turelli *et al.*

**Figure 4-Figure Supplement 3: LINE-1 TE types are similarly enriched for H3K9me3 in humans and chimpanzees. (A)** The proportion of a subset of L1 family members that do not overlap an orthologous H3K9me3 region (magenta), show similar H3K9me3 enrichment in both species (grey: Shared), are enriched in human (blue: Human-enriched), or are enriched in chimpanzee (brown: Chimpanzee-enriched) are shown. Elements are ordered from youngest to oldest. **(B)** The proportion of each L1 family member that overlaps an orthologous H3K9me3

region and is either Shared, Human-enriched or Chimpanzee-enriched. Curated list of L1 elements from Castro-Diaz *et al.*

**Figure 5-Figure Supplement 1: Longer TE families overlap with orthologous H3K9me3 regions more often than shorter families. (A)** The length of TEs in each of the TE classes stratified by H3K9me3 enrichment category. **(B)** The length of TEs in eleven major TE families ordered by the proportion of family members overlapping H3K9me3. **(C)** The proportion of TE family members that overlap an orthologous H3K9me3 region, and the median TE length in each TE family ordered by the proportion of TEs within a family overlapping H3K9me3 from highest to lowest. Pearson's correlation coefficient is represented as  $r$ . **(D)** The median length of TEs not overlapping H3K9me3 regions and the median length of TEs that do overlap an H3K9me3 region. Asterisk denotes a significant difference between the Shared and Unmarked categories (Wilcoxon-rank sum test;  $*P < 0.001$ )

**Figure 5-Figure Supplement 2: TE copy number is weakly anti-correlated with silencing by H3K9me3. (A)** The number of instances of all family members within a family. **(B)** The relationship between the proportion of TEs in each family that overlap an orthologous H3K9me3 region, and the number of instances of TE family members. Pearson's correlation coefficient is represented as  $r$ .

**Figure 5-Figure Supplement 3: TEs, within a class, that are similarly enriched for H3K9me3 in both species tend to be less diverged from the consensus TE sequence than TEs that are not marked. (A)** The divergence from the consensus sequence of TEs in each of the TE classes. **(B)** The divergence from the consensus sequence of TEs in eleven major TE families. Families are ordered by the proportion of TE instances that overlap H3K9me3 regions (highest-to-lowest from

left-to-right). Asterisk denotes a significant difference between the Shared and Unmarked categories (Wilcoxon-rank sum test; \* $P < 0.001$ ).

**Figure 5-Figure Supplement 4: TEs, within a class, that are similarly enriched for H3K9me3 in both species tend to be further away from the nearest TSS than TEs that are not marked.**

**(A)** The absolute distance between each TE in each of the TE classes and the nearest TSS. **(B)** The absolute distance between each TE in each of the TE families and the nearest TSS. Families are ordered by the proportion of TE instances that overlap H3K9me3 regions (highest-to-lowest from left-to-right). Asterisk denotes a significant difference between the Shared and Unmarked categories (Wilcoxon-rank sum test; \*  $P < 0.001$ ).

**Figure 5-Figure Supplement 5: TEs preferentially silenced by H3K9me3 in human are depleted for regions of active chromatin in human relative to TEs preferentially silenced in chimpanzee.** Orthologous TEs categorized by H3K9me3 silencing category that overlap chromatin states as determined by the ChromHMM algorithm in the H1 hESC line, HepG2 liver cancer cell line, and GM12878 lymphoblastoid cell line.

**Figure 5-Figure Supplement 6: TEs preferentially silenced by H3K9me3 in human are depleted for regulatory regions in human relative to TEs preferentially silenced in chimpanzee.** The proportion of orthologous TEs within each of the four H3K9me3 silencing categories (Shared, Human-enriched, Chimpanzee-enriched and Unmarked) that overlap with human ChIP-seq and DNase I hypersensitivity data. Only a small subset of orthologous TEs overlap human regulatory regions but plots are scaled so that differences between categories can be visualized. The number of instances in each category are indicated above each bar. Asterisk

denotes a significant difference between categories (Pearson's Chi-squared test;  $*P < 0.01$ ,  $**P < 0.001$ ,  $***P < 0.0001$ ).

**Figure 6-Figure Supplement 1: RNA-seq data clusters by species.** (A) The number of aligned RNA-seq read-pairs for each individual in each species. (B) Spearman correlation of read counts between individuals at orthologous autosomal genes. (C) PCA analysis of  $\log_2\text{cpm}$  values at 17,354 genes with  $>0$  counts in  $>10$  individuals for humans from a Yoruba population (YRI: light blue), humans from a Caucasian population (CAU: dark blue), and chimpanzees (black).

**Figure 6-Figure Supplement 2: TEs overlapping H3K9me3 regions upstream of genes associate with decreased gene expression.** Quantile normalized gene expression levels for genes with TEs within 1, 10 and 20 kb windows of the TSS that do not overlap (TE - H3K9me3) or overlap (TE + H3K9me3) with at least one H3K9me3 region. Data is shown for human (blue) and chimpanzee (black). Asterisk denotes a significant difference between categories (Wilcoxon-rank sum test;  $P < 0.001$ ).

**Figure 6-Figure Supplement 3: TEs preferentially enriched for H3K9me3 in one species only have a minimal effect on gene expression divergence at the TSS.** (A) Gene expression divergence between human and chimpanzee for genes with TEs overlapping H3K9me3 within 1 kb upstream of the TSS. Genes are categorized by H3K9me3 enrichment at  $\text{FDR} < 0.01$ . (B) Interspecies divergence in H3K9me3-mediated silencing of TEs in a 1 kb window upstream of the TSS versus gene expression divergence. Genes which have a H3K9me3 log fold divergence of  $>1$  or  $<-1$ , and a gene expression divergence log fold change of  $>1$  or  $<-1$  are shown in red. (C) As in (A) except using a 20 kb window. (D) As in (B) except using a 20 kb window. (E) As in (A) except using a 40 kb window. (F) As in (B) except using a 40 kb window. There are no significant differences between categories.



**Figure 6-Figure Supplement 4: H3K9me3 preferentially enriched in one species at the TSS associates with gene expression divergence.** Gene expression divergence between human and chimpanzee at genes with H3K9me3 regions overlapping the TSS irrespective of the presence of a TE. Genes are categorized by the proportion of orthologous H3K9me3 regions that are enriched in each species (Shared, Human-enriched, Chimpanzee-enriched). Asterisk denotes a significant difference between categories (Wilcoxon-rank sum test;  $P < 0.01$ ).

**Figure 6-Figure Supplement 5: Genes that are differentially expressed between human and chimpanzee are not associated with inter-species differences in H3K9me3 enrichment at TEs distal to the TSS. (A)** Genes are stratified into three classes: those that are similarly expressed between humans and chimpanzees (dark grey), those that are up-regulated in human at  $FDR < 1\%$  (turquoise), and those that are up-regulated in chimpanzee at  $FDR < 1\%$  (dark brown). The proportion of TEs marked by H3K9me3 in 1, 10, 20 kb windows upstream of the TSS that are shared, enriched in human and enriched in chimpanzee are represented as pie charts. Box plots represent human-chimpanzee divergence of H3K9me3 enrichment overlapping TEs for each gene category in each window. **(B)** Proportion of genes that are differentially expressed between species when varying the FDR threshold to identify differentially enriched H3K9me3 windows 10 kb upstream of genes (orange). Proportion of H3K9me3 windows 10 kb upstream of genes that are differentially enriched between species when varying the threshold to identify differentially expressed genes (red). Asterisk denotes a significant difference between categories (Wilcoxon-rank sum test;  $P < 0.01$ ).

**Figure 6-Figure Supplement 6: TEs in human enhancer regions do not associate with inter-species gene expression divergence.** Inter-species gene expression divergence of orthologous

genes closest to TEs overlapping H3K9me3 regions and overlapping human H3K4me1/H3K4me3 regions identified in H1 hESCs. The closest orthologous genes to all human H3K4me1/H3K4me3 regions in H1 cells are used for comparison. Asterisk denotes a significant difference between categories (Wilcoxon rank-sum test;  $P < 0.01$ ).

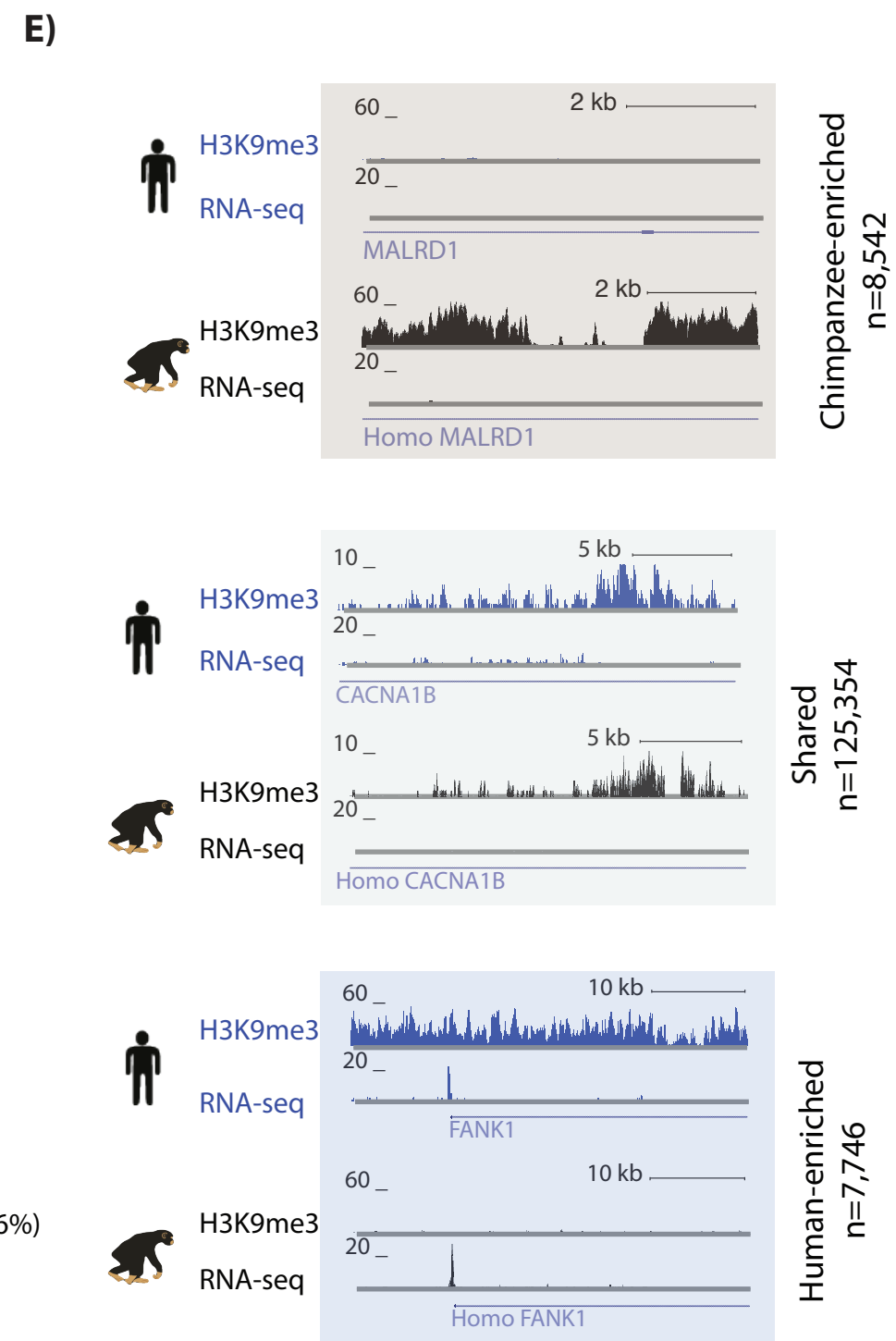
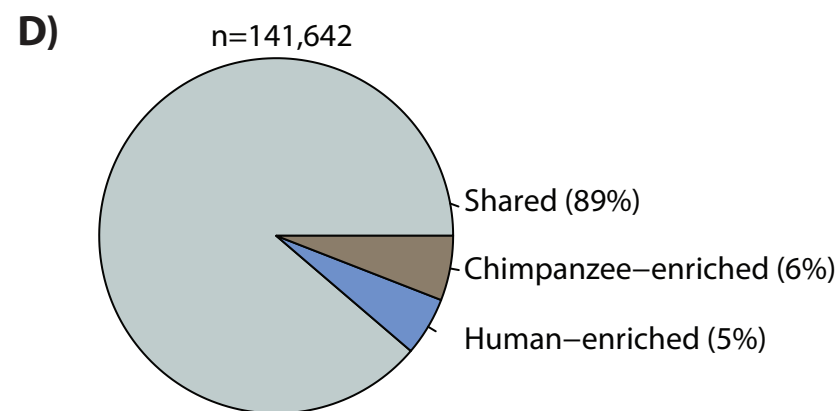
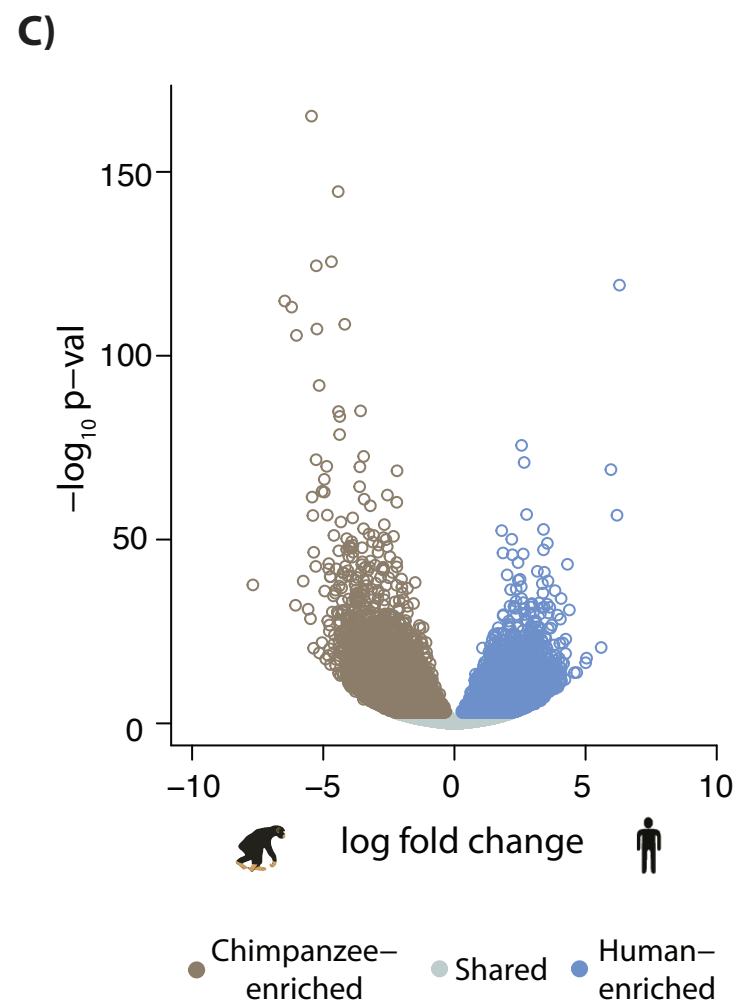
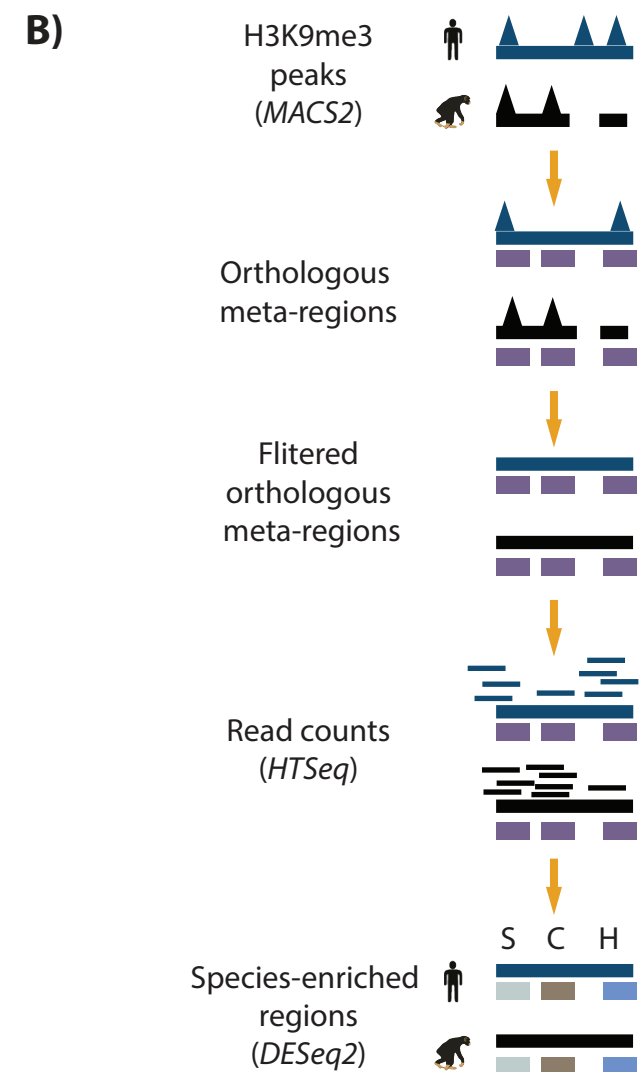
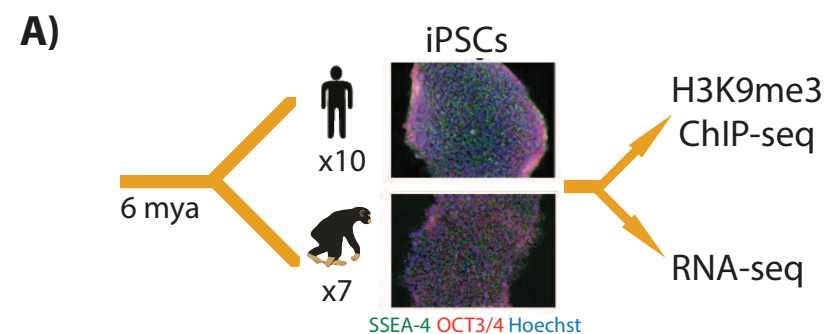
**Figure 6-Figure Supplement 7: Non-orthologous TE silencing distal to the TSS does not consistently influence gene expression divergence between species. (A)** Schematic representation of the non-orthologous (n.o) TE analysis. **(B)** Gene expression divergence for orthologous genes with at least one non-orthologous TE in 1, 10, 20 kb windows upstream of the TSS. Genes are categorized as non-silent (at least one TE does not overlap H3K9me3) or silent (all TEs overlap H3K9me3). Asterisk denotes a significant difference between categories (Wilcoxon-rank sum test;  $P < 0.01$ ).

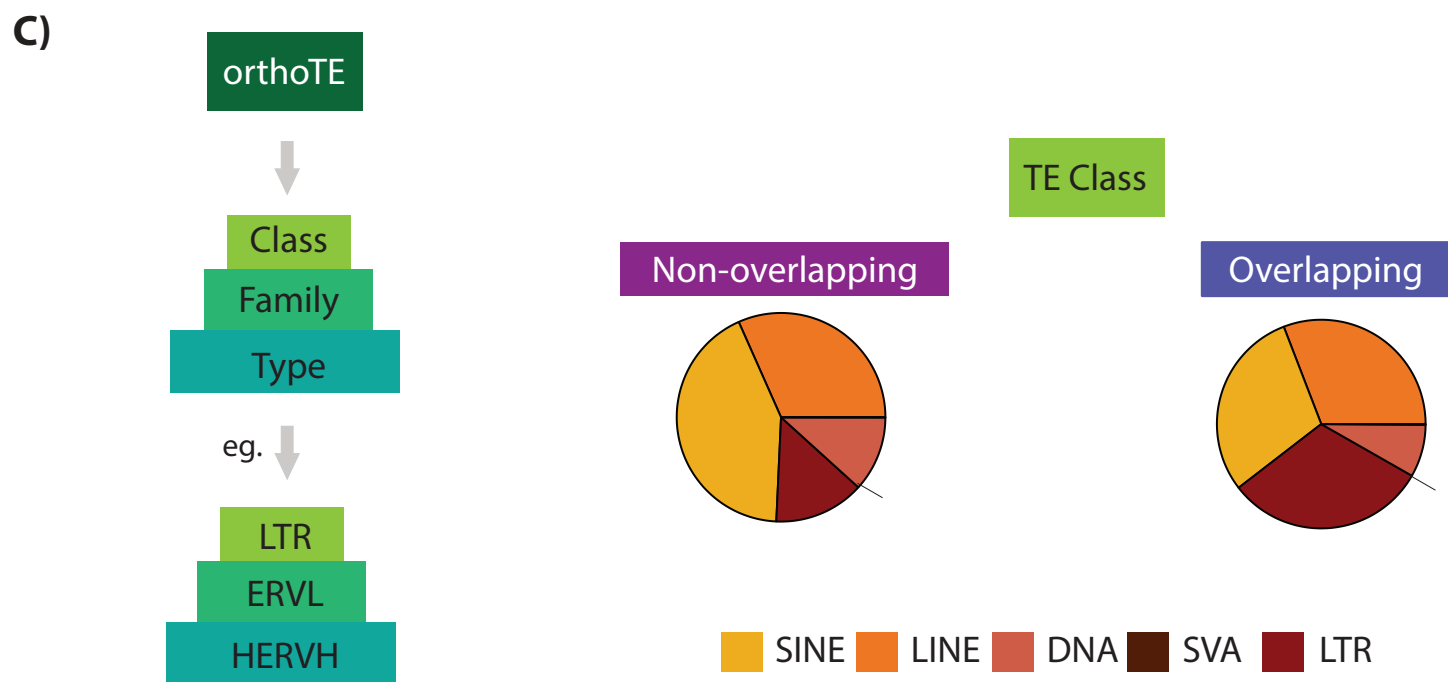
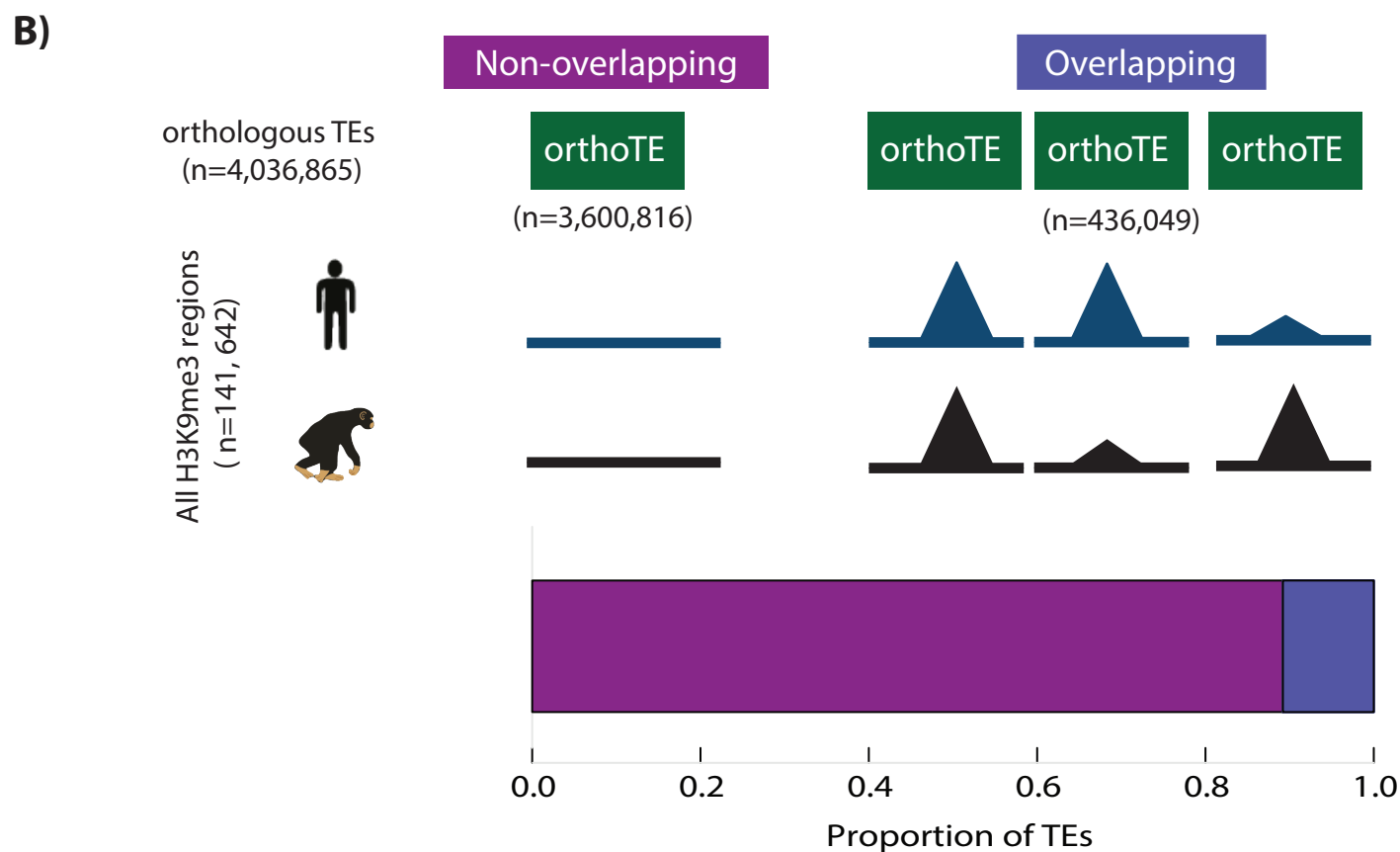
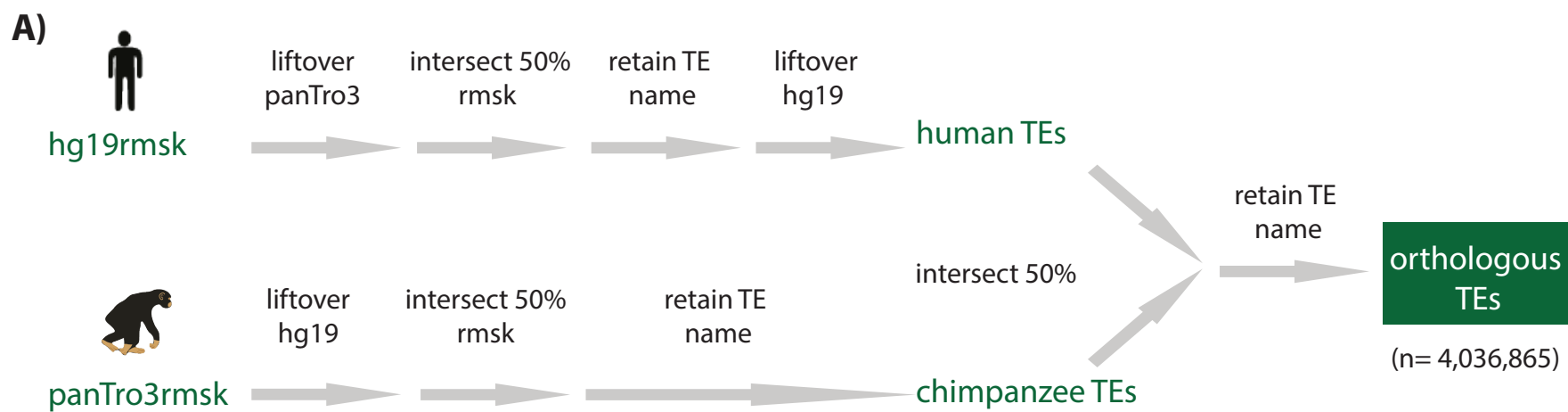
**Figure 6-Figure Supplement 8: Orthologous KRAB-ZNF genes overlap with orthologous H3K9me3 regions.** The proportion of orthologous KRAB-ZNF genes curated by Kapopoulou *et al.* that do not overlap an orthologous H3K9me3 region (magenta), overlap an orthologous region and show similar H3K9me3 enrichment in both species (grey: Shared), are enriched in human (blue: Human-enriched), or are enriched in chimpanzee (brown: Chimpanzee-enriched) are shown. **(B)** The proportion of KRAB-ZNF genes that overlap an orthologous H3K9me3 region and is either Shared, Human-enriched or Chimpanzee-enriched.

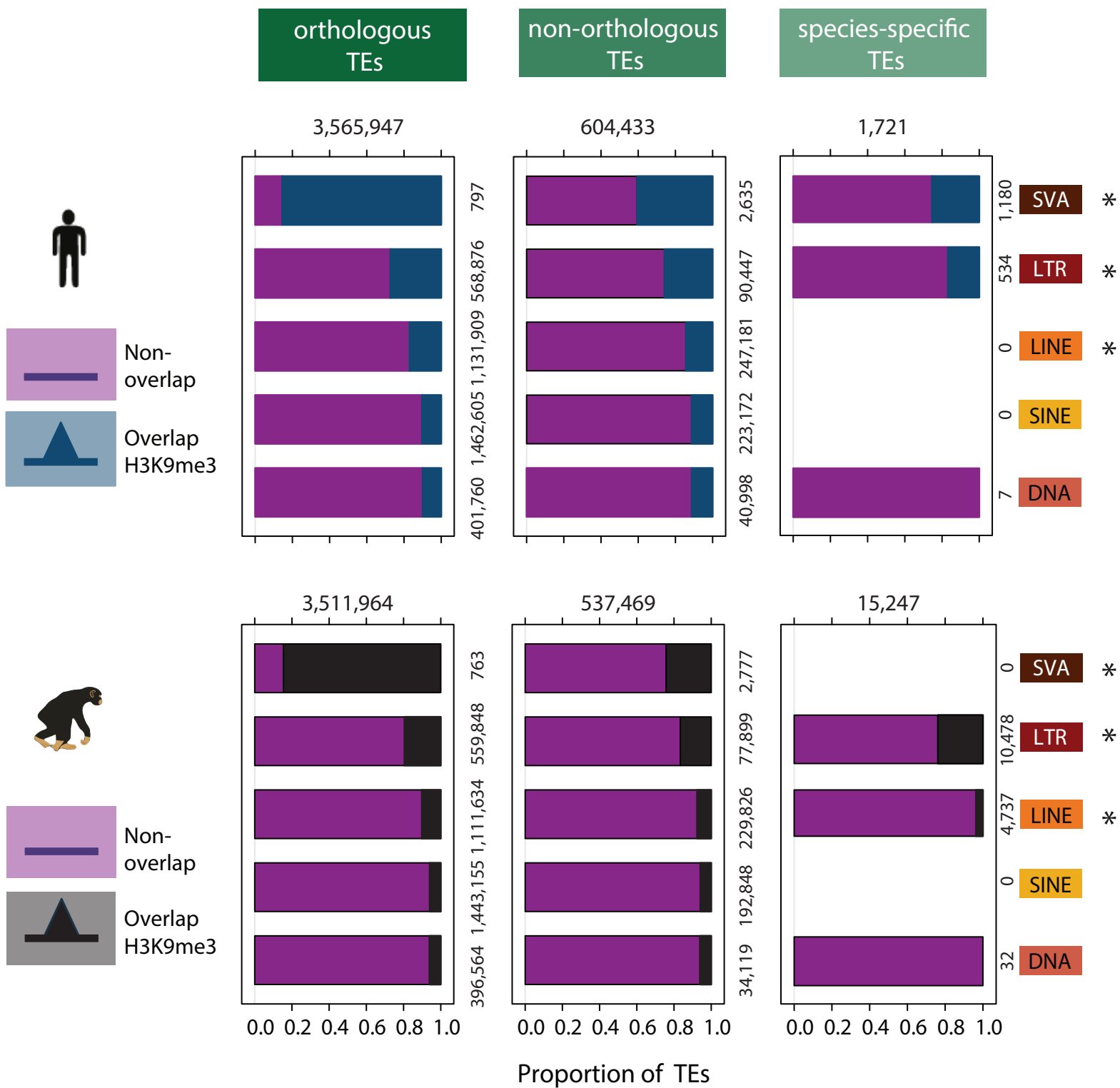
**Figure 6-Figure Supplement 9: KRAB-ZNF genes are no more likely to be differentially expressed between species than all genes. (Ai)** The expression of 256 orthologous KRAB-ZNF genes curated by Kapopoulou *et al.* Genes that are identified as being significantly differentially expressed at an FDR  $< 1\%$  are represented in red (131/256 genes=51%). **(Aii)** The expression of

all orthologous genes (8,420/17,354 genes differentially expressed = 49%). **(B)** KRAB-ZNF gene expression divergence stratified by H3K9me3 enrichment at the TSS. Genes with higher expression in chimpanzee and significantly lower H3K9me3 in chimpanzee compared to human (blue box): ZNF582, ZNF471, ZNF585A, ZNF667, ZNF583. Genes with higher expression in human and significantly lower H3K9me3 in human compared to chimpanzee (brown box): ZFP82, ZNF662, ZNF429, ZNF100, ZNF43.

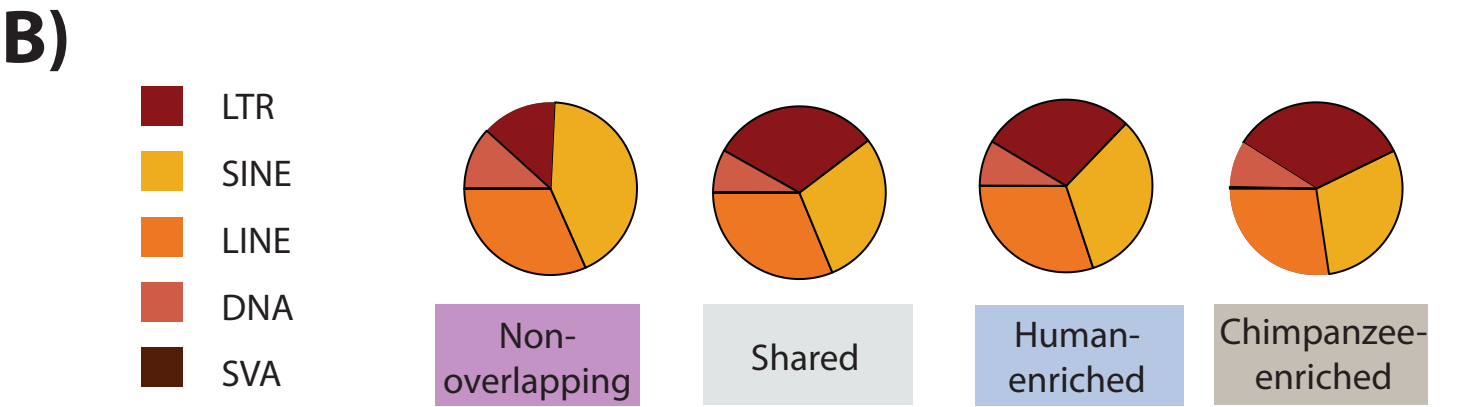
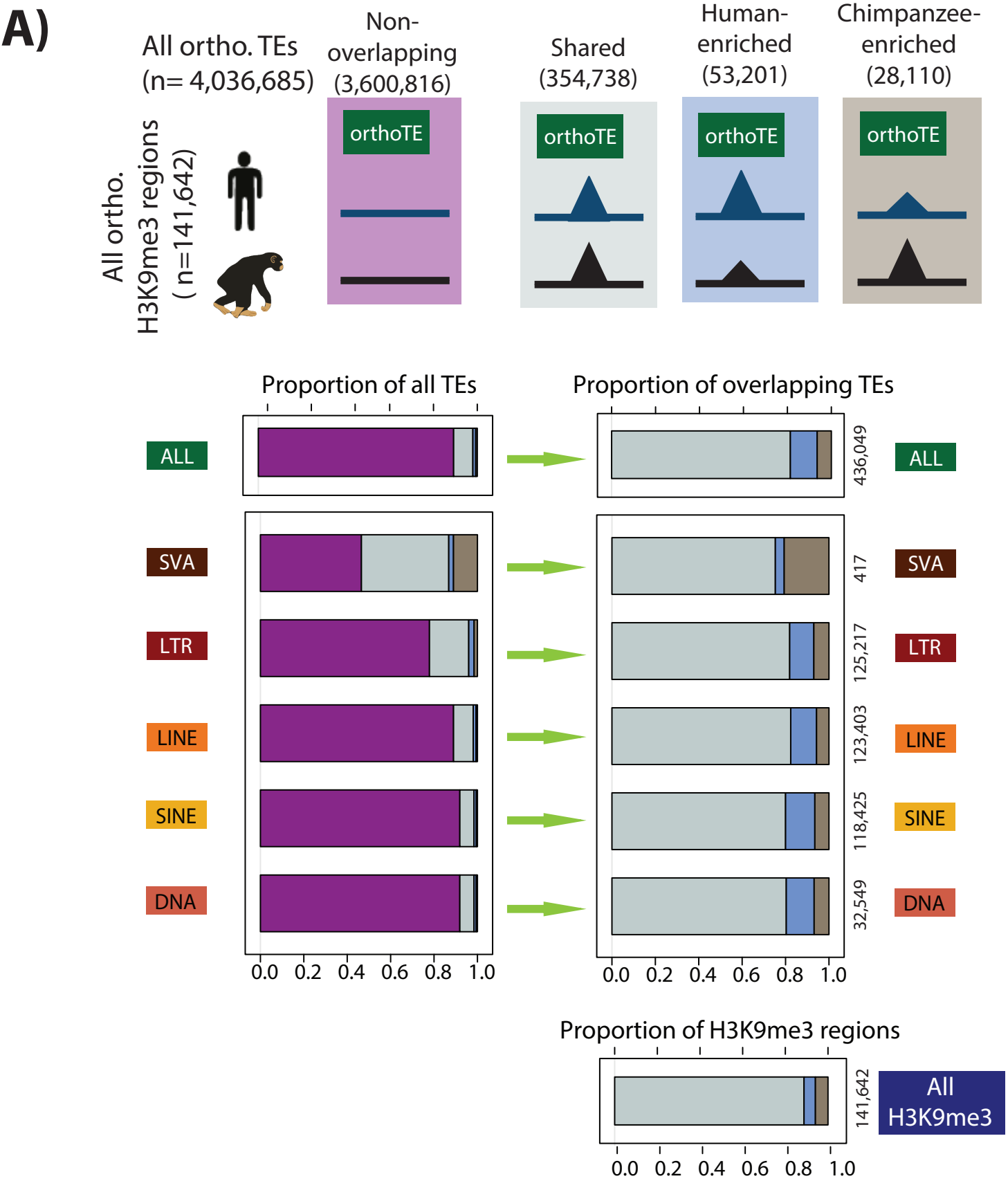
**Figure 6-Figure Supplement 10: TEs preferentially silenced by H3K9me3 in chimpanzee are not more often associated with increased levels of H3K27me3 compared to TEs preferentially silenced by H3K9me3 in human.** The proportion of orthologous TEs within each of the four H3K9me3 silencing categories (Shared, Human-enriched, Chimpanzee-enriched and Unmarked) that overlap with human H3K27me3 ChIP-seq data in H1 hESCs. The number of instances in each category are indicated above each bar. Asterisk denotes a significant difference between categories (Pearson's Chi-squared test; \*\*\* $P < 0.0001$ ).

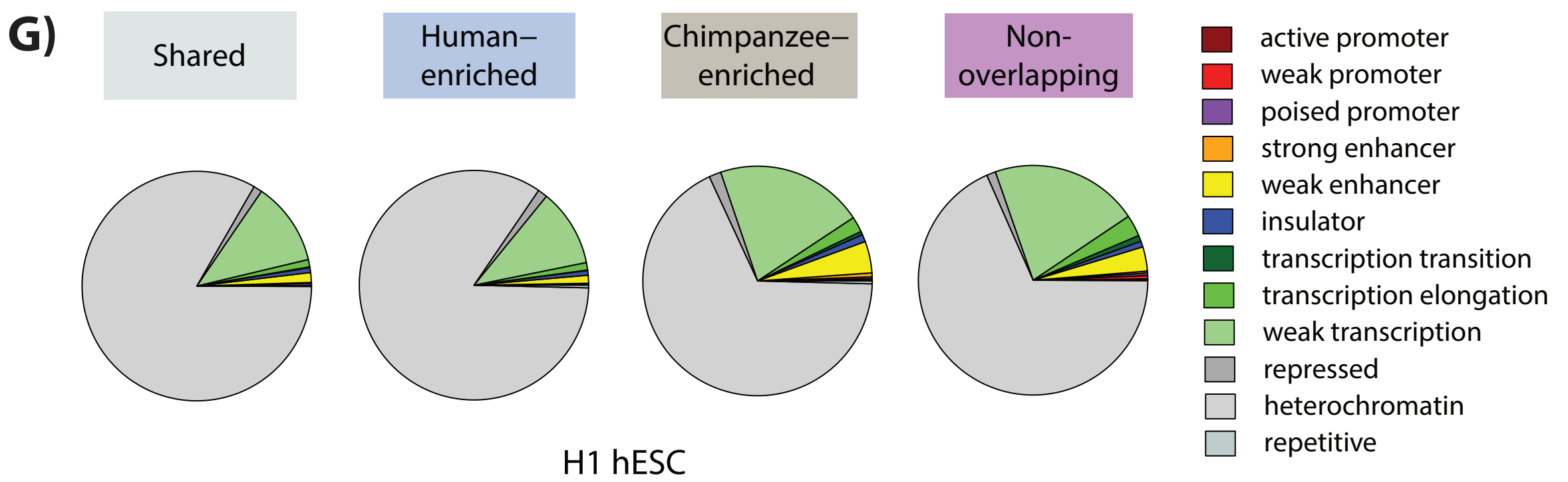
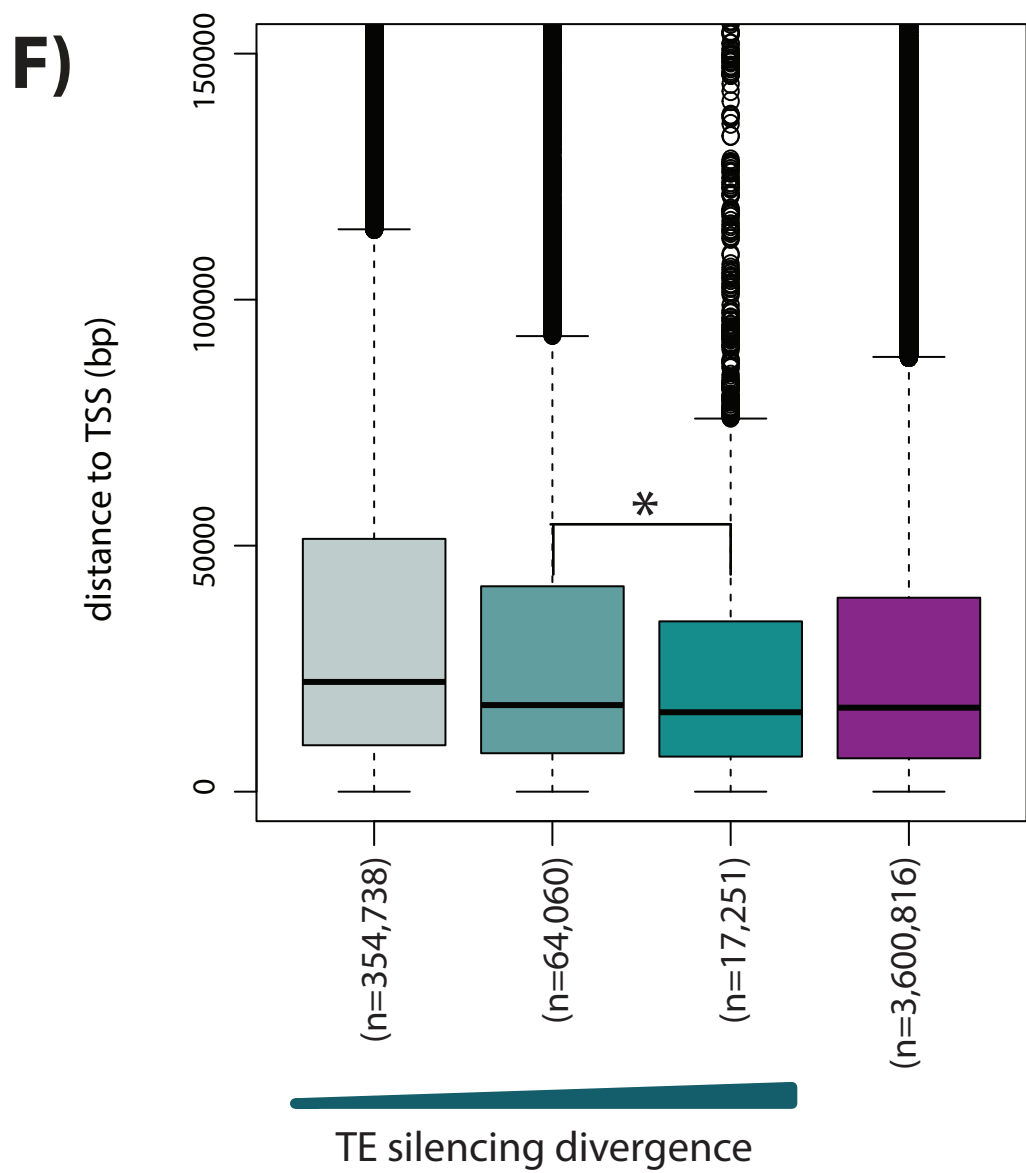
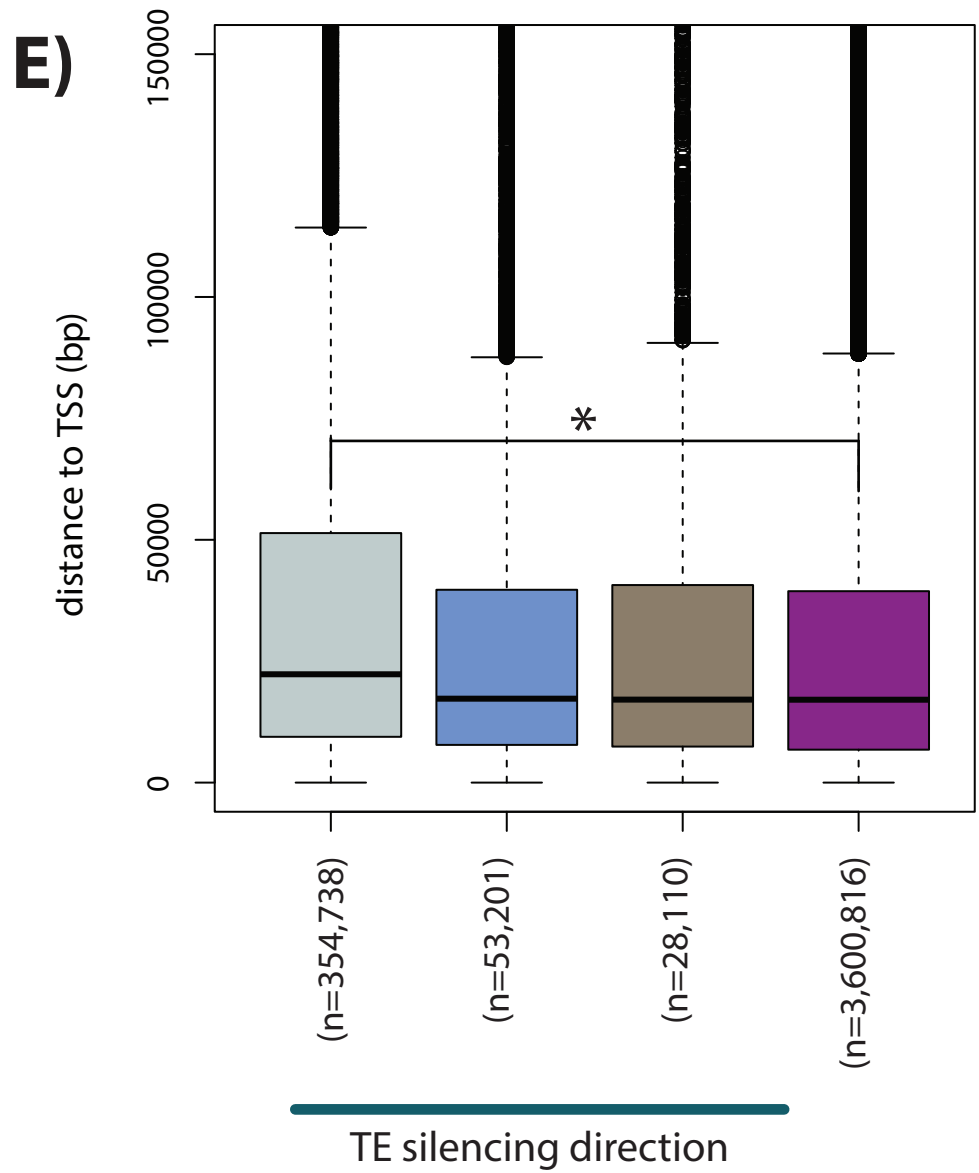
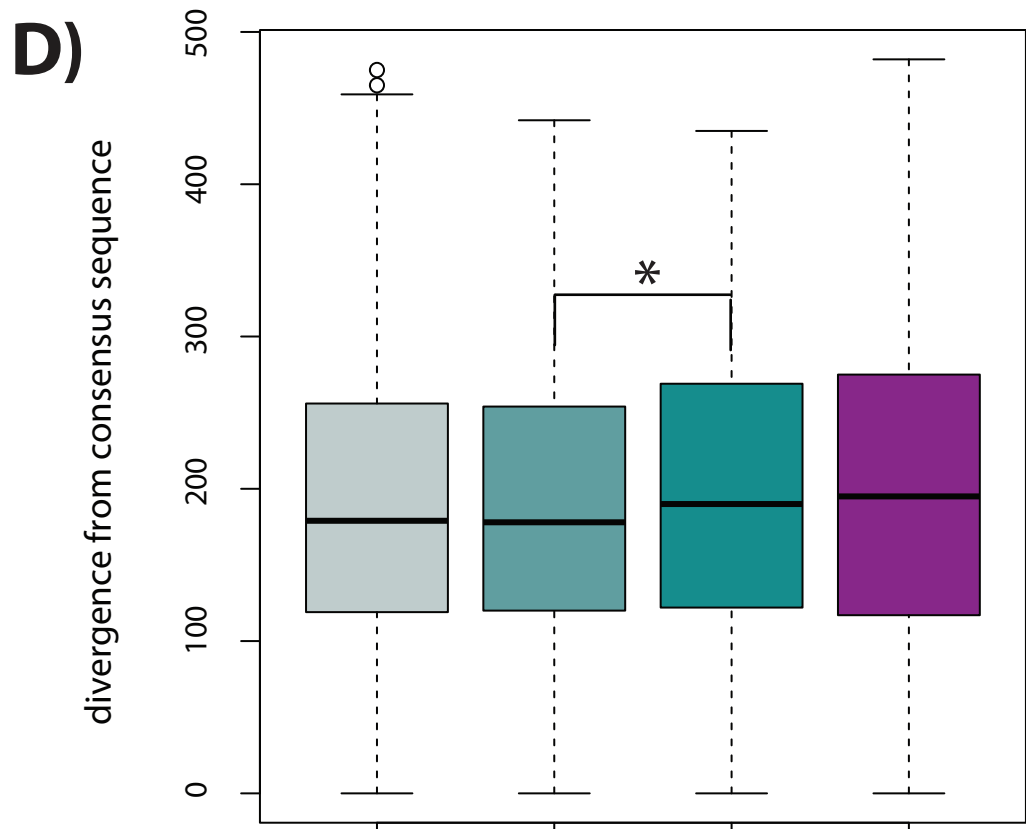
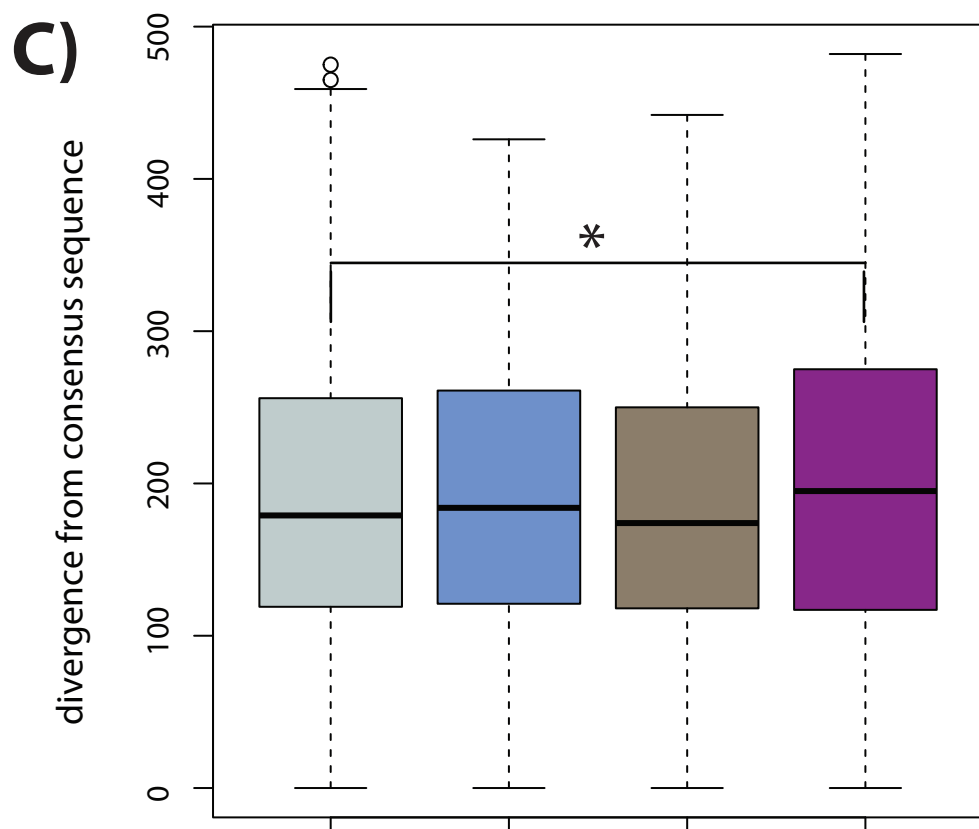
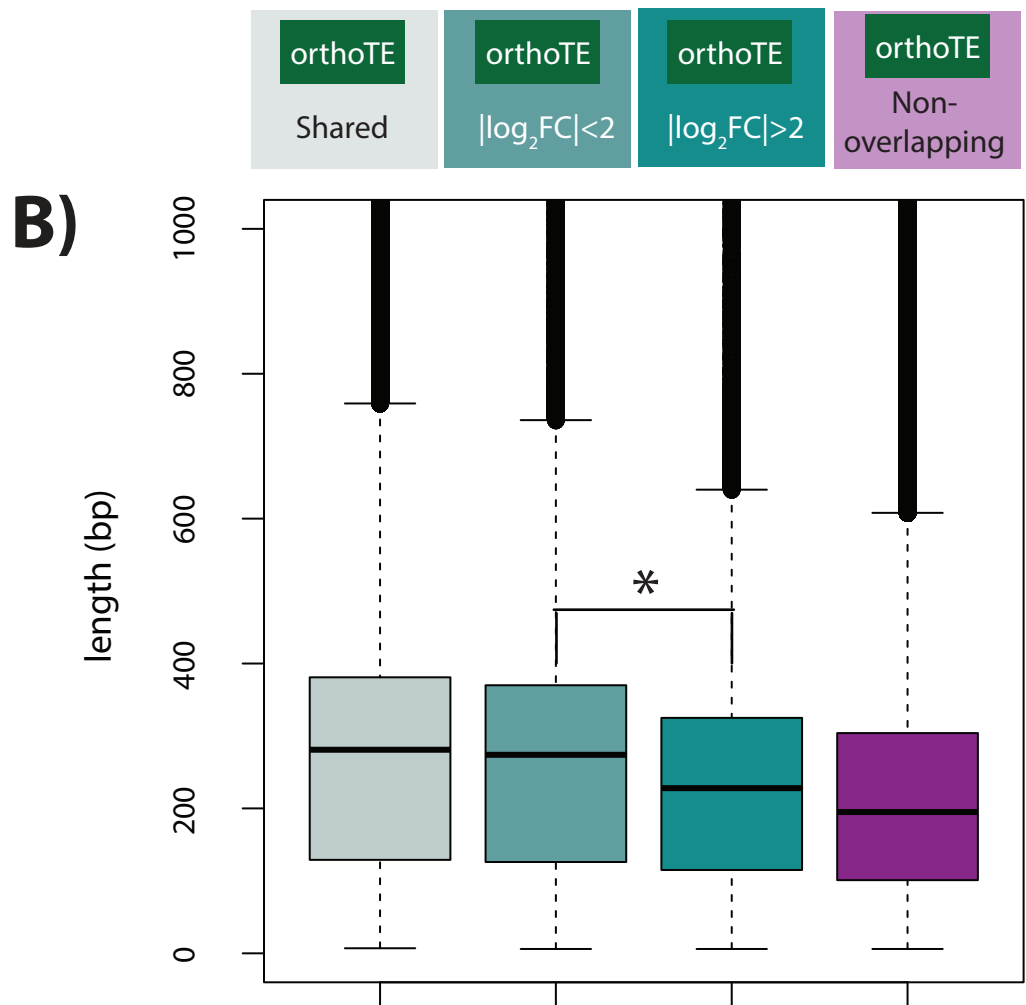
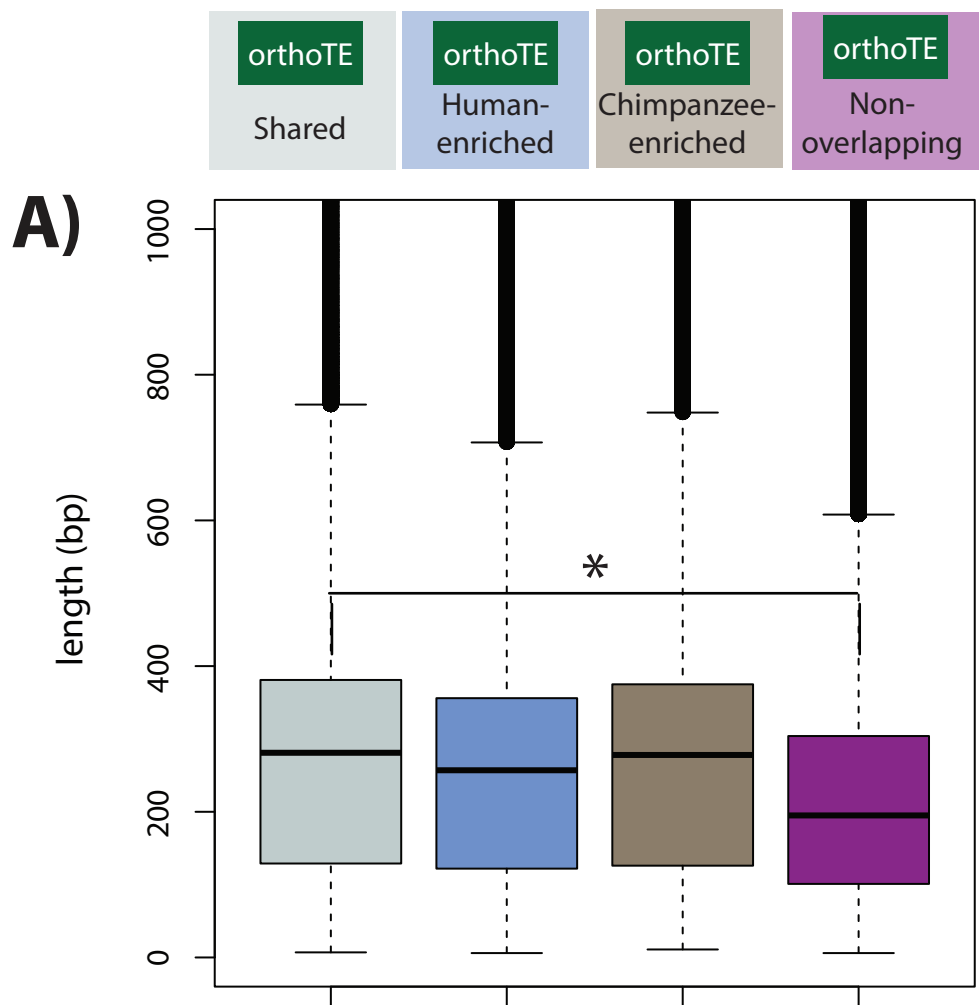




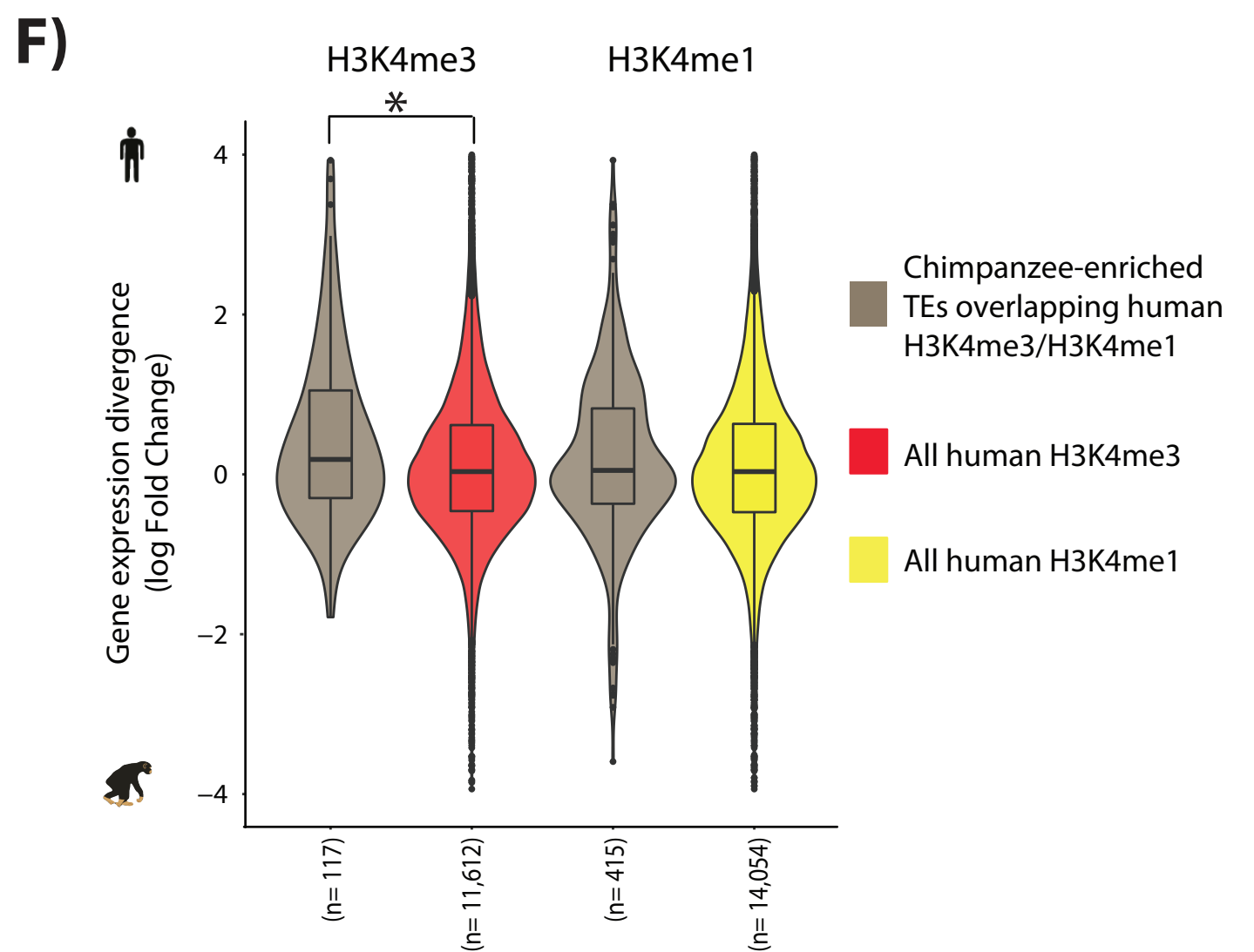
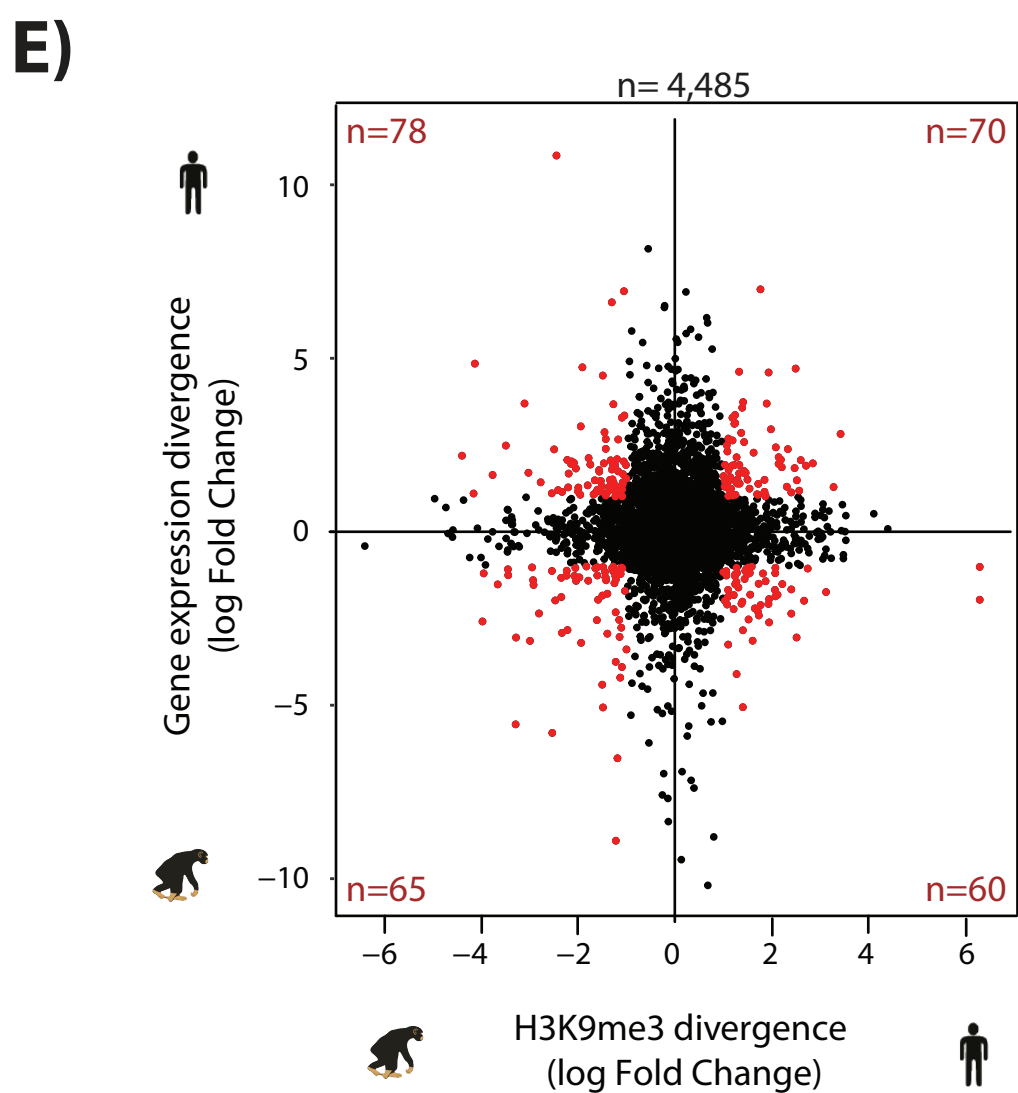
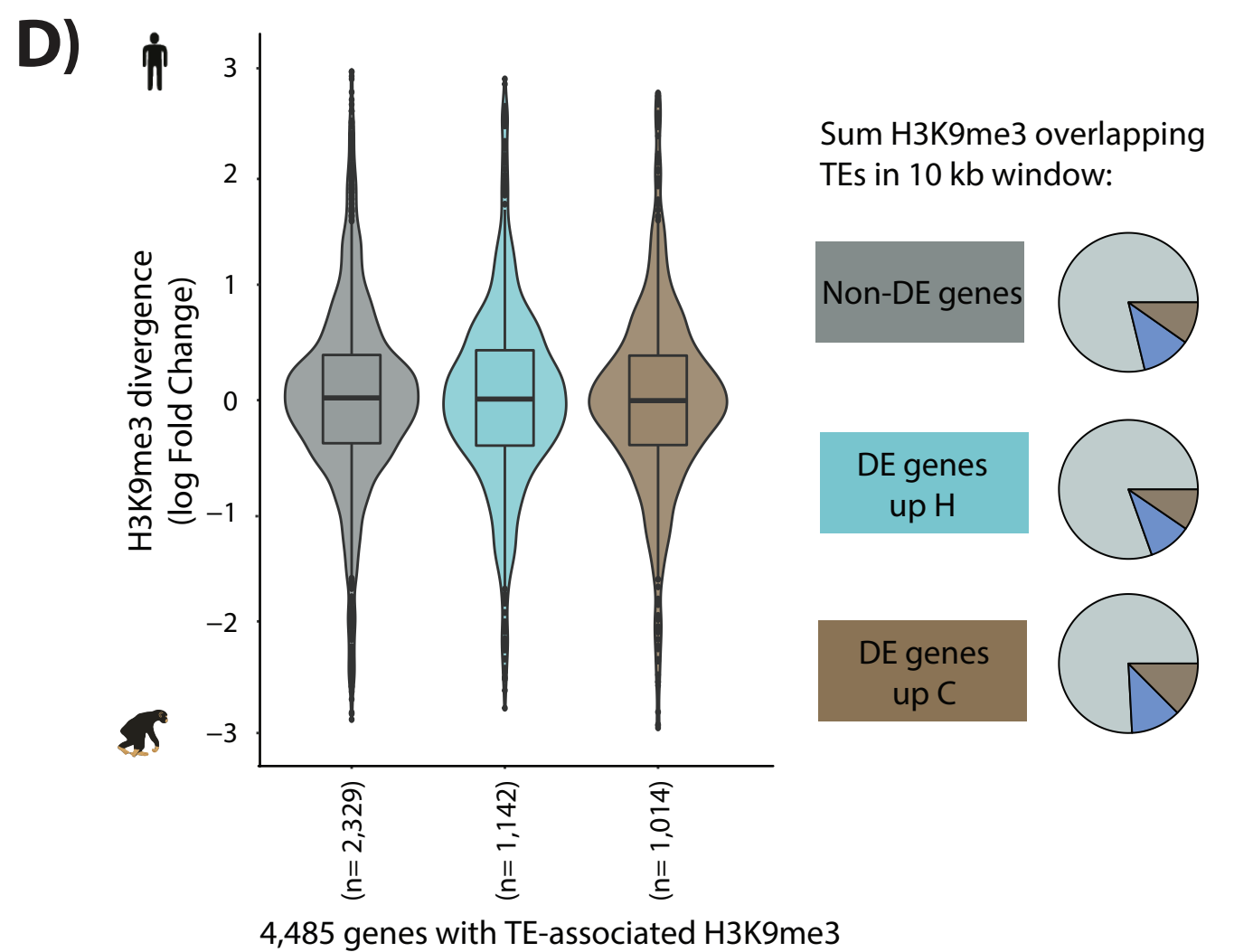
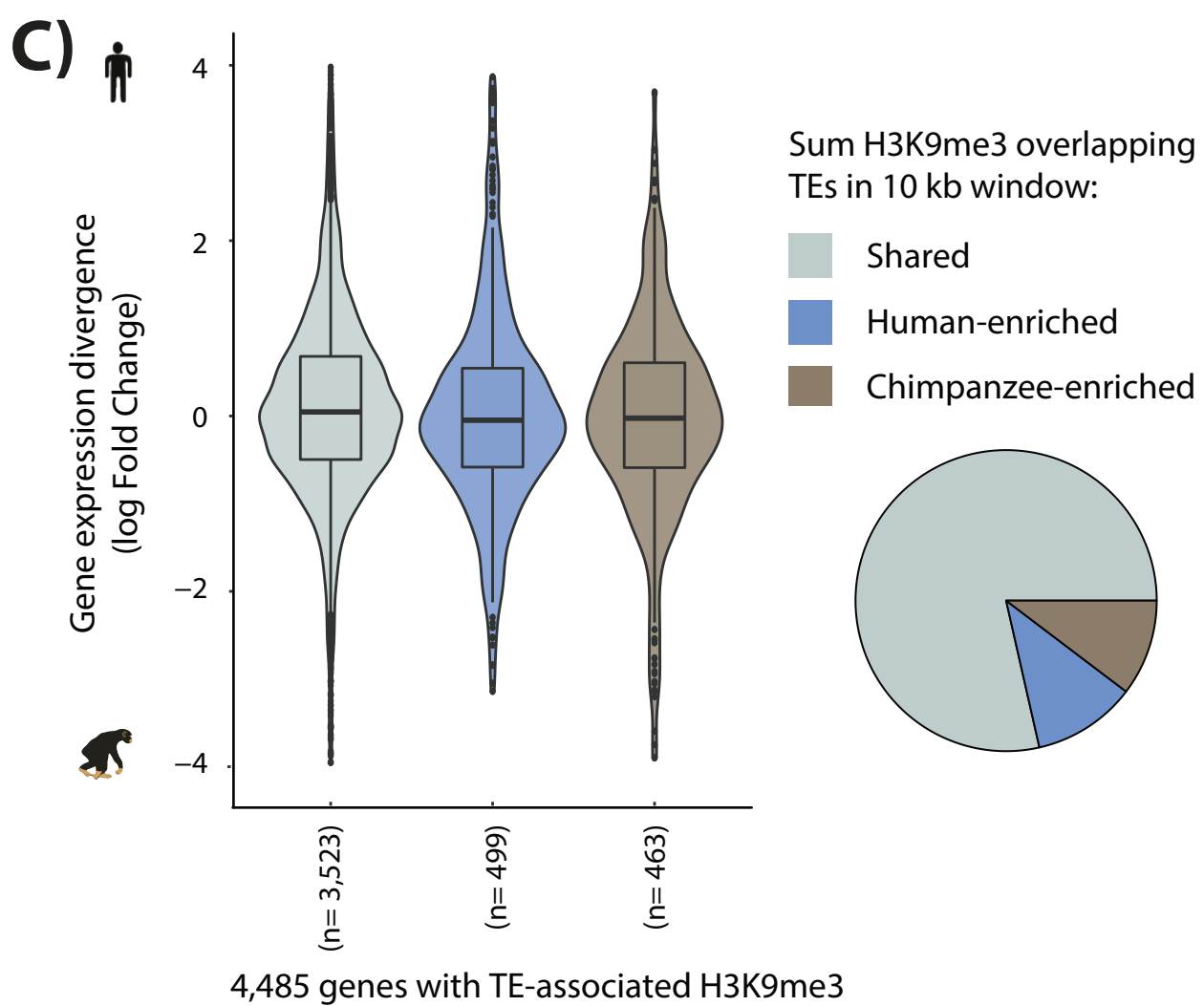
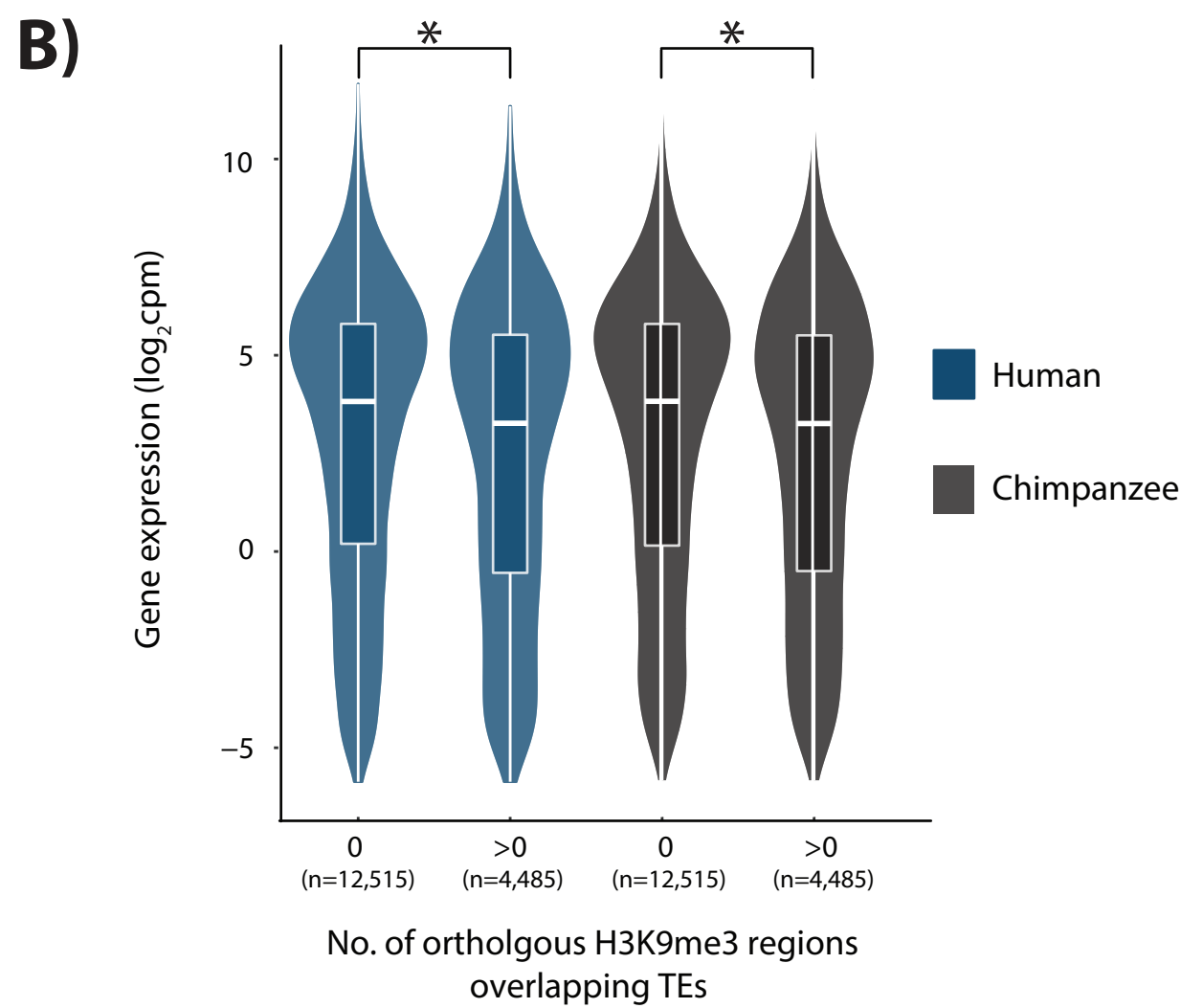
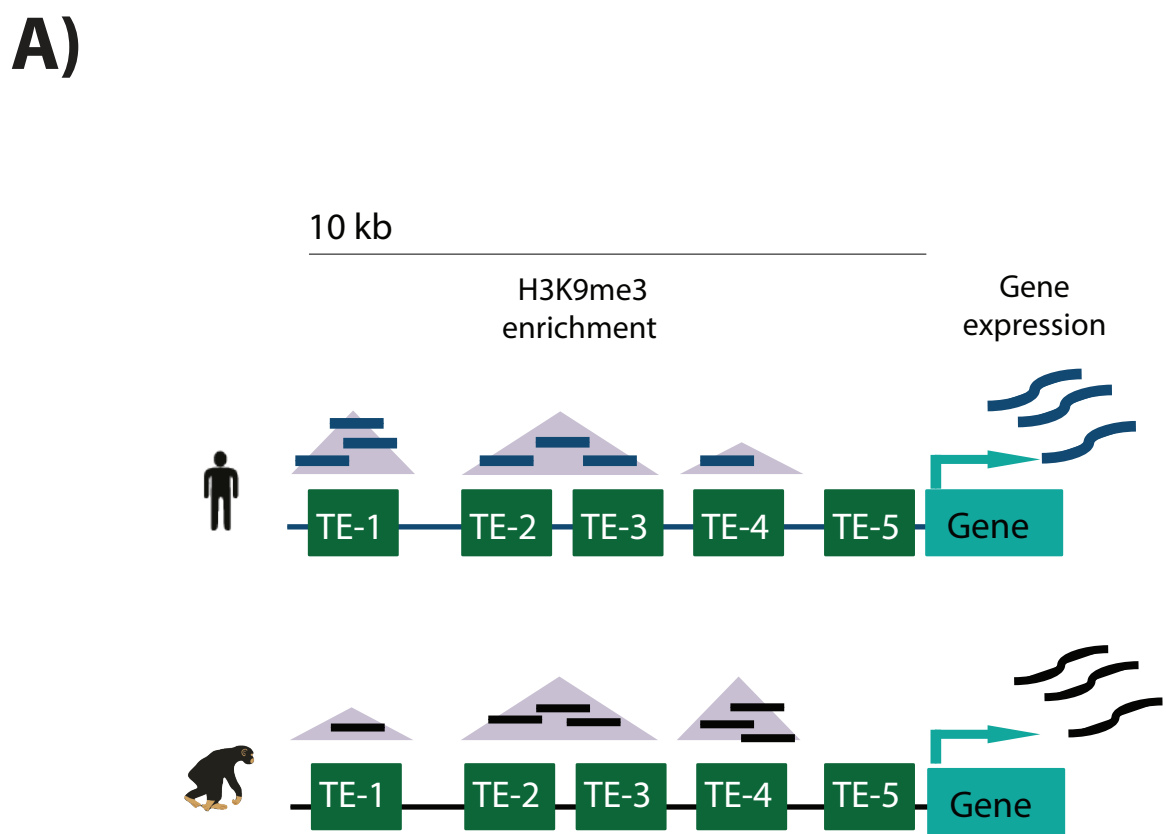


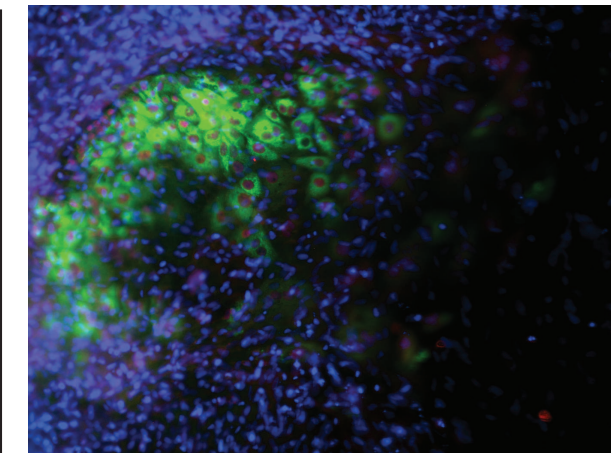
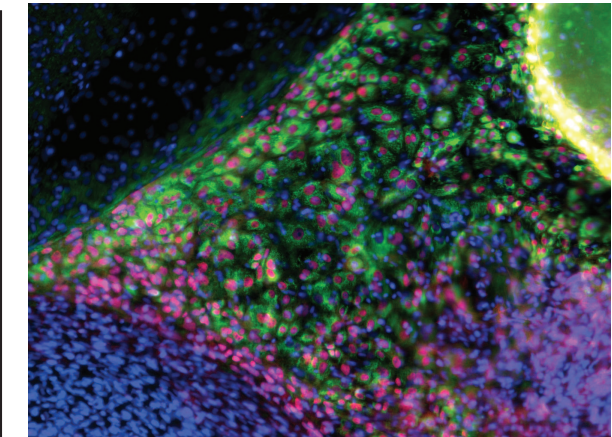
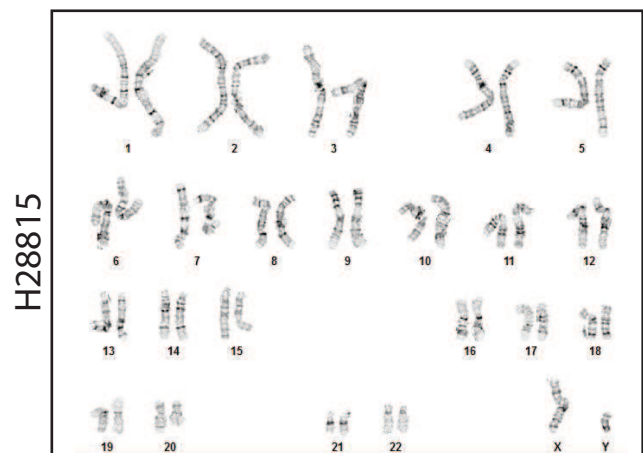
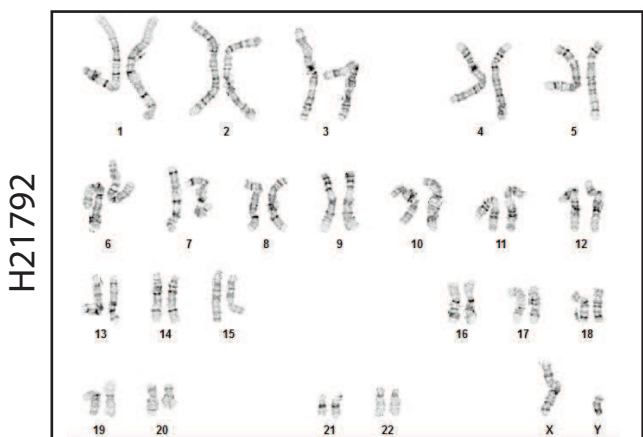
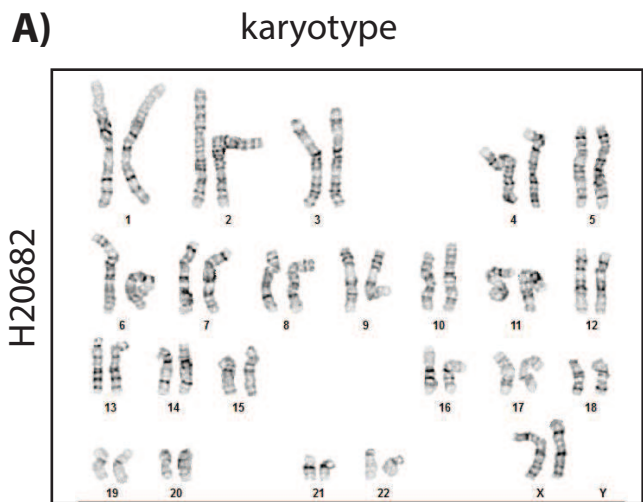
decreasing orthology



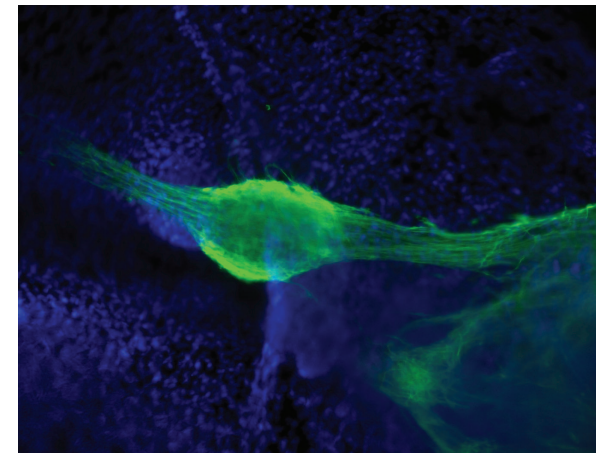
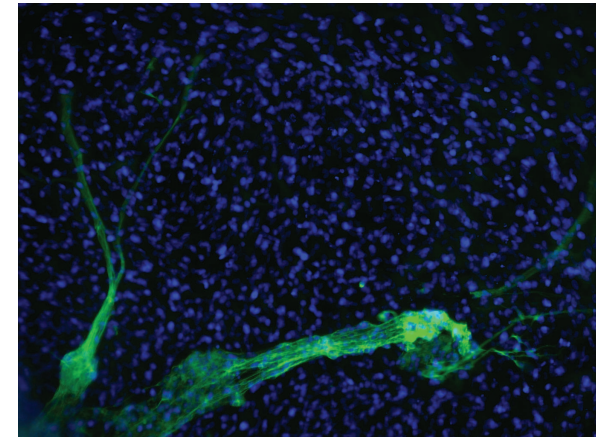
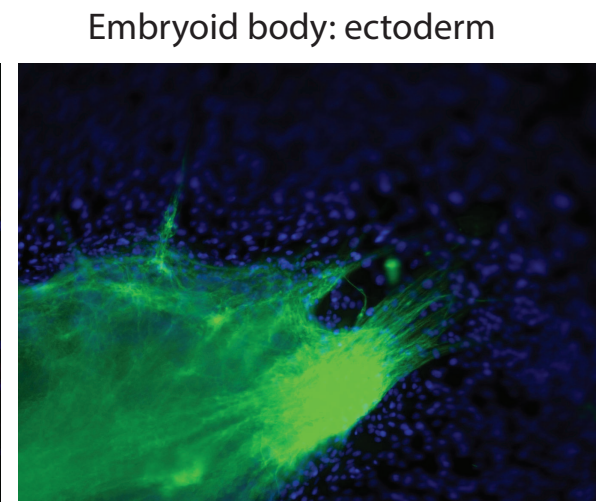




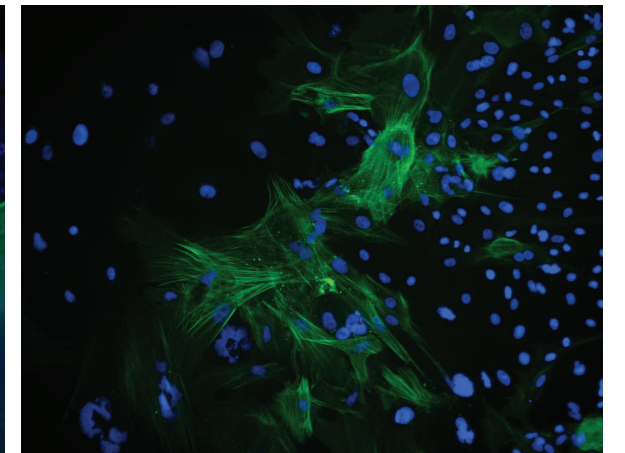
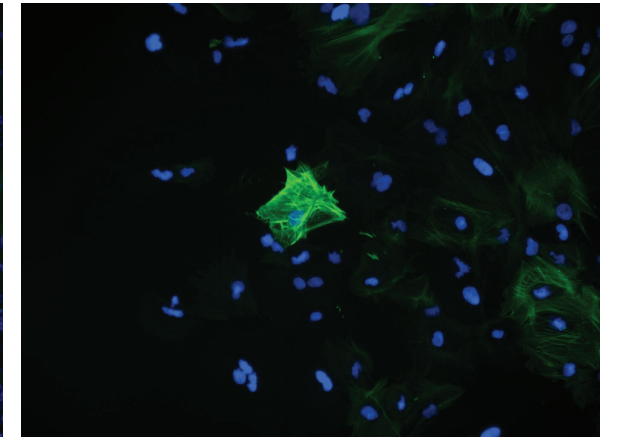




AFP/HNF3b/Hoechst

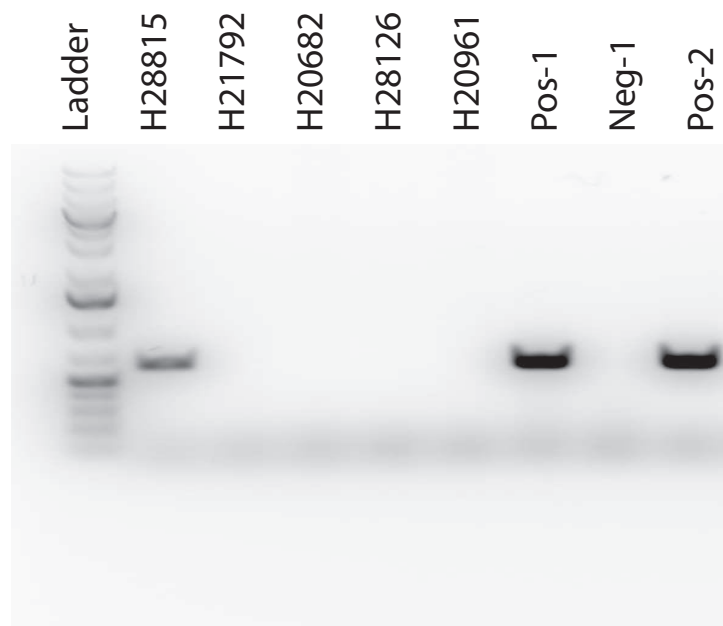


MAP2/Hoechst



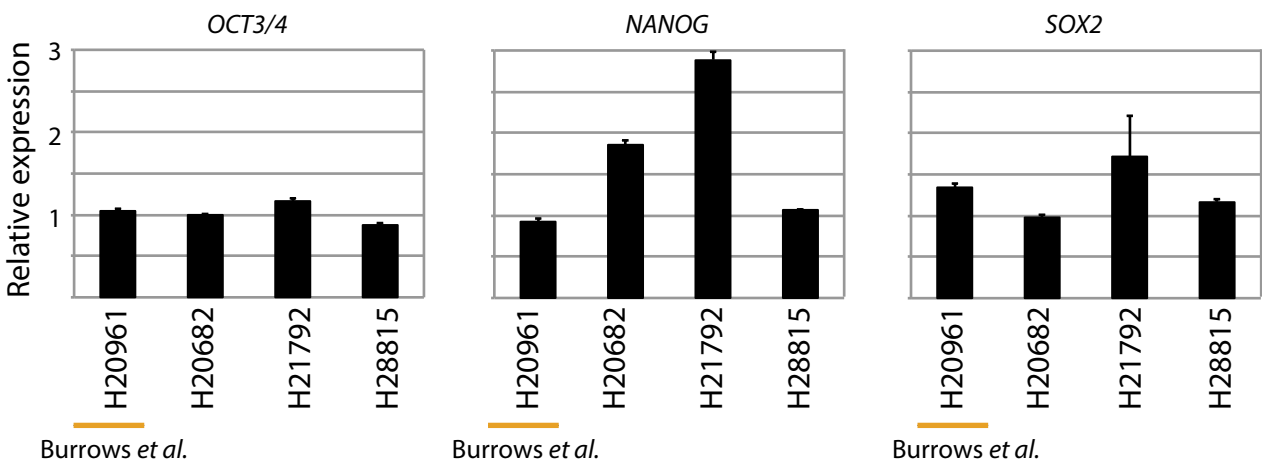
SMA/Hoechst

**C)** Burrows *et al.*

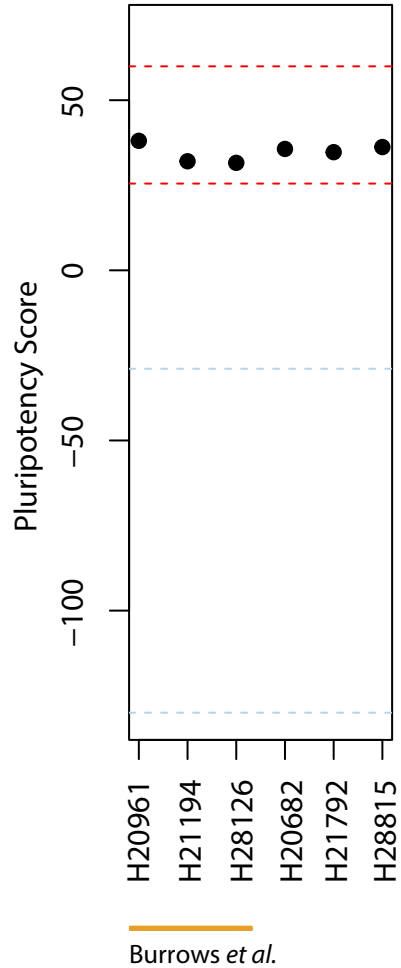




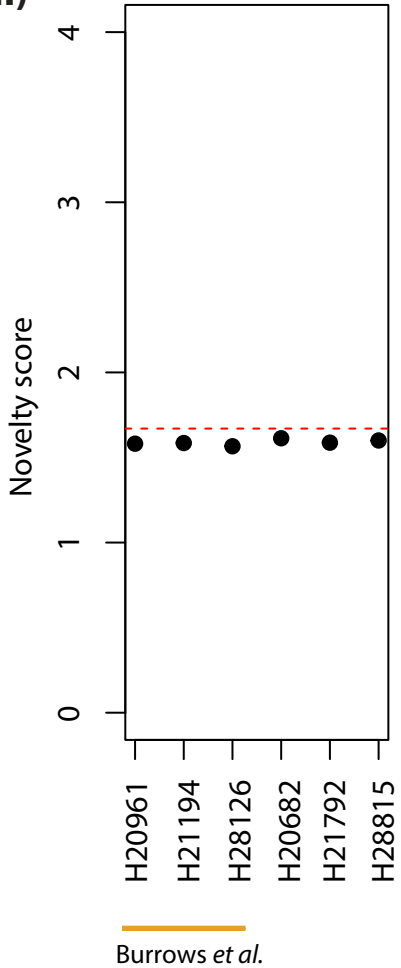
**A)**

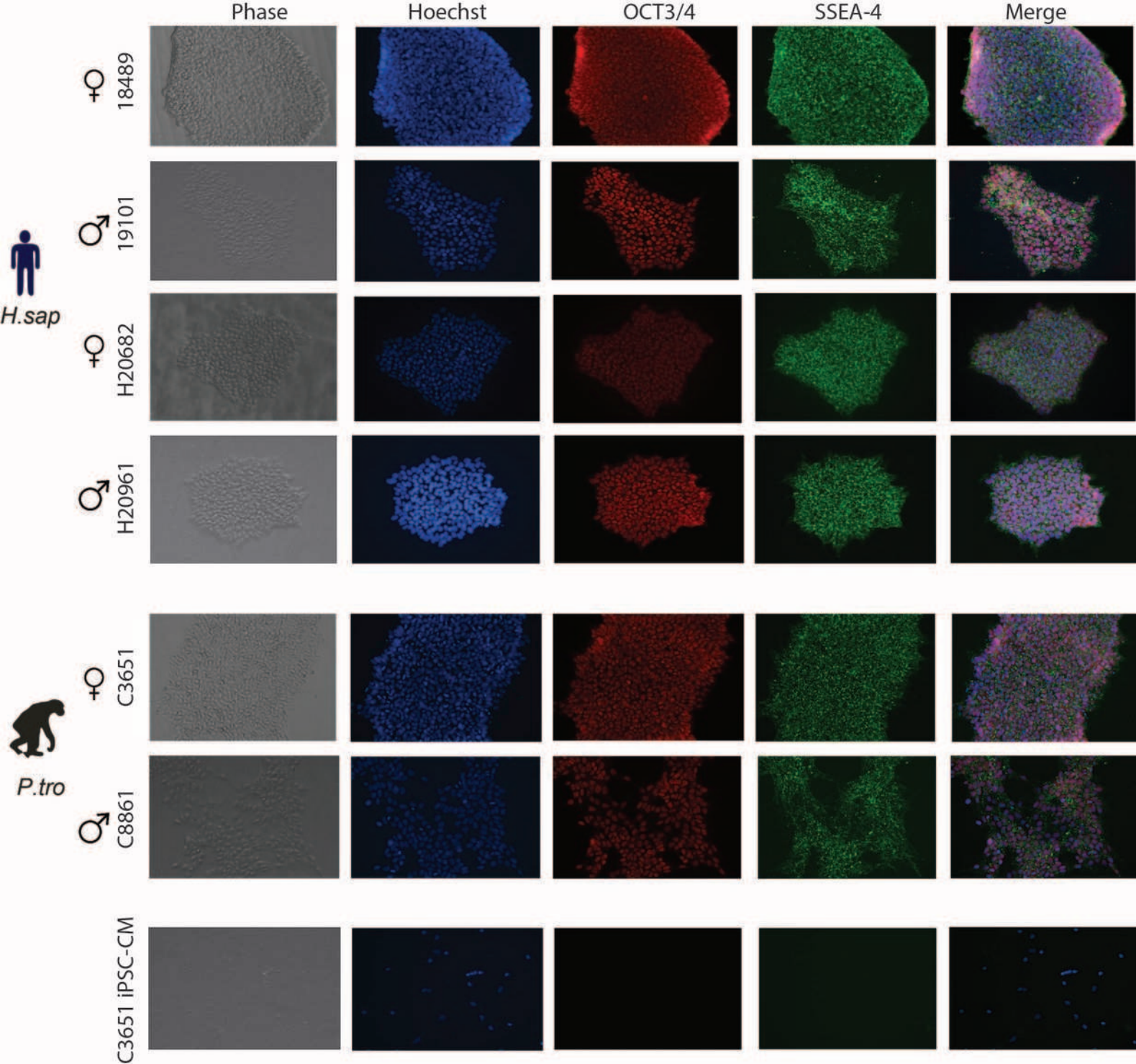


**Bi)**



**Bii)**

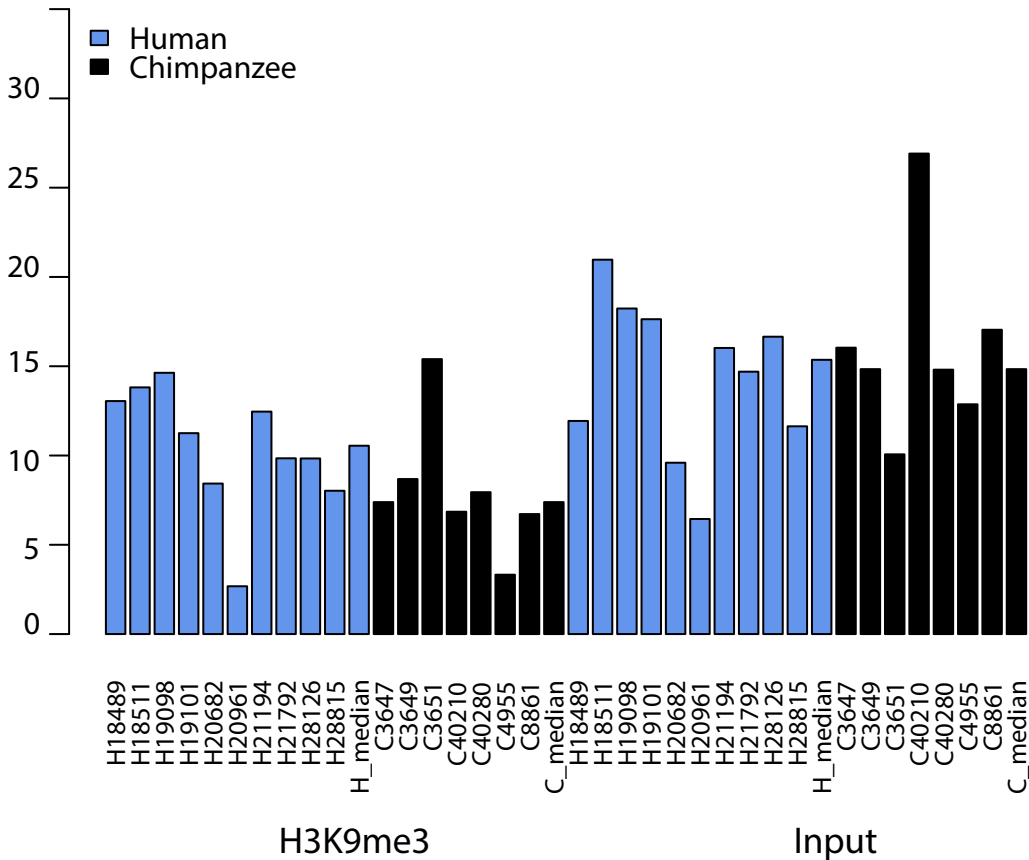


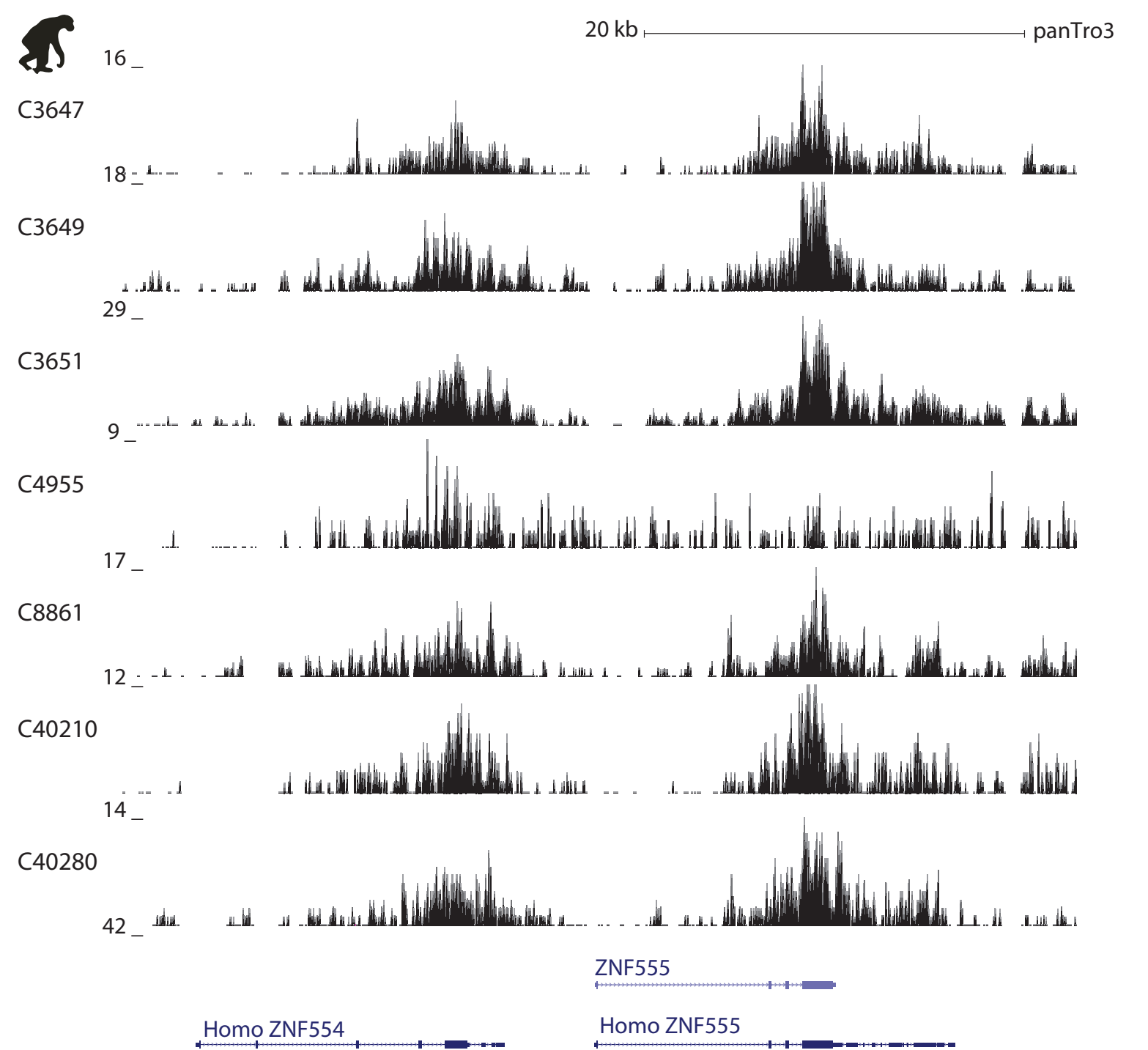
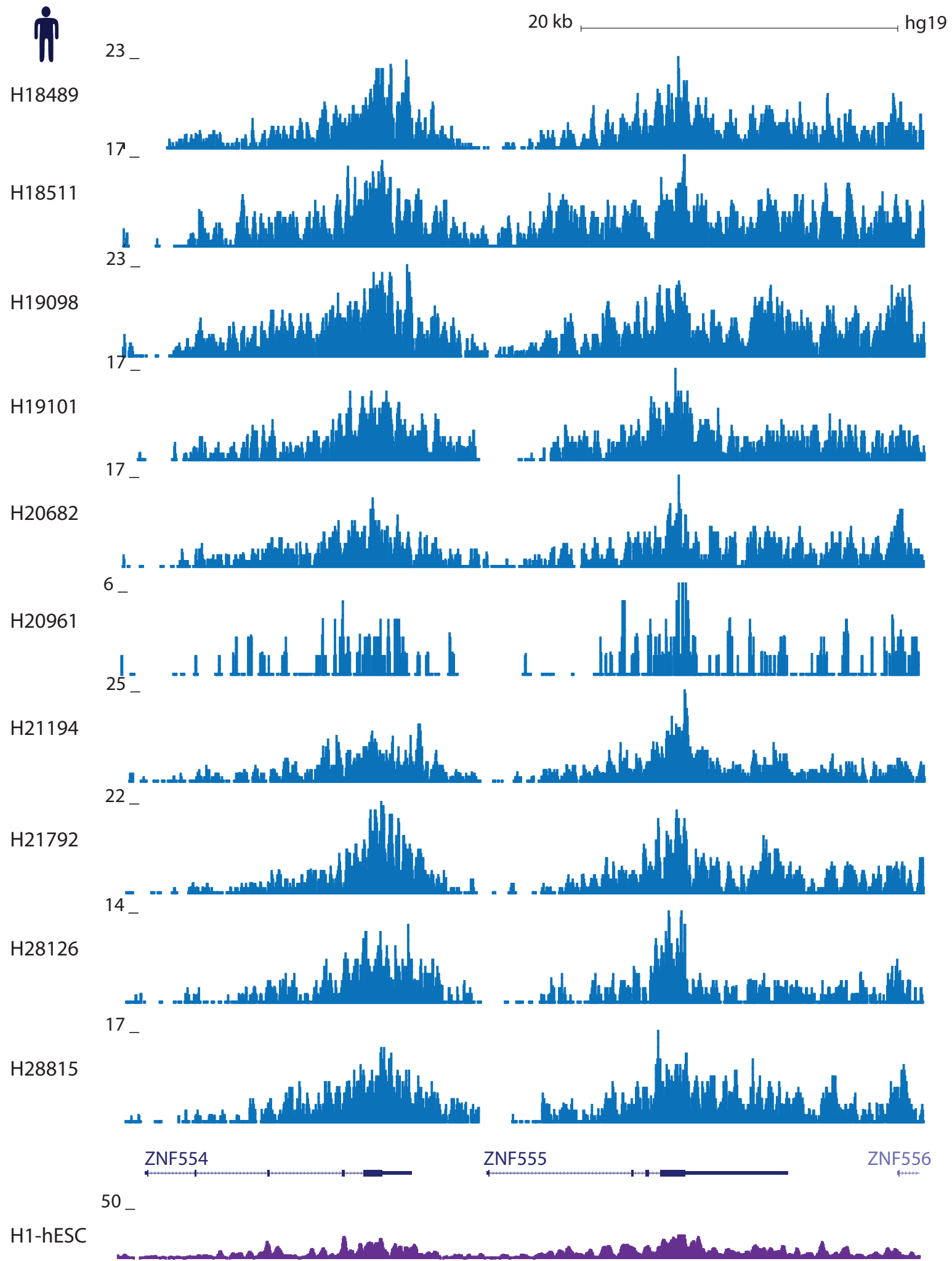


No. of aligned ChIP-seq read pairs

(millions)

Human  
Chimpanzee





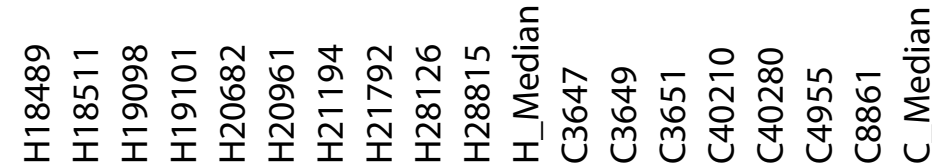
No. of peaks

120000  
100000  
80000  
60000  
40000  
20000  
0

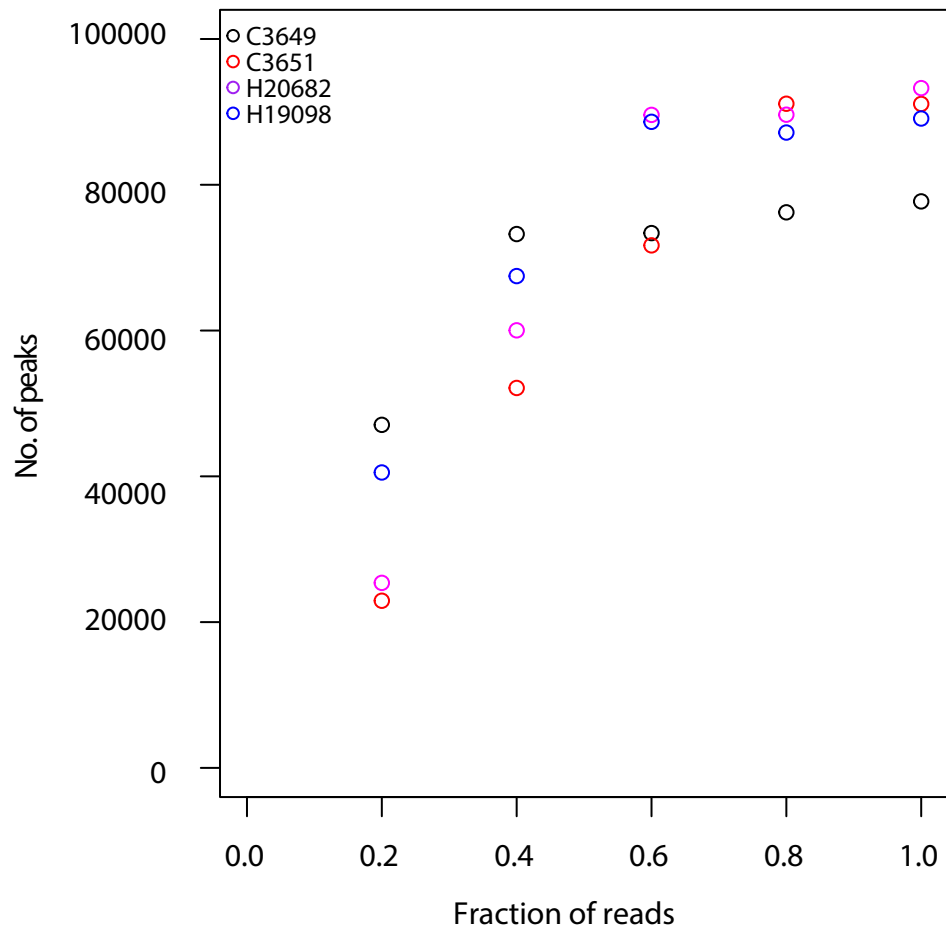
■ FDR0.1  
■ FDR0.05  
■ FDR0.01  
■ FDR0.001  
■ FDR0.0001

H18489 H18511 H19098 H19101 H20682 H20961 H21194 H21792 H28126 H28815 H\_Median C3647 C3649 C3651 C40210 C40280 C4955 C8861 C\_Median

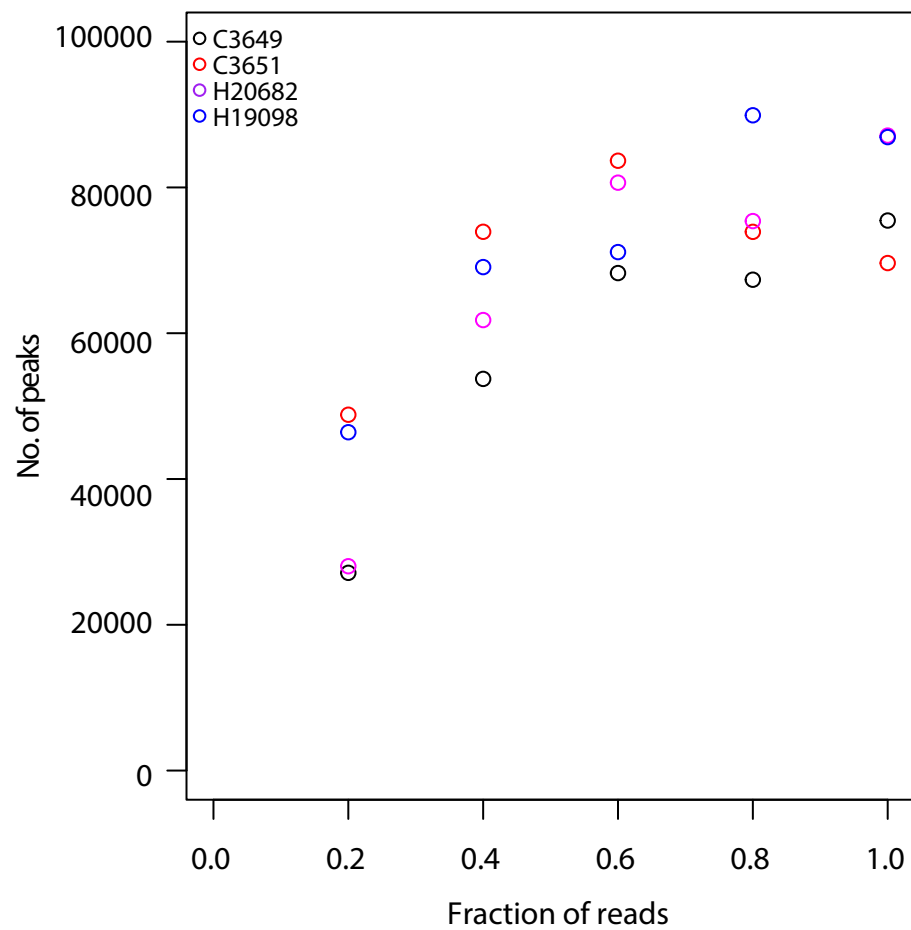
Individual



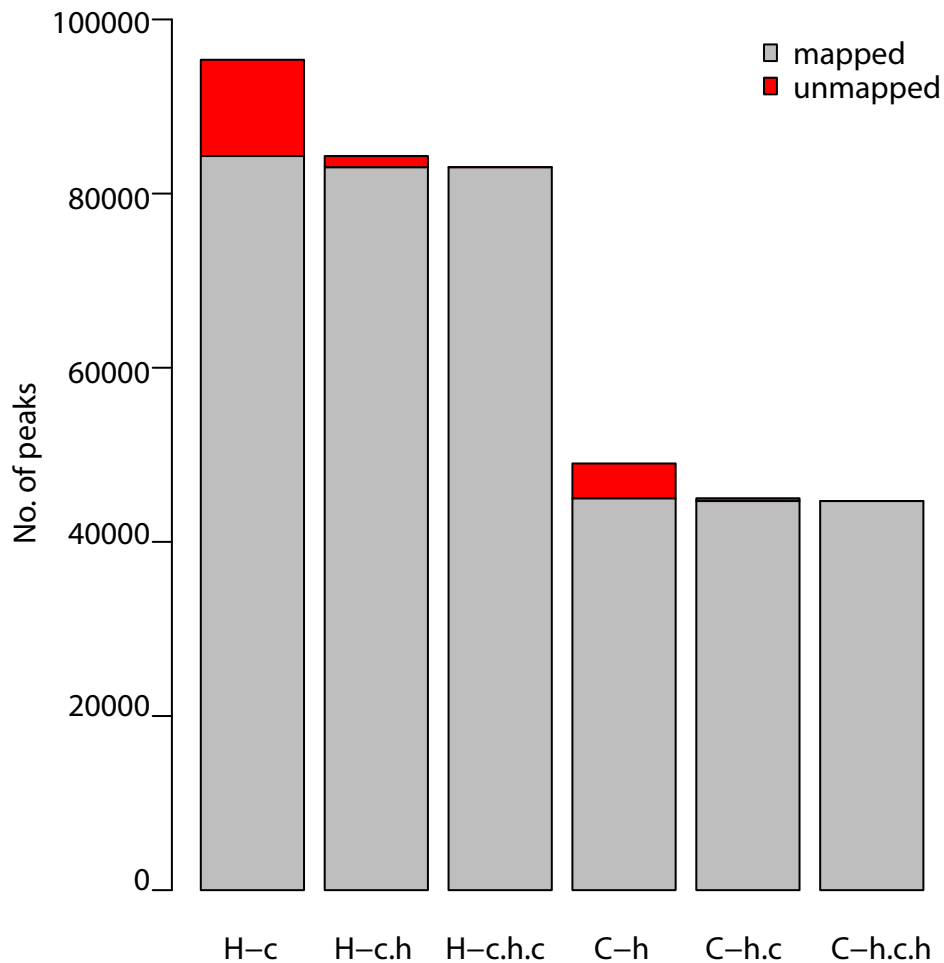
**A)** Sub-sample ChIP reads



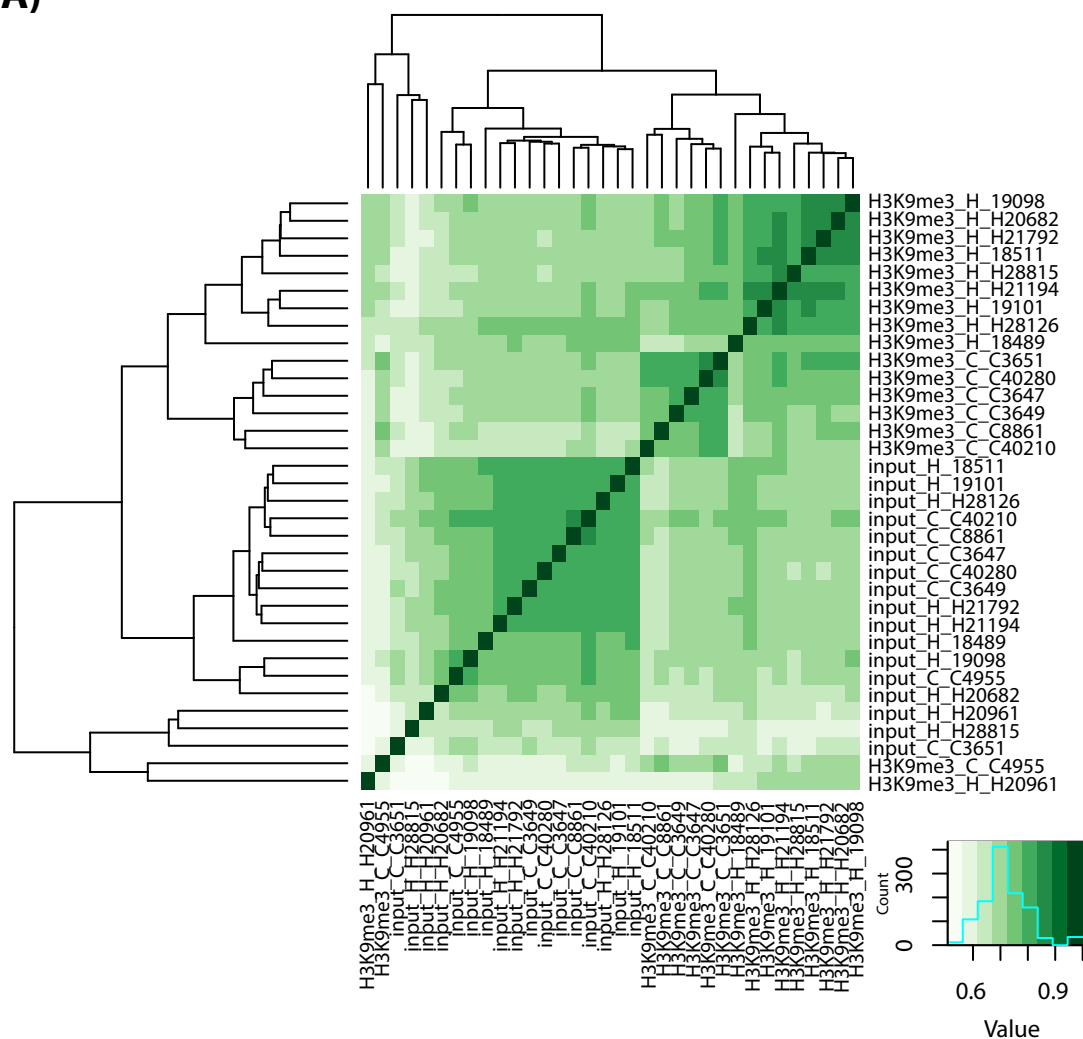
**B)** Sub-sample Input reads



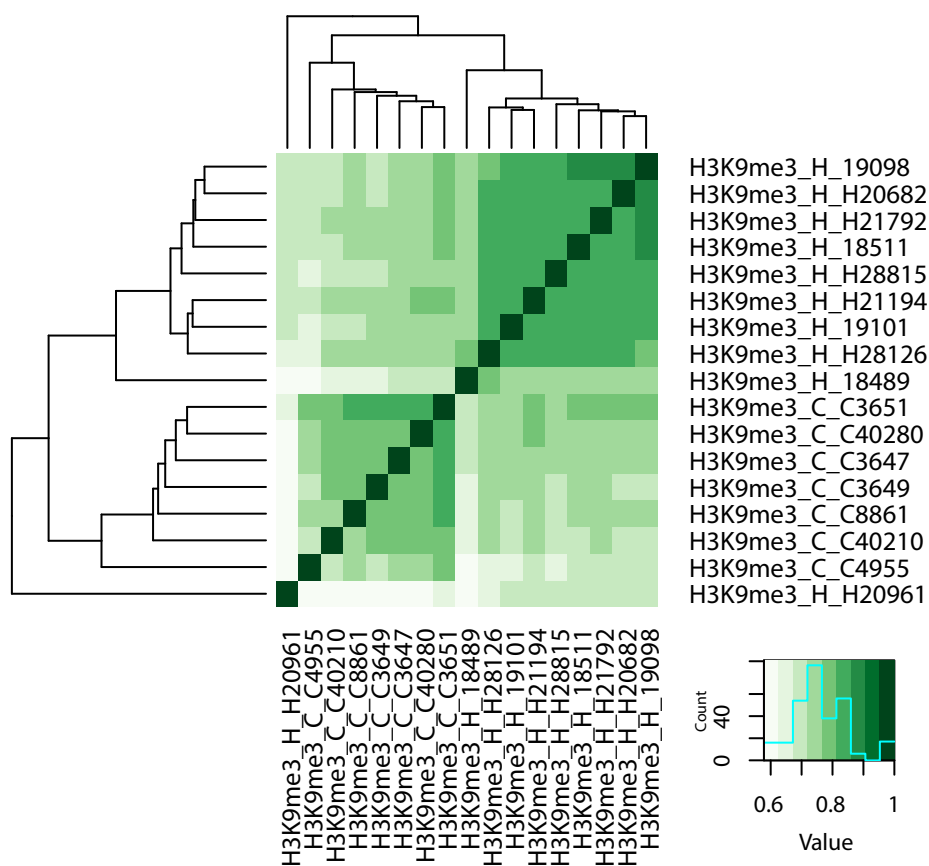




A)

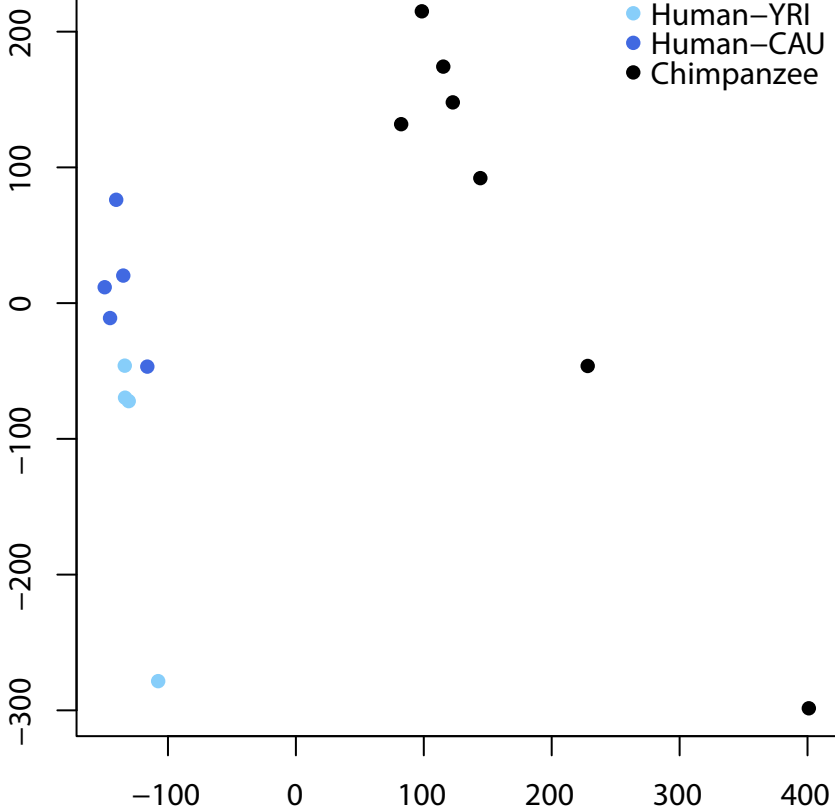


B)



PC2 (13.07% of variance)

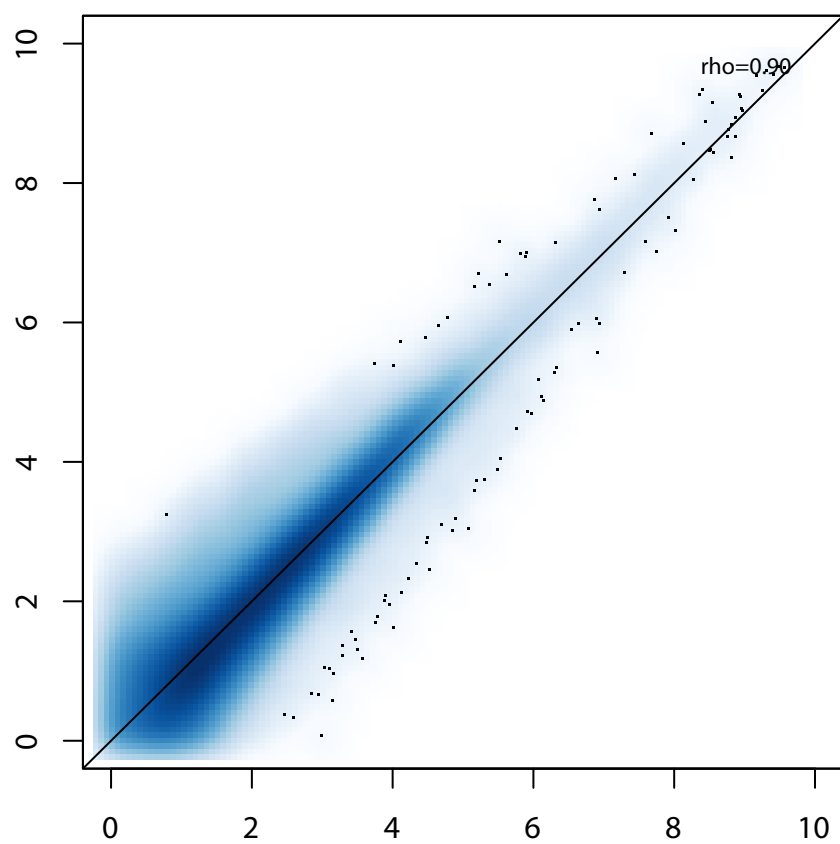
Human-YRI  
Human-CAU  
Chimpanzee



PC1 (18.23% of variance)

**A)**

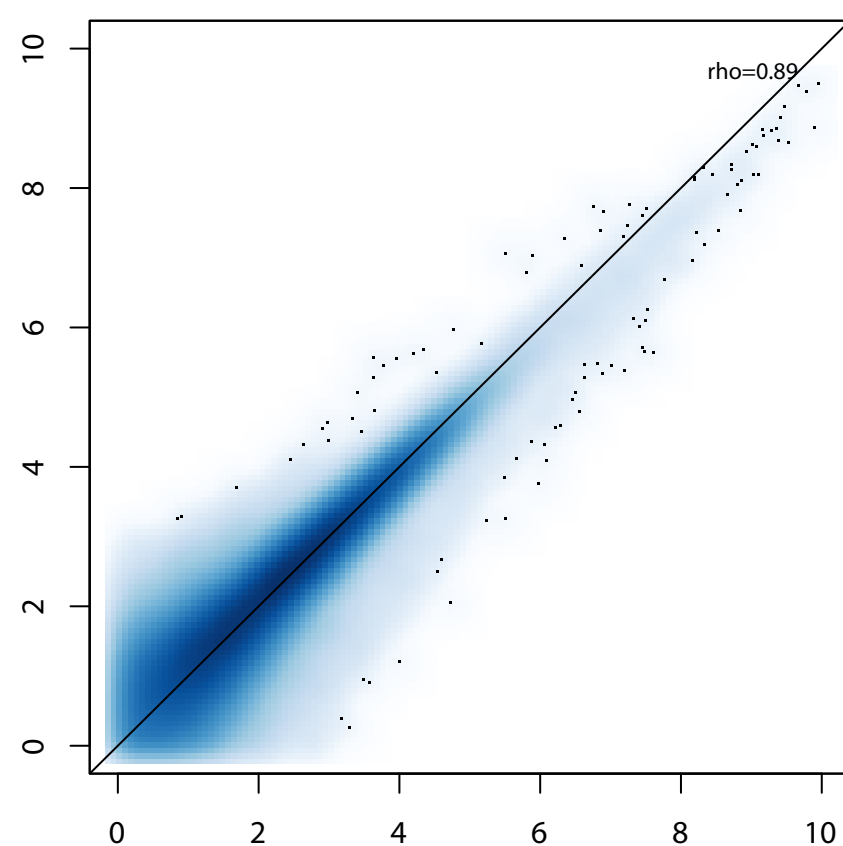
human log count mean (n=4)



human log count mean (n=5)

**B)**

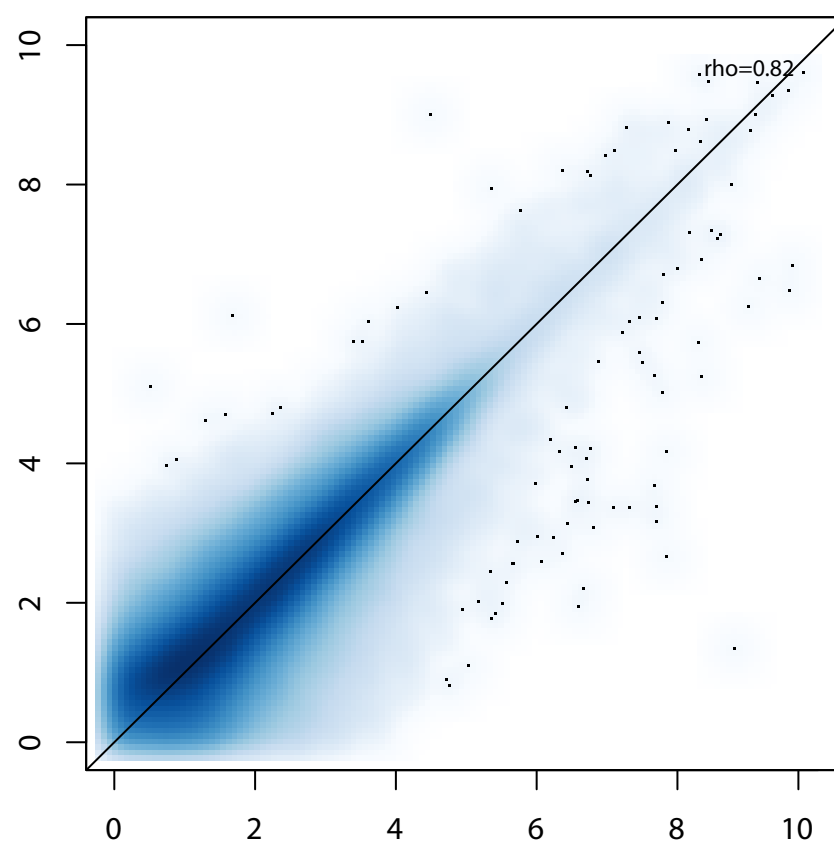
chimp log count mean (n=3)



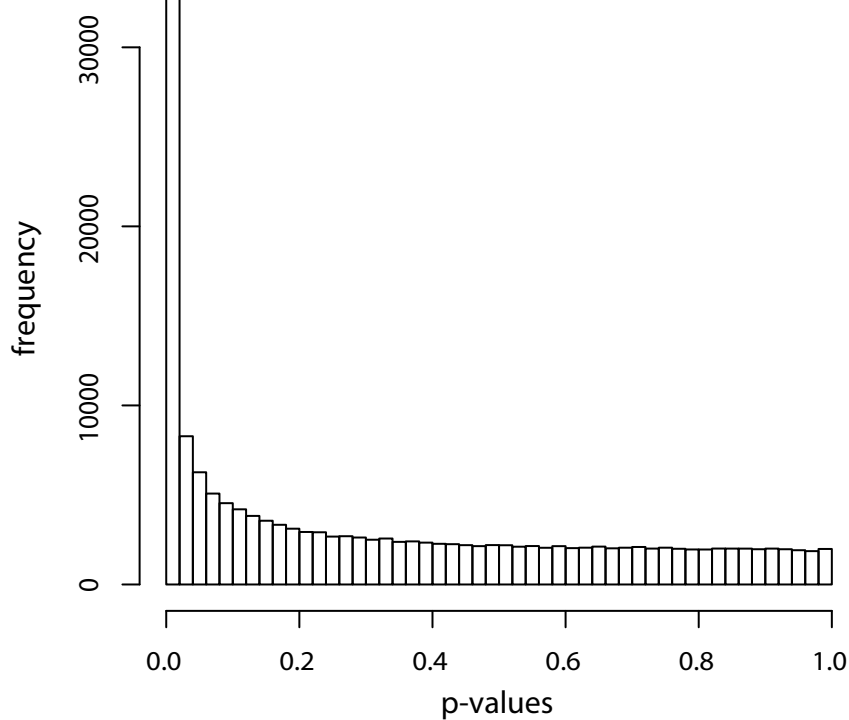
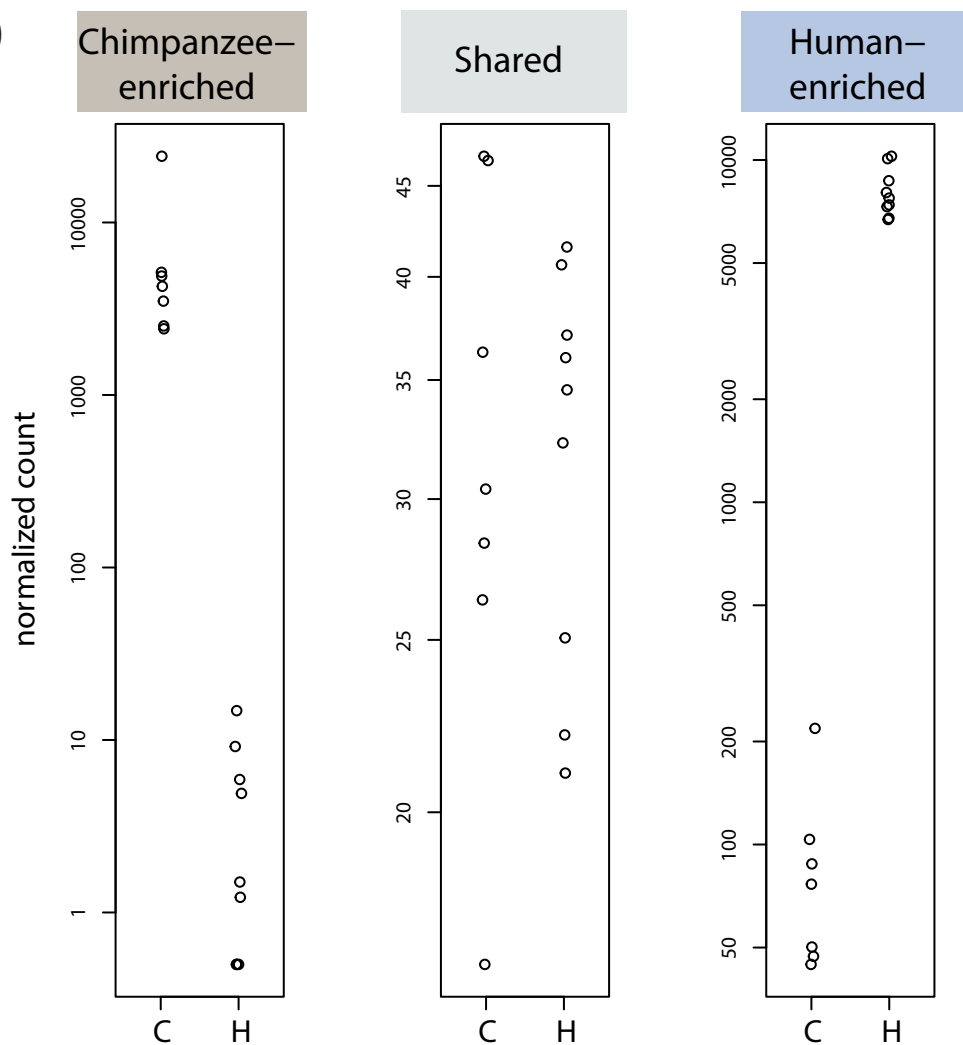
chimp log count mean (n=4)

**C)**

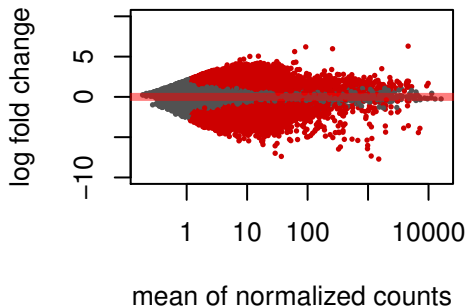
human log count mean (n=9)



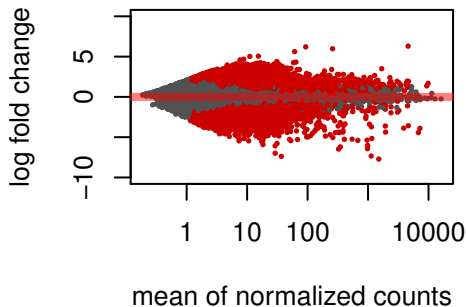
chimp log count mean (n=7)

**A)****B)**

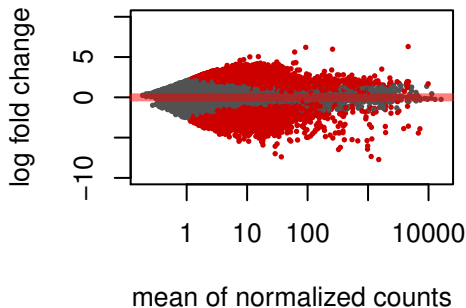
**0.22 DE (FDR=0.1)**



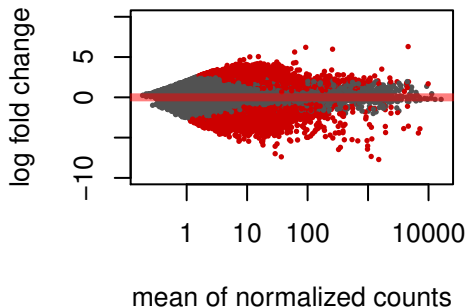
**0.17 DE (FDR=0.05)**

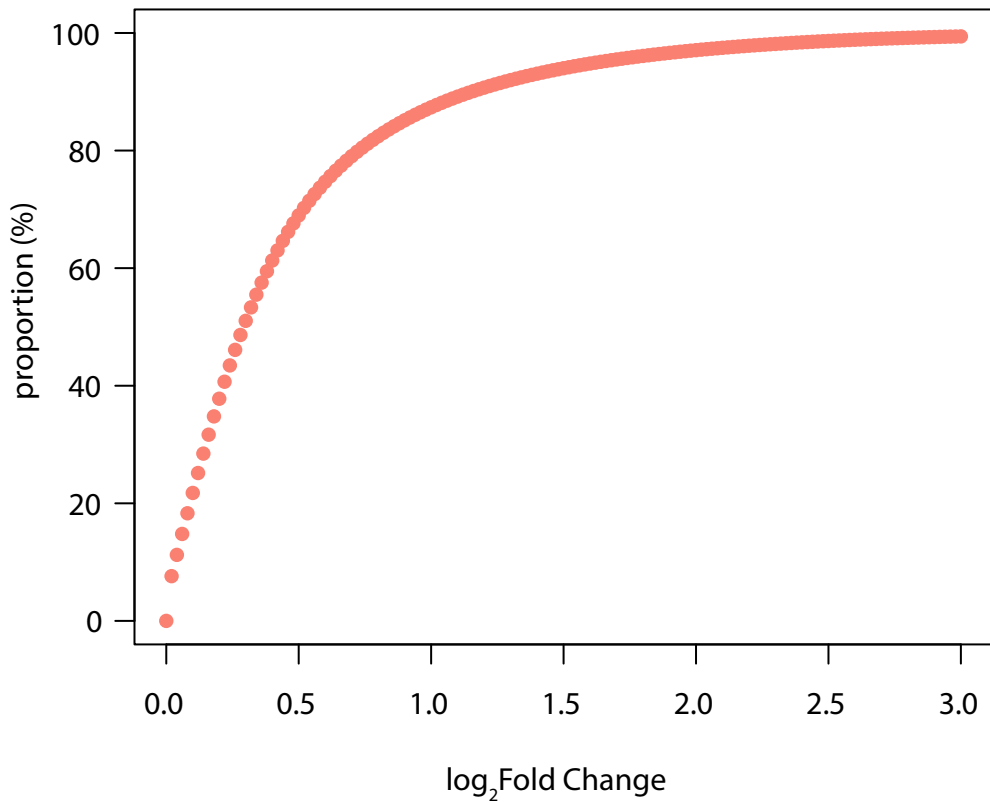


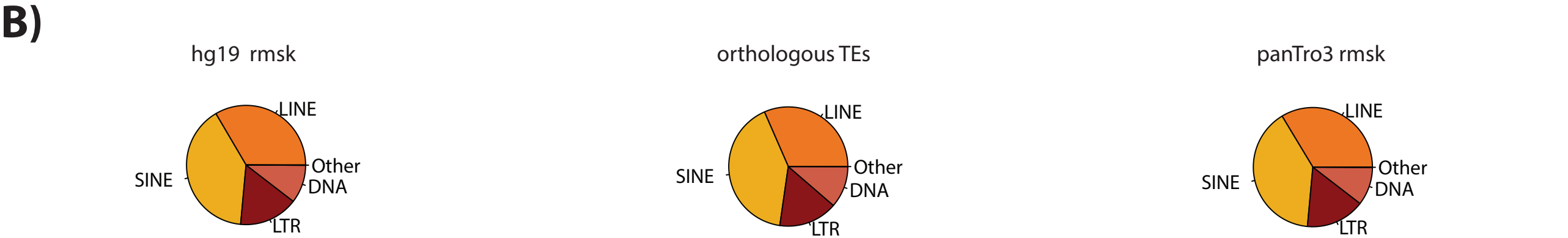
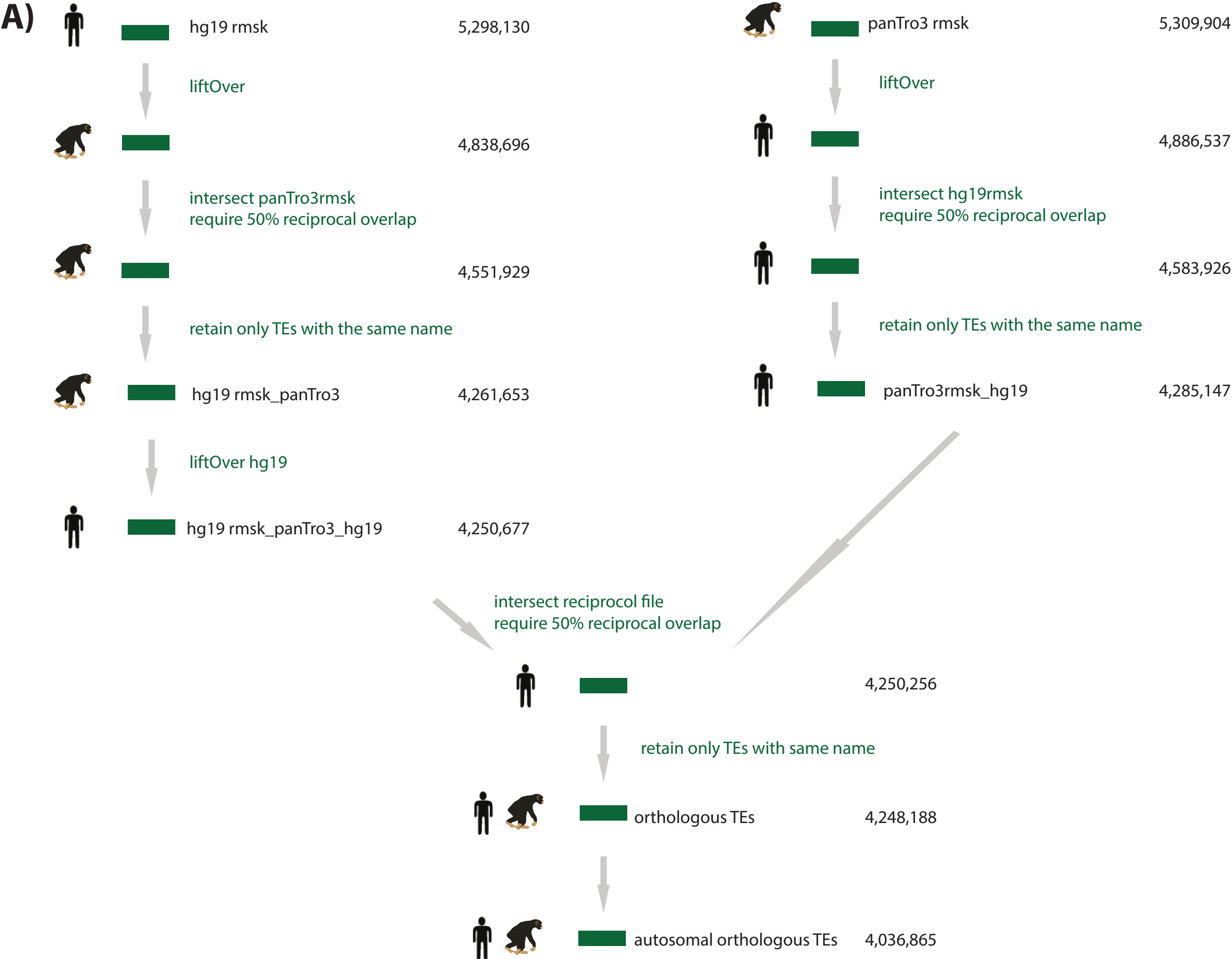
**0.11 DE (FDR=0.01)**



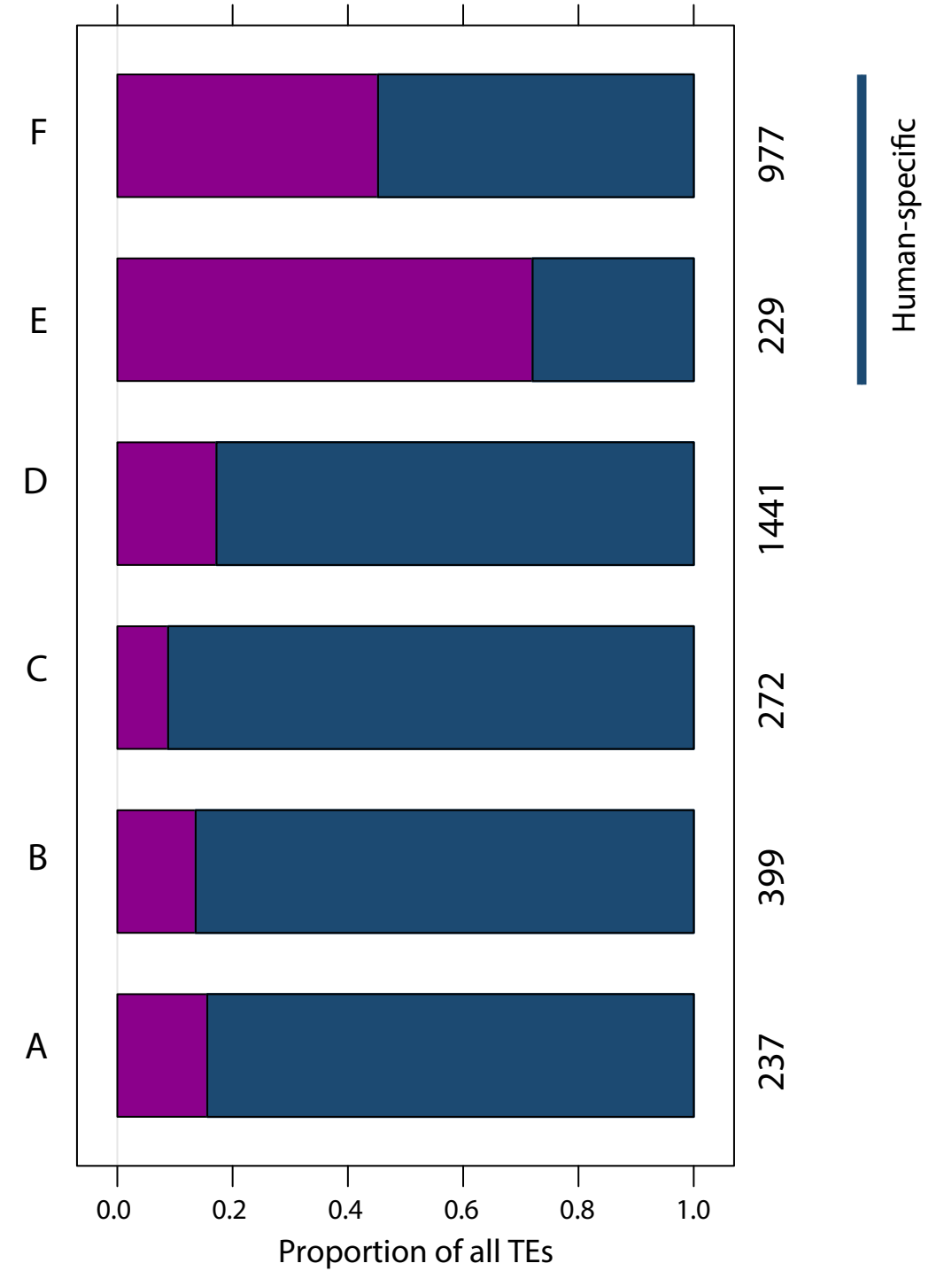
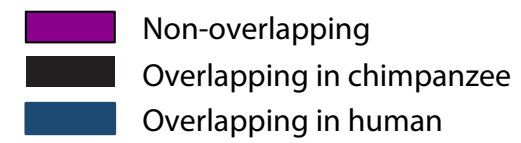
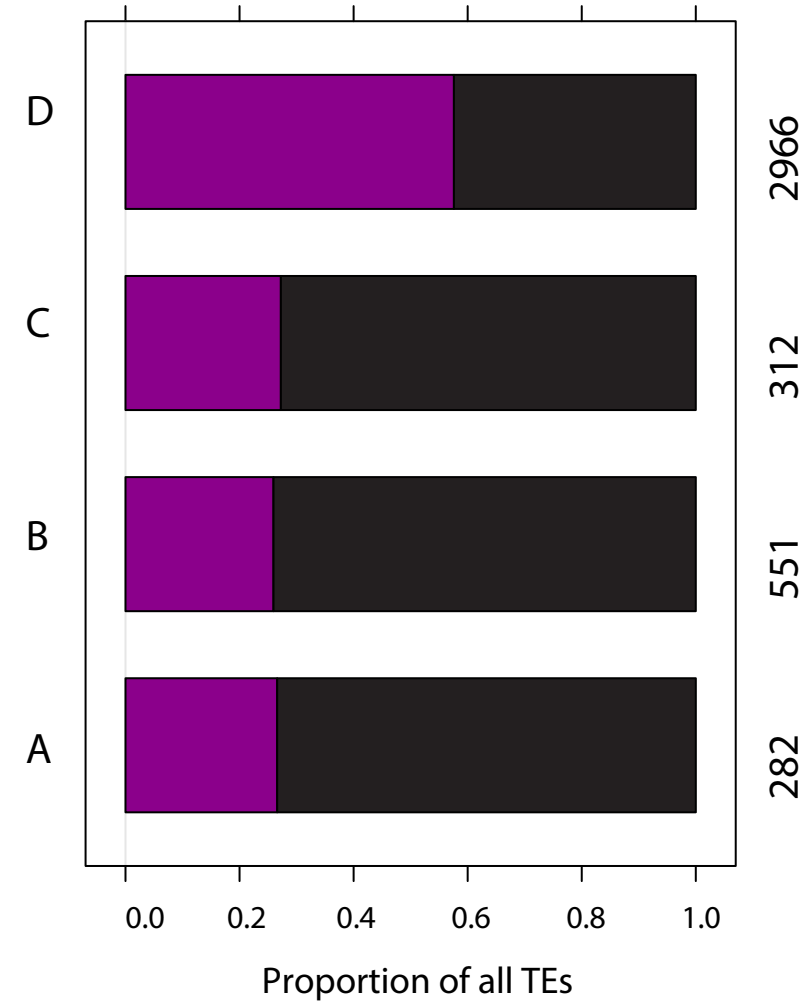
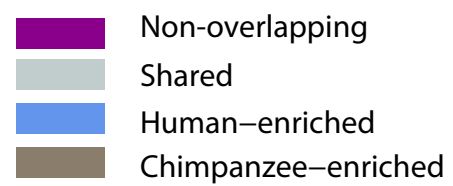
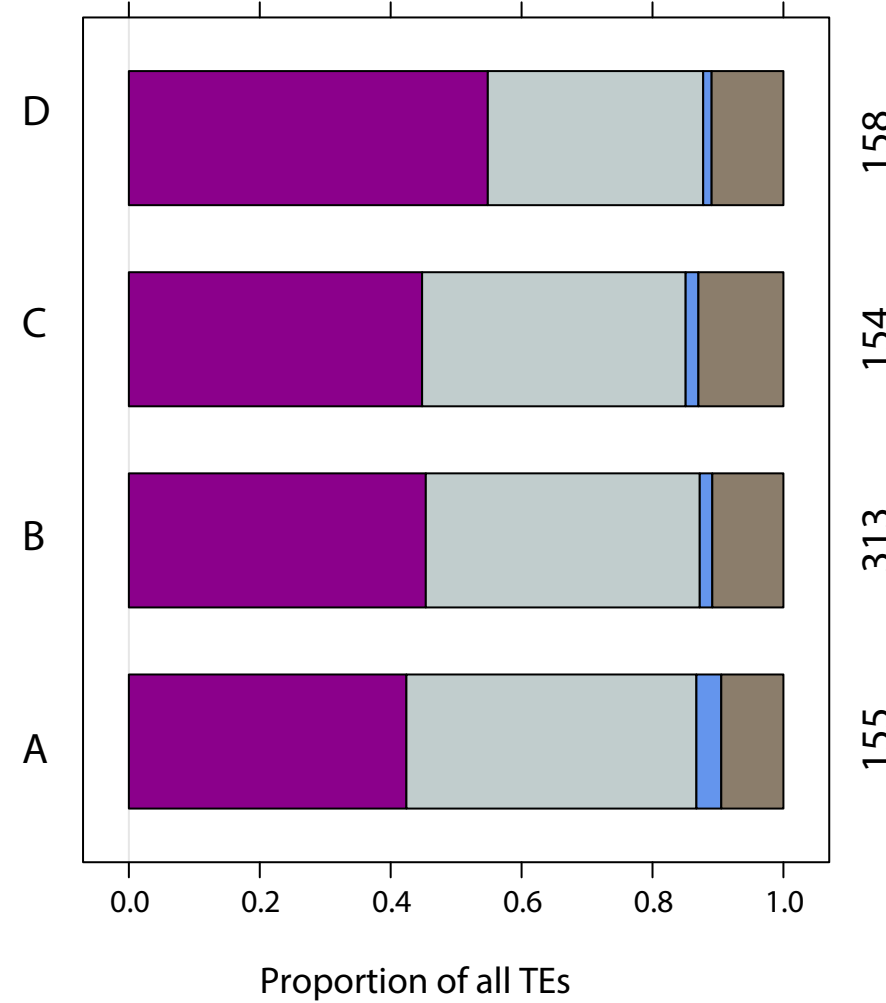
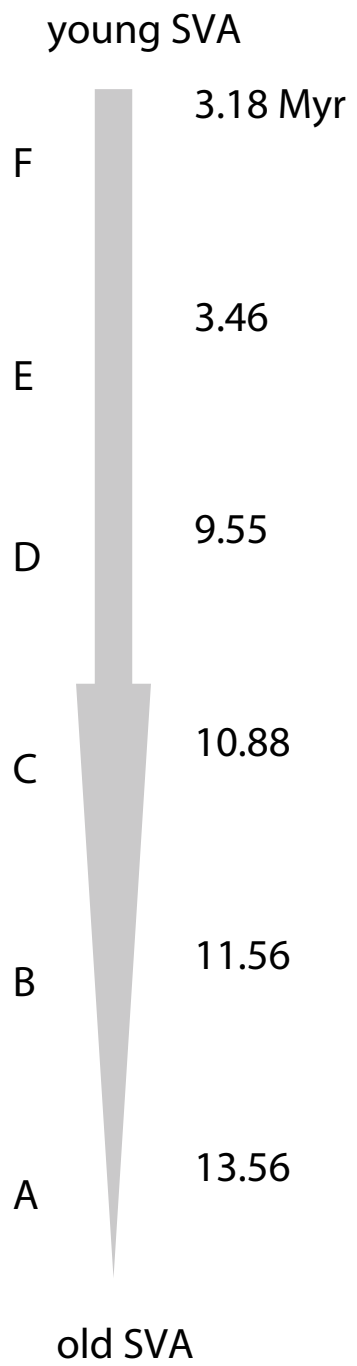
**0.07 DE (FDR=0.001)**





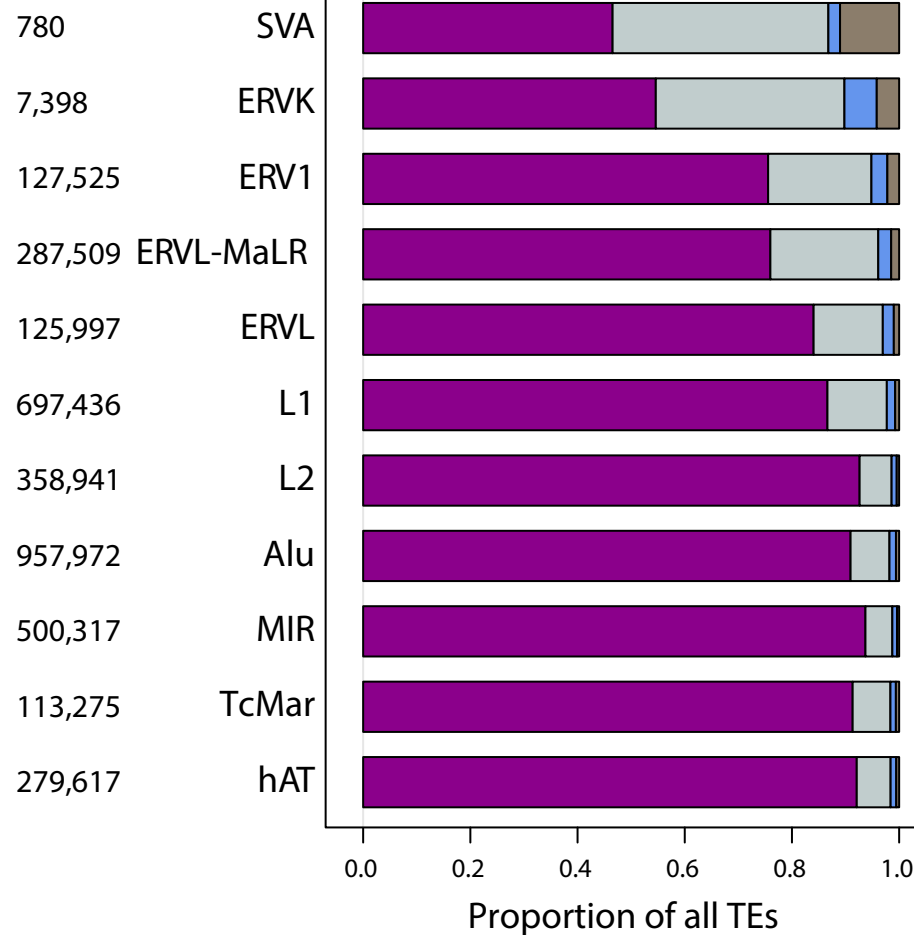
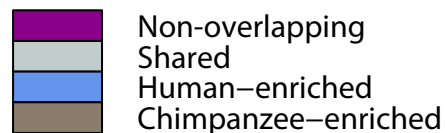
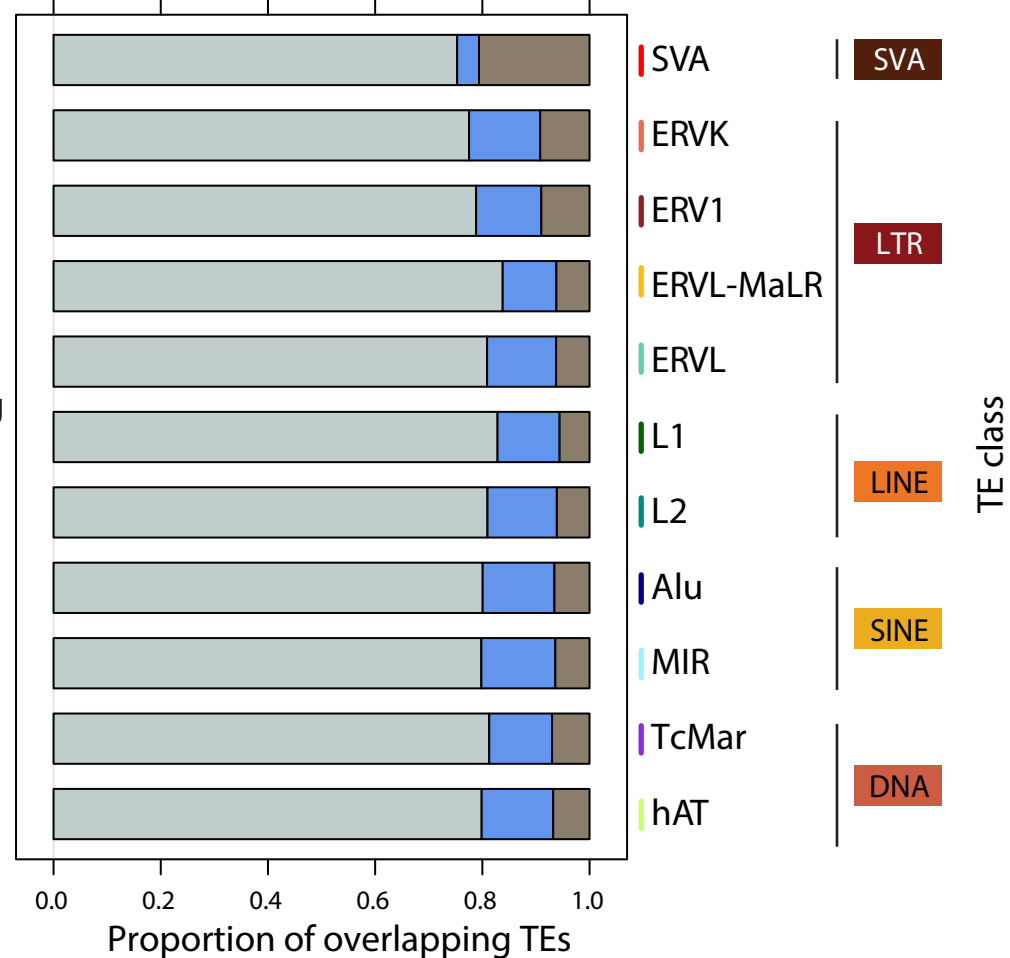


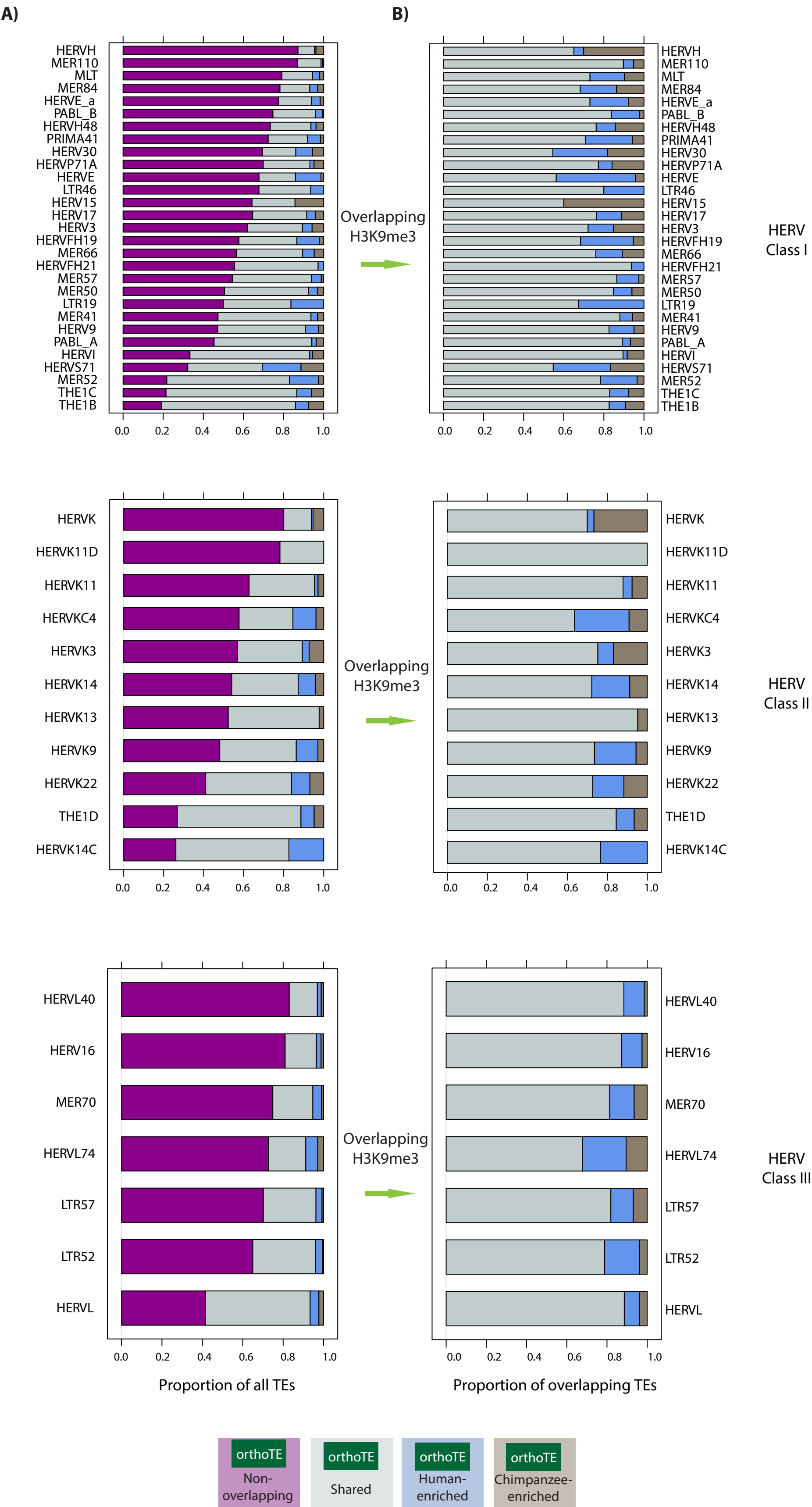




**A)**

TE family

Overlapping  
H3K9me3**B)**



HERVL40

HERV16

MER70

HERVL74

LTR57

LTR52

HERVL

0.0

0.2

0.4

0.6

0.8

1.0

Overlapping  
H3K9me3

HERVL40

HERV16

MER70

HERVL74

LTR57

LTR52

HERVL

0.0

0.2

0.4

0.6

0.8

1.0

HERV  
Class III

Proportion of all TEs

Proportion of overlapping TEs

orthoTE

Non-overlapping

orthoTE

Shared

orthoTE

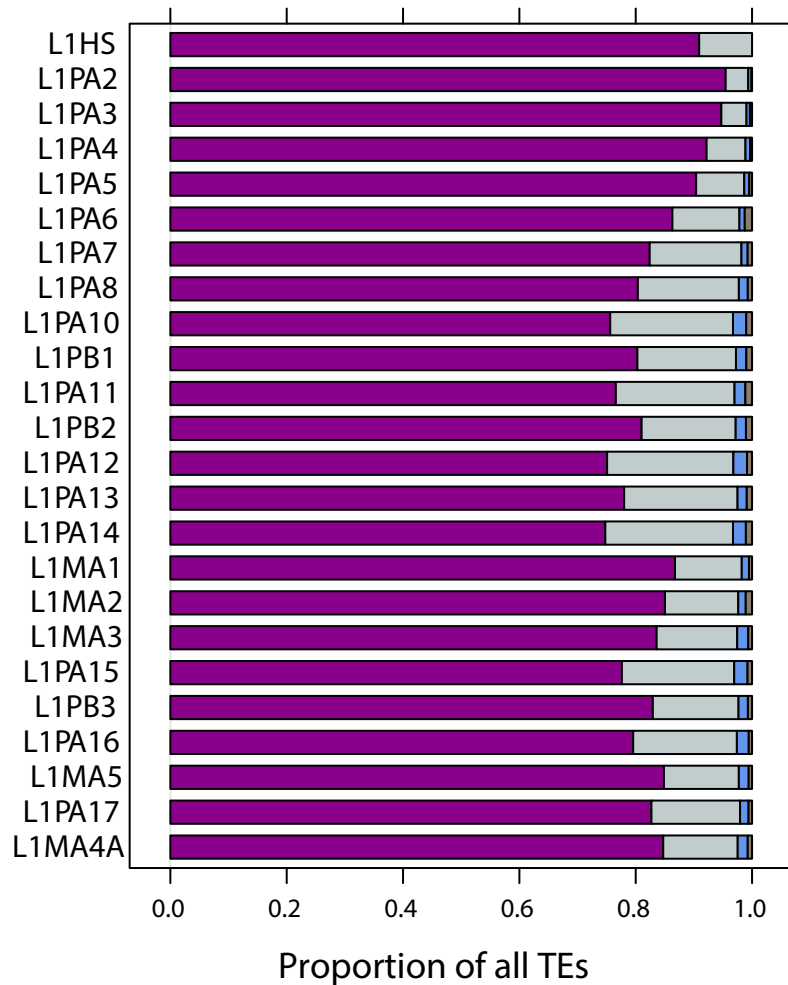
Human-enriched

orthoTE

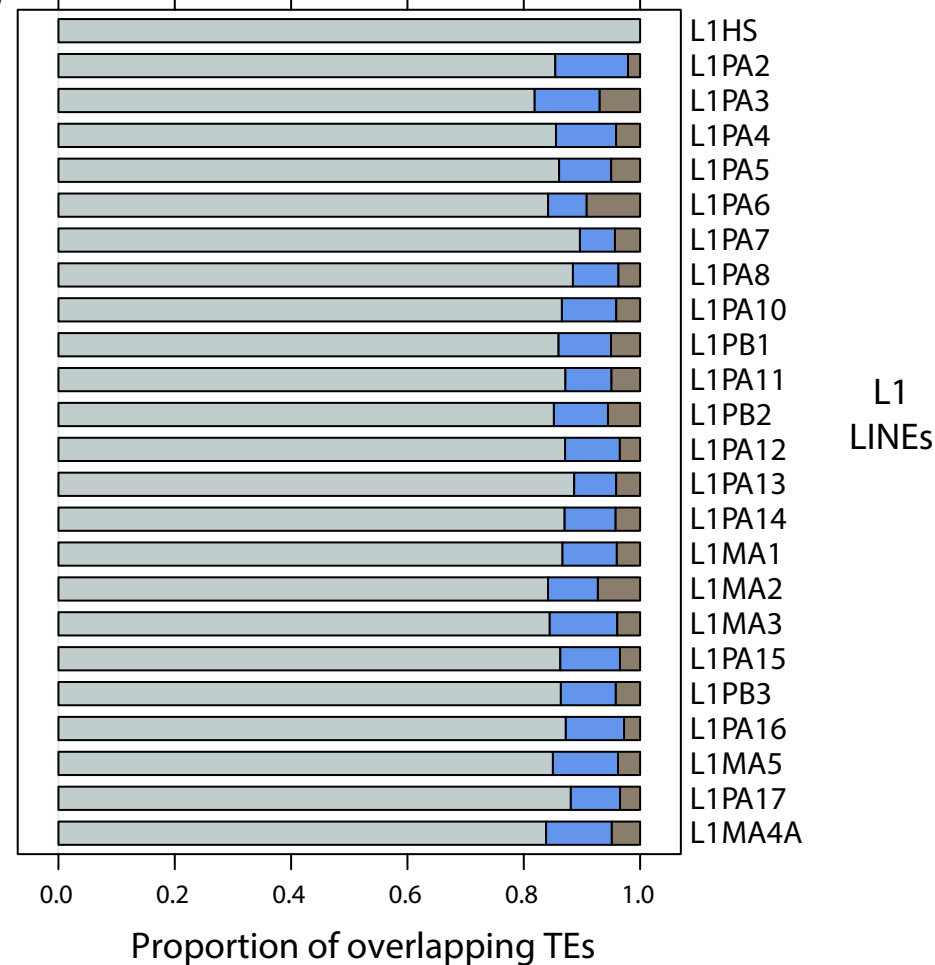
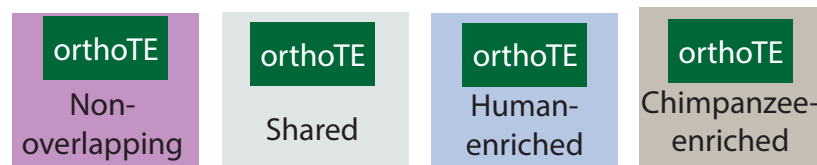
Chimpanzee-enriched

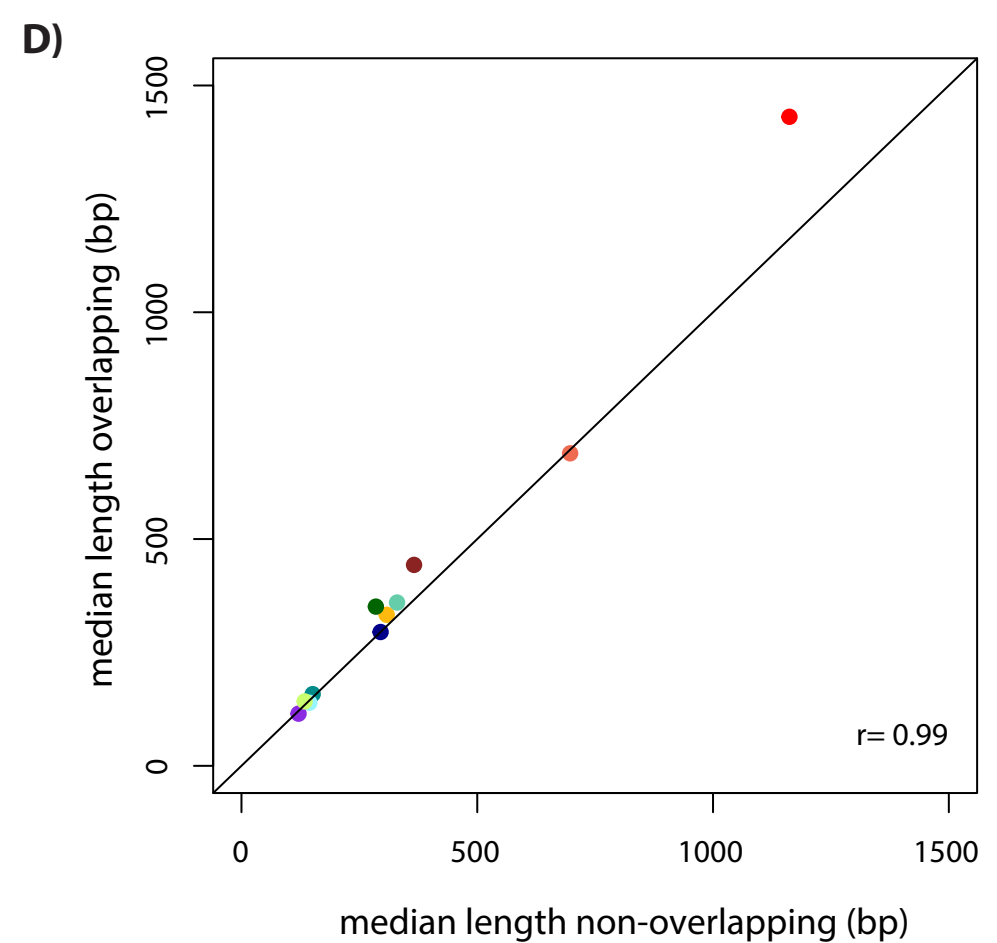
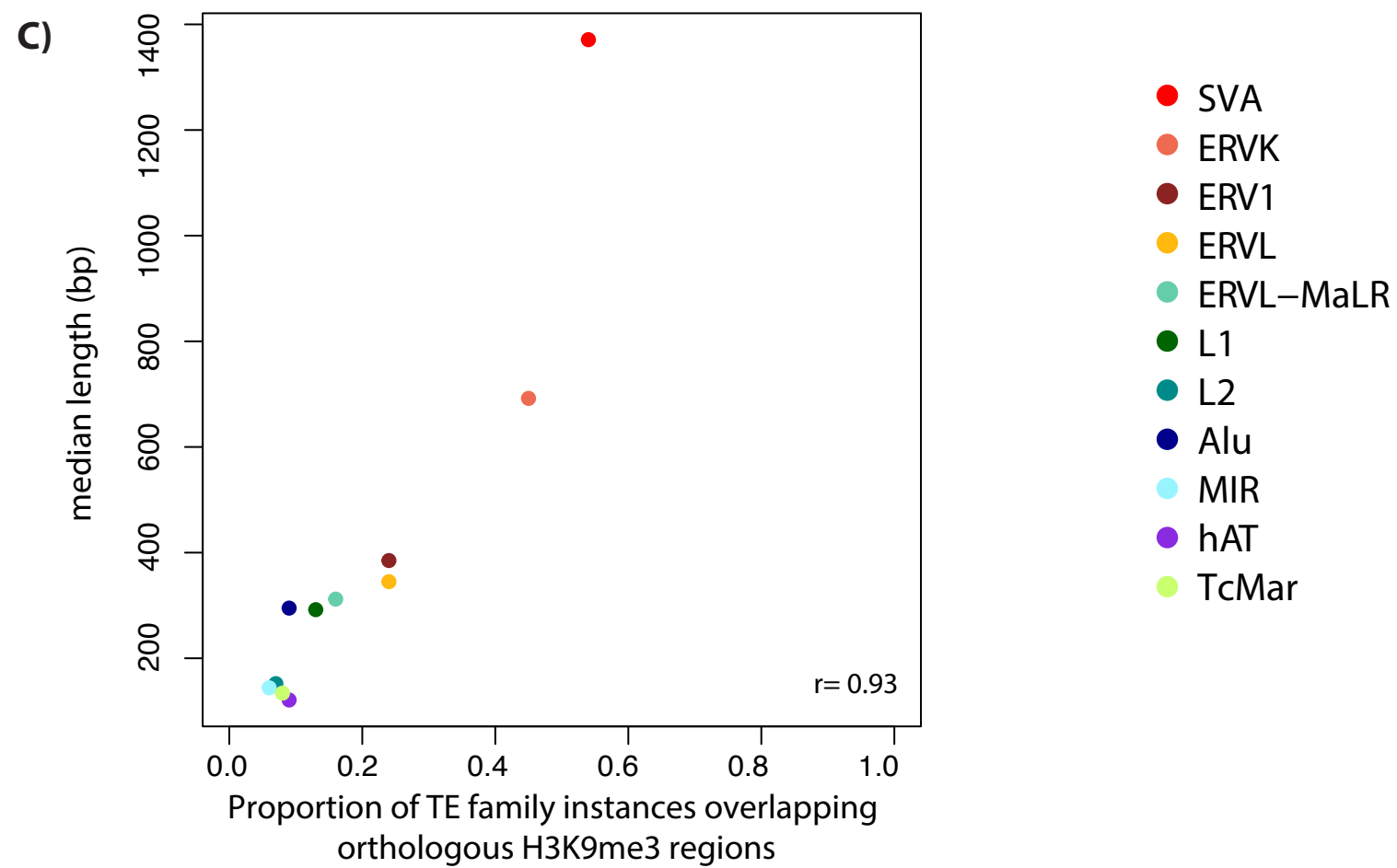
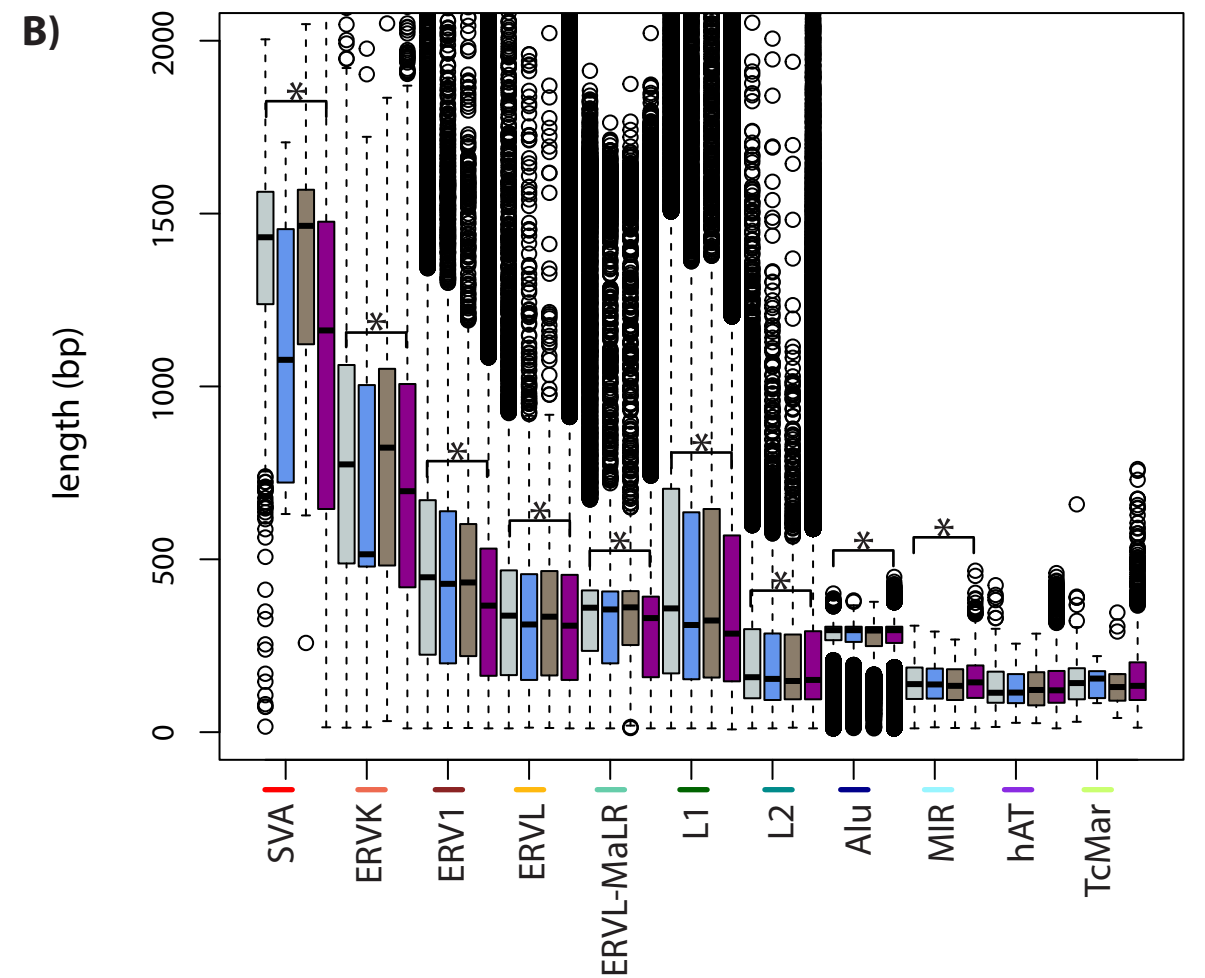
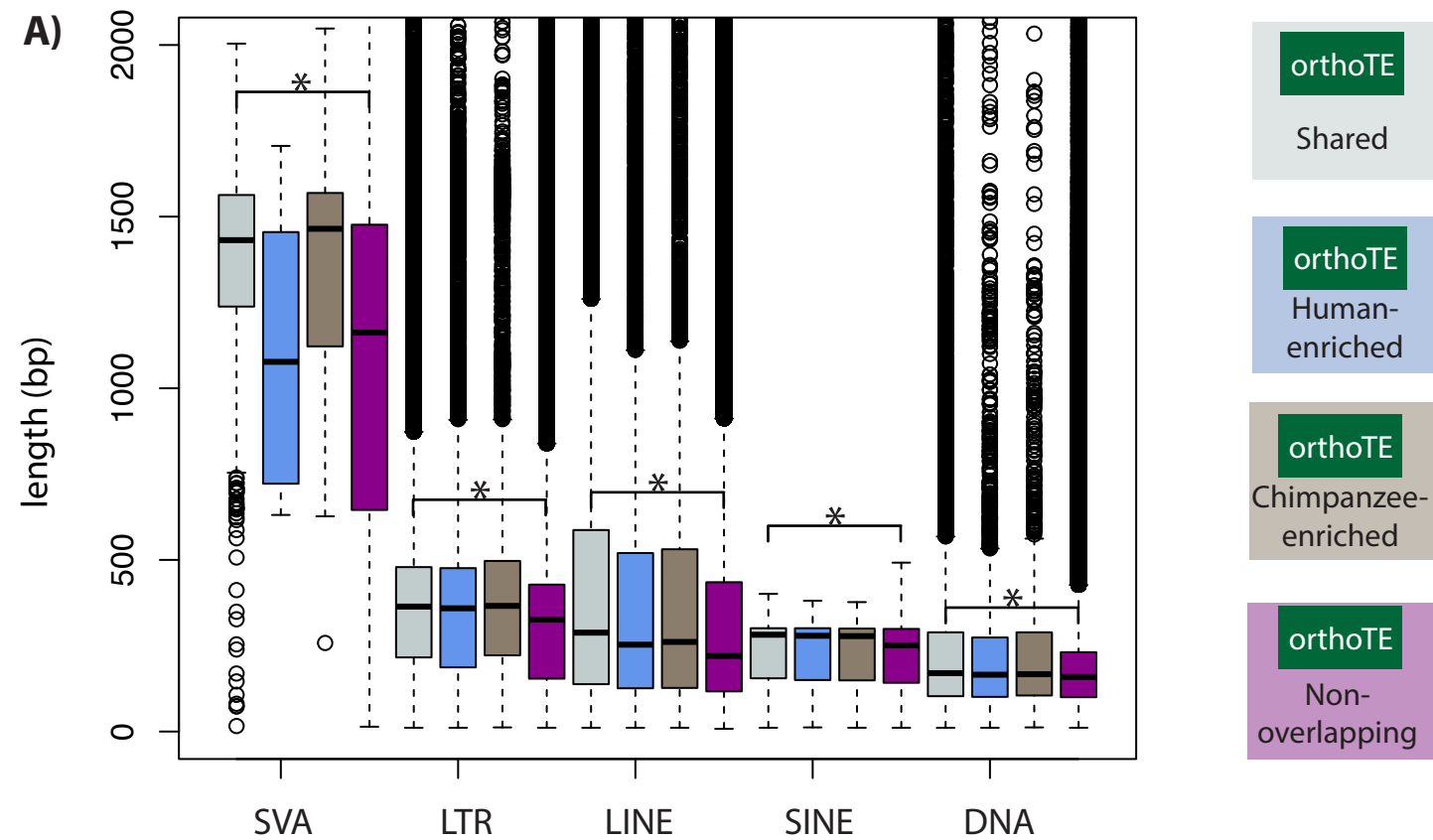
A)

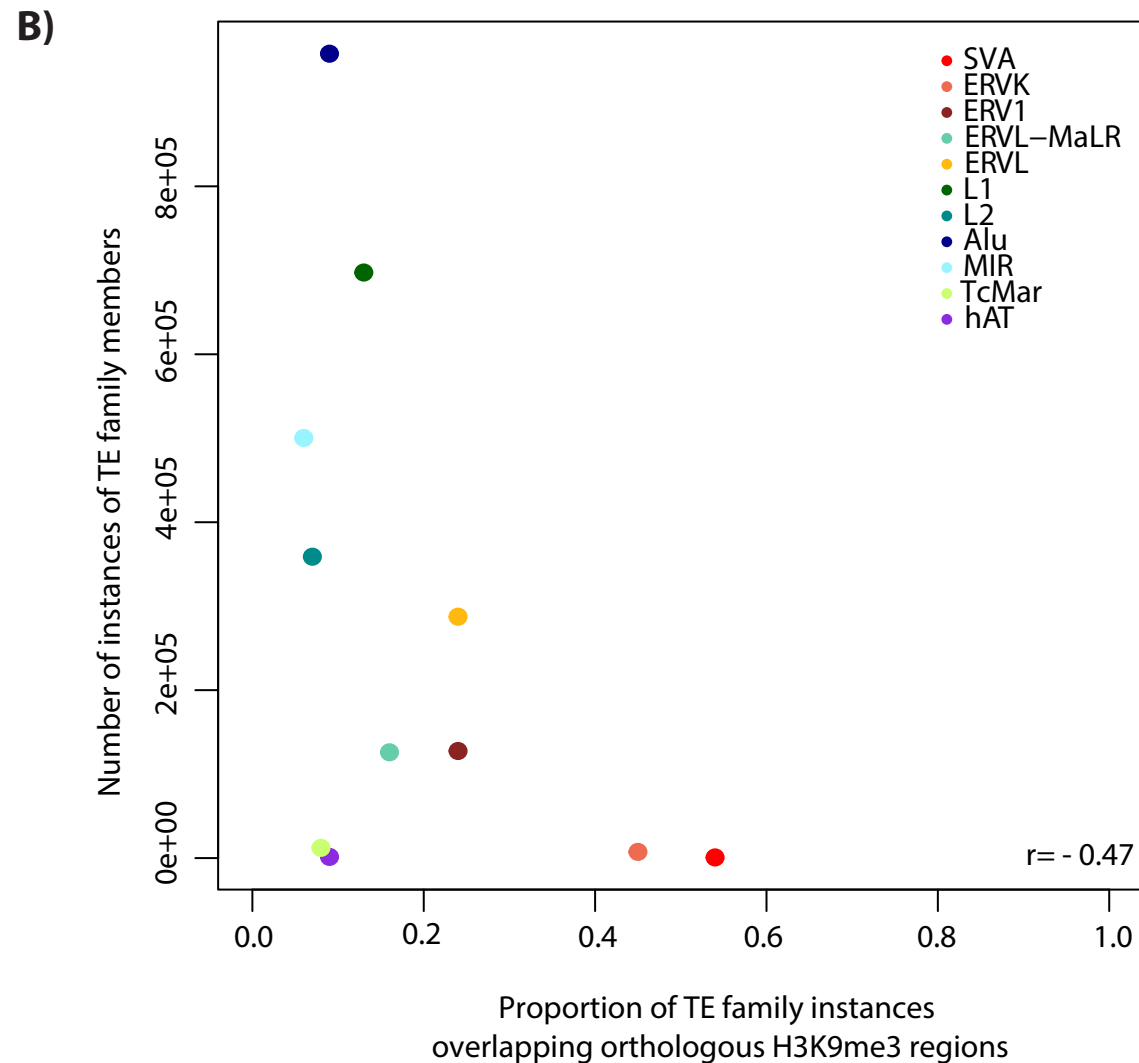
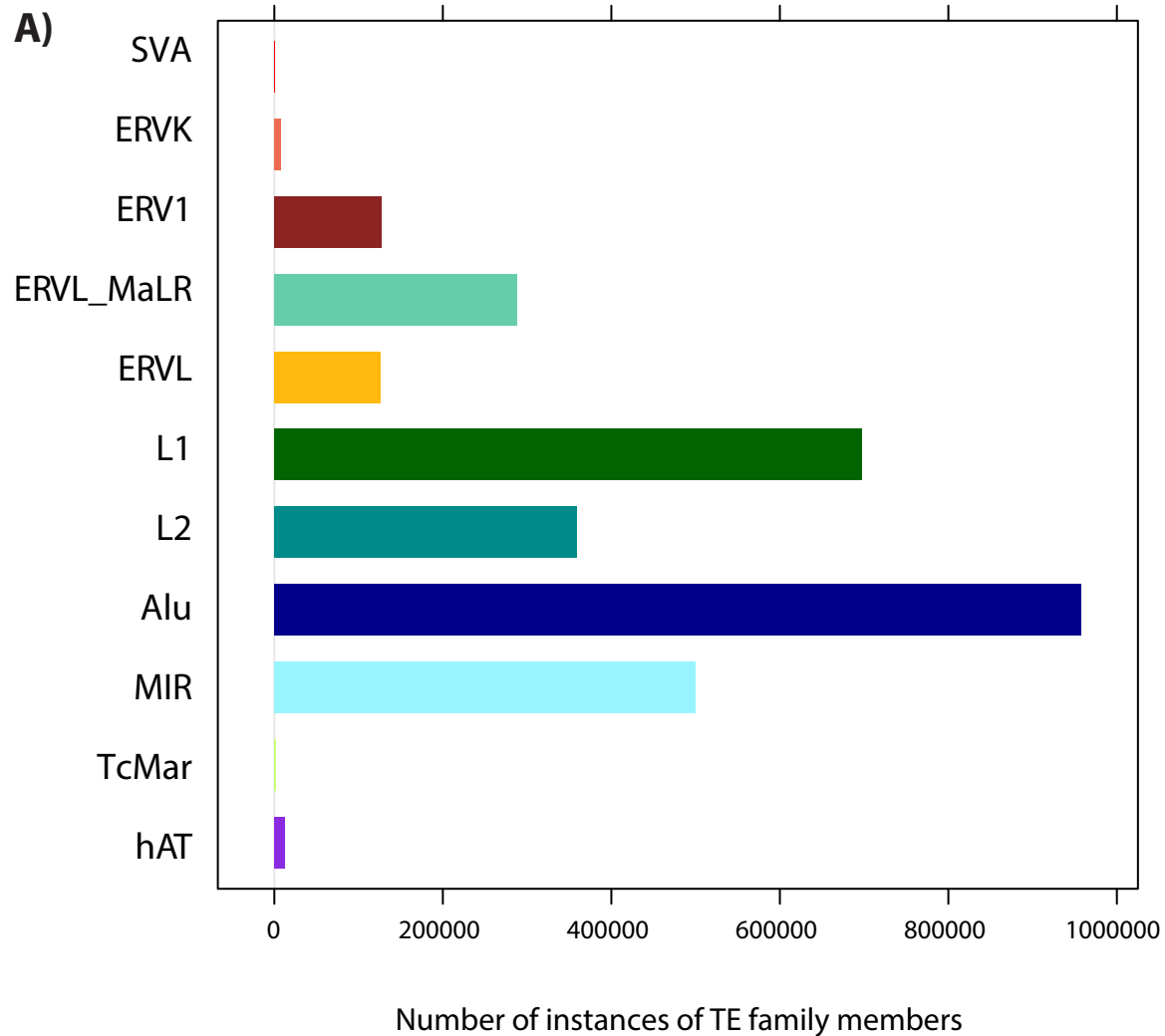
Increasing evolutionary age

Overlapping  
H3K9me3

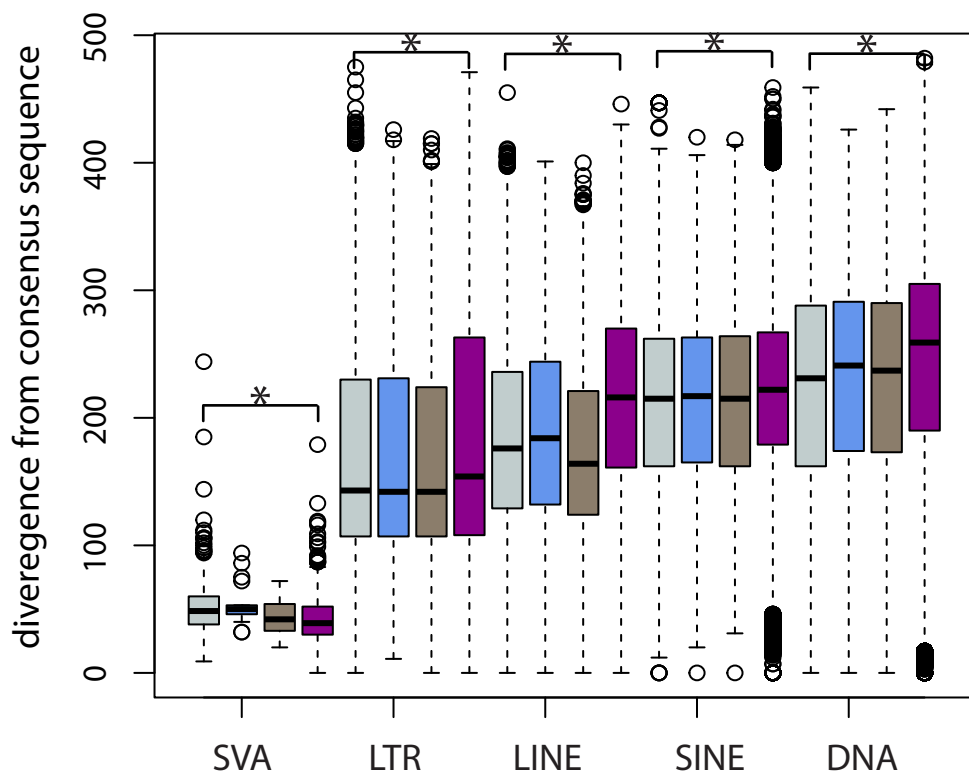
B)

L1  
LINEs

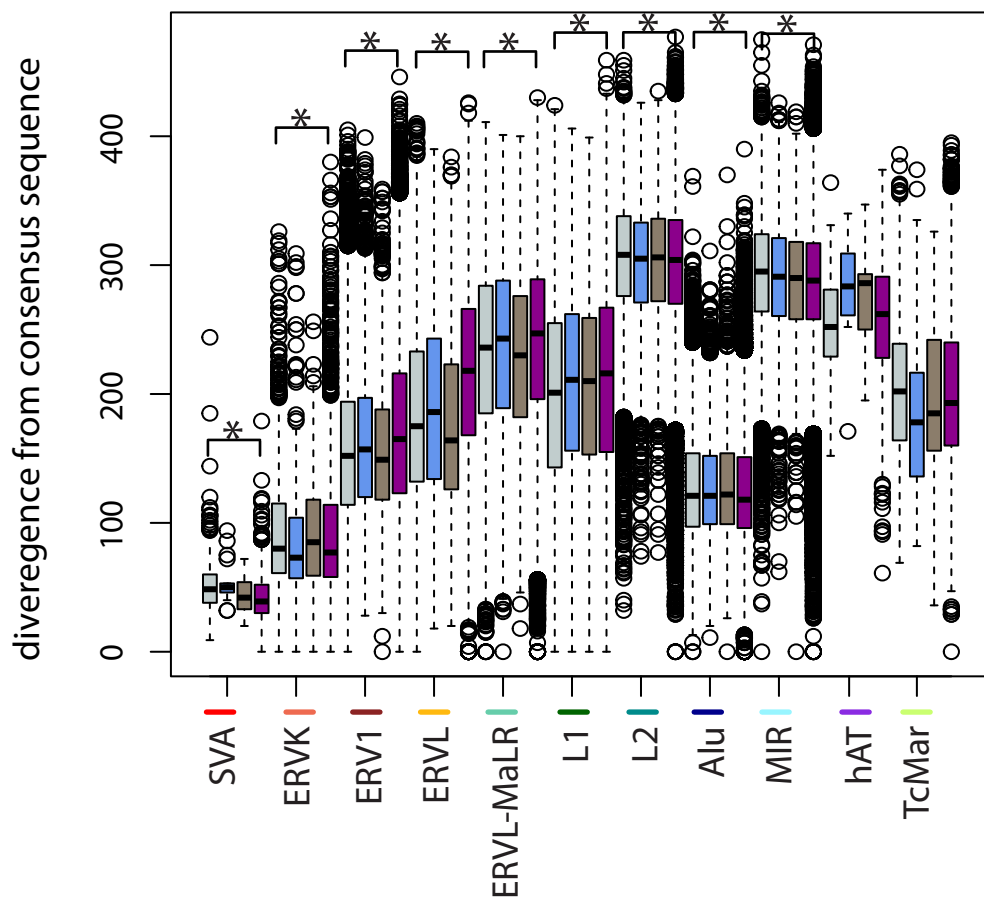




A)



B)



orthoTE

Shared

orthoTE

Human-enriched

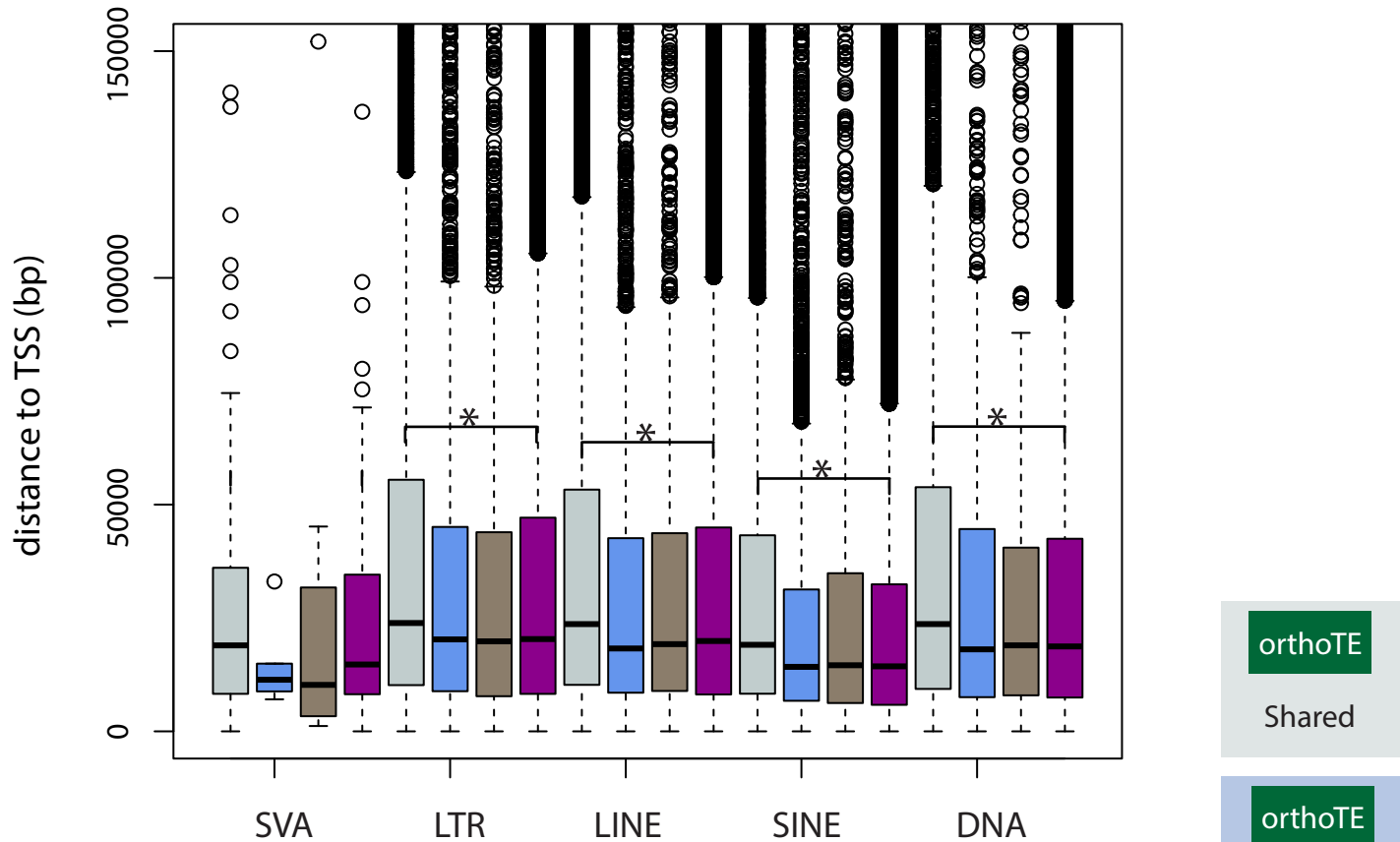
orthoTE

Chimpanzee-enriched

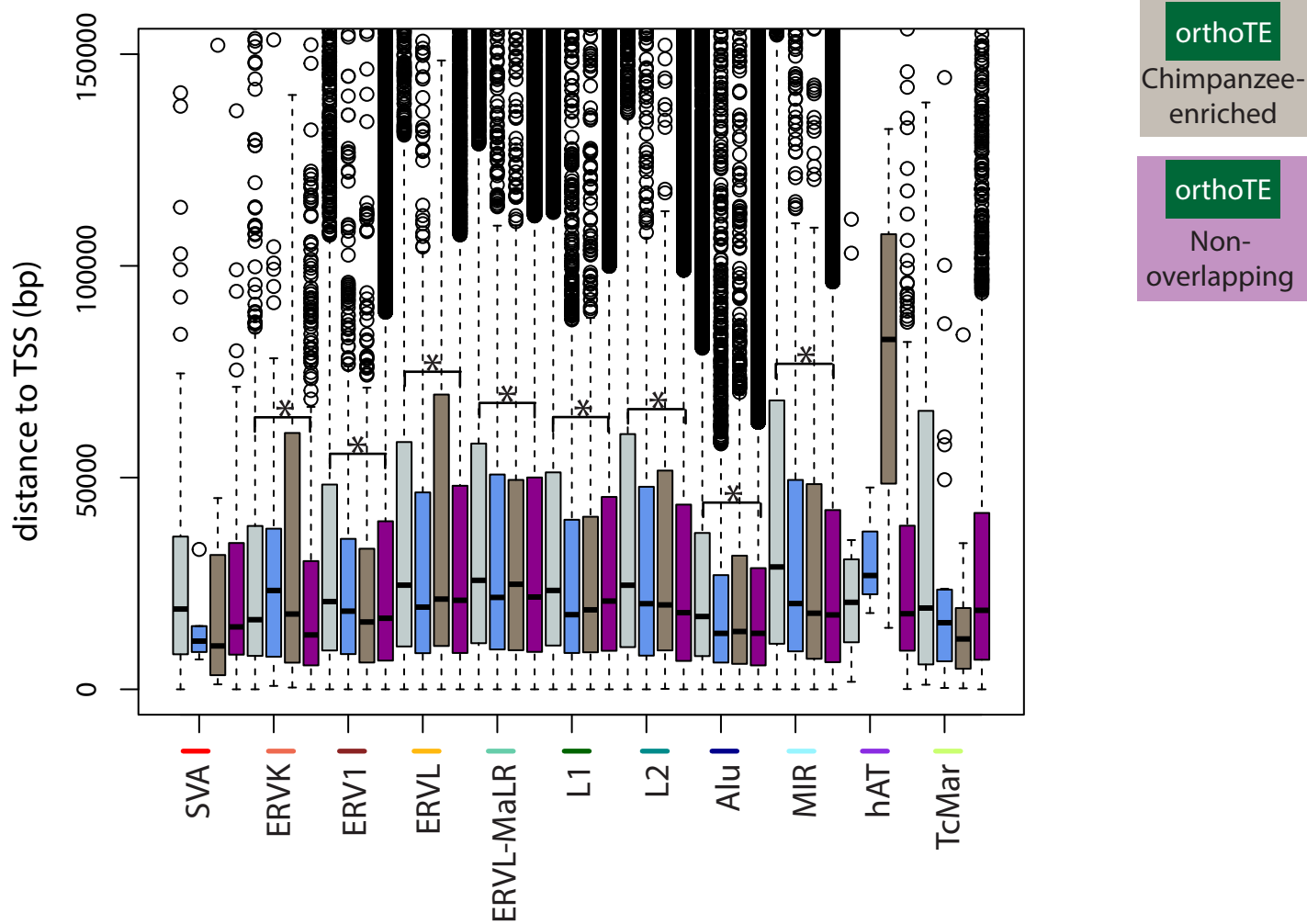
orthoTE

Non-overlapping

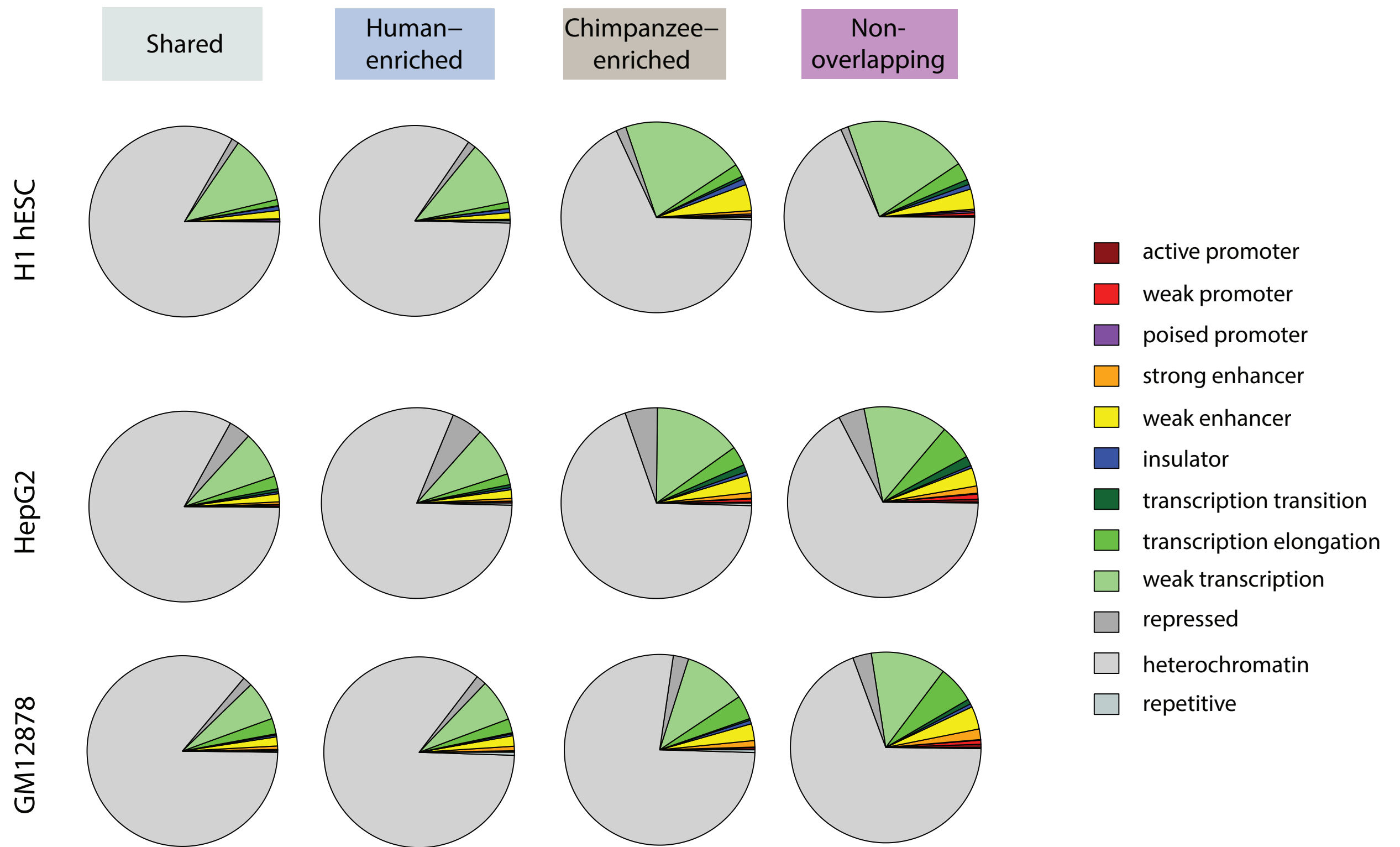
A)



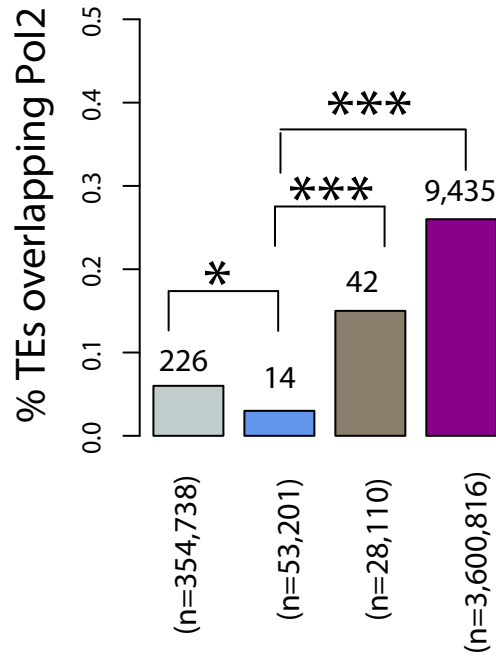
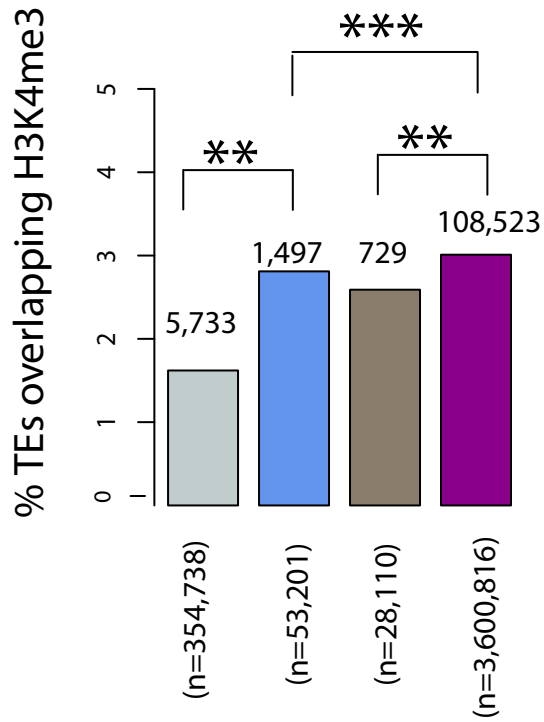
B)



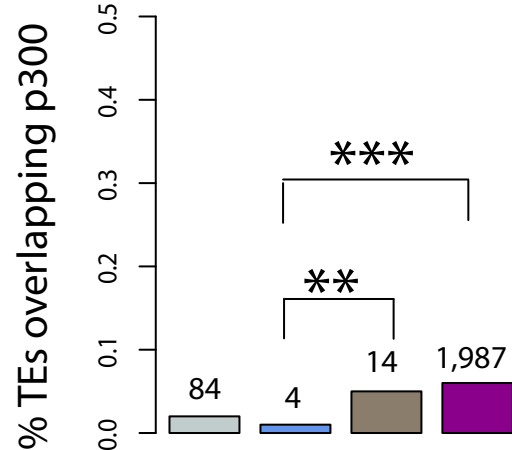
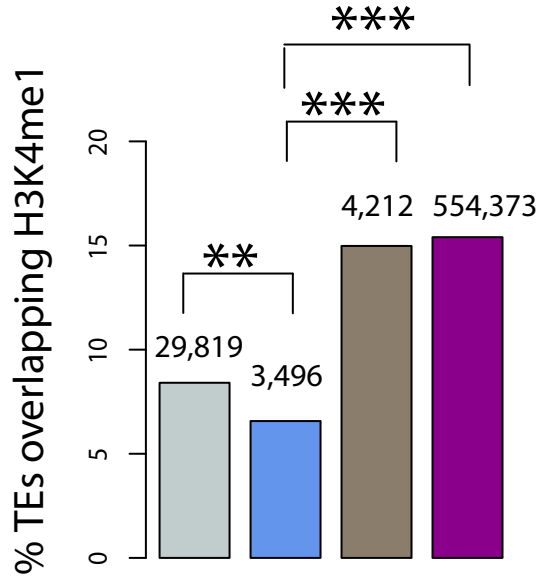




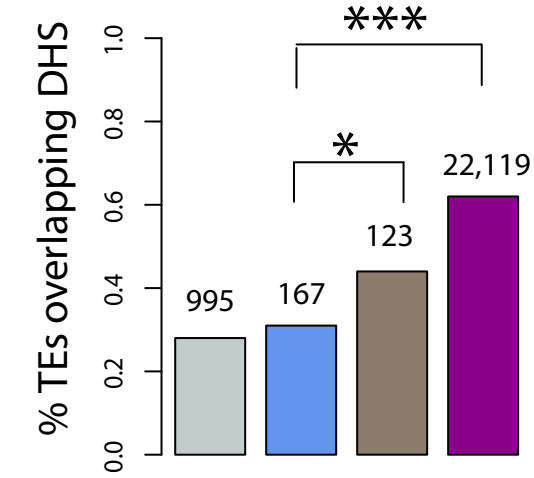
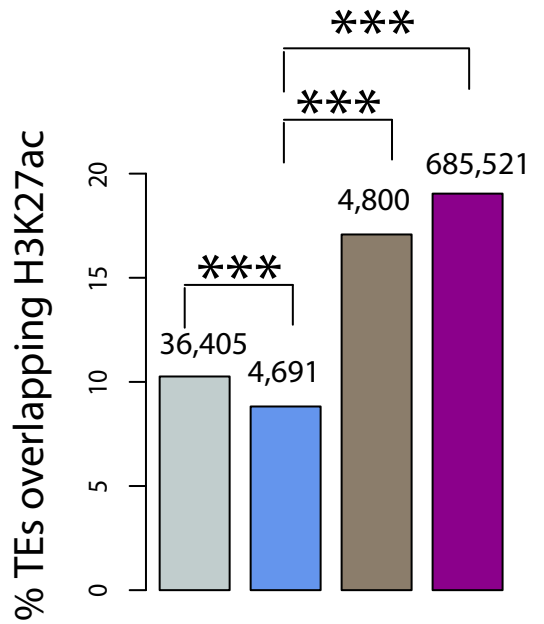
## Promoter



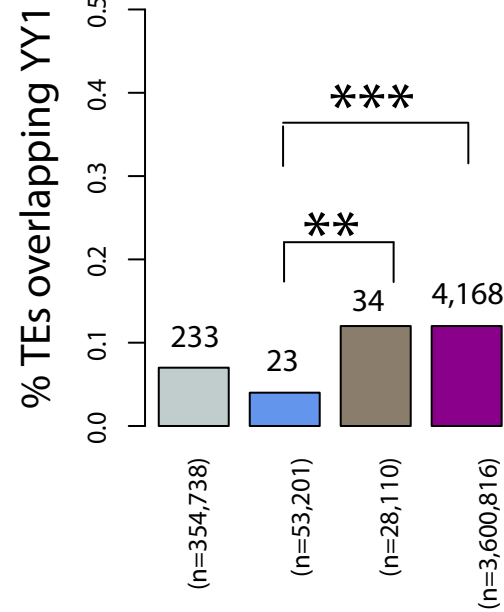
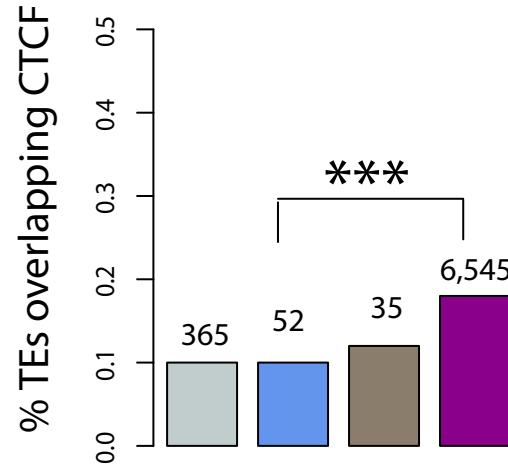
## Enhancer



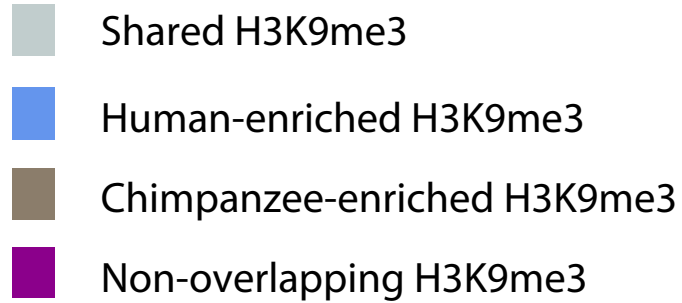
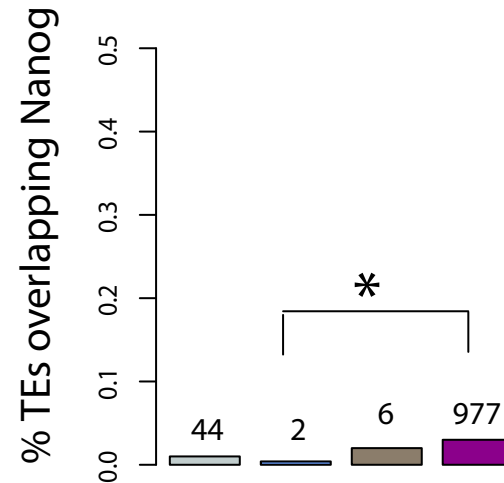
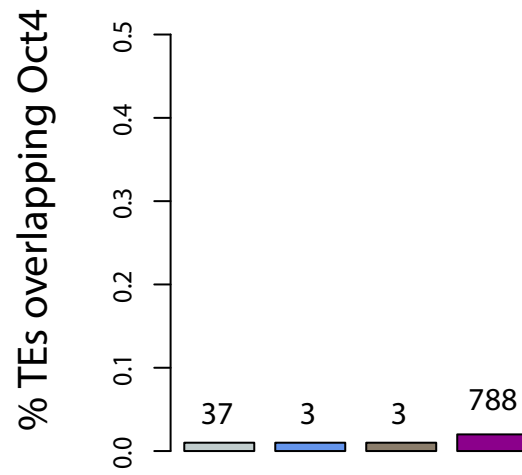
## Active chromatin

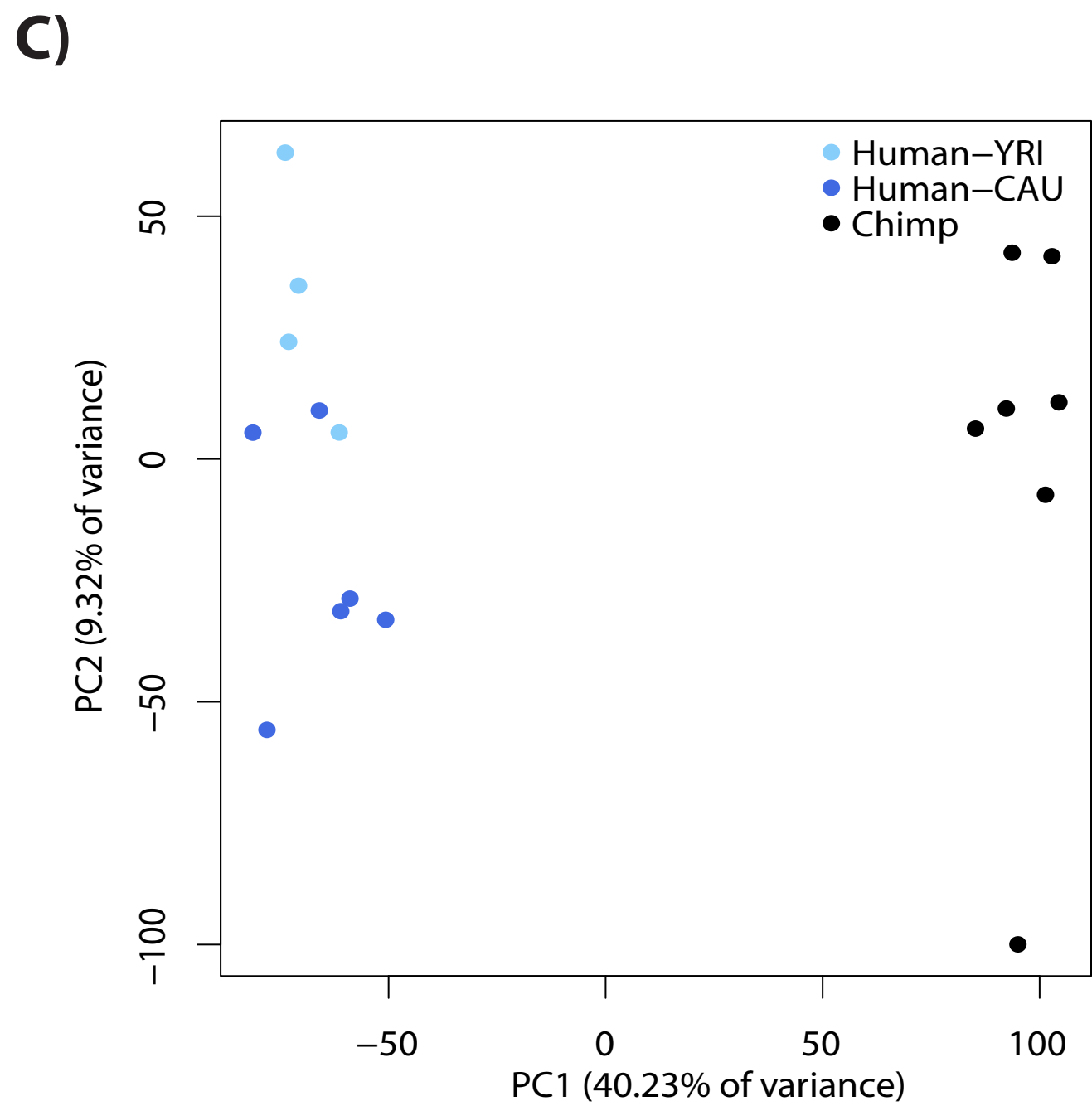
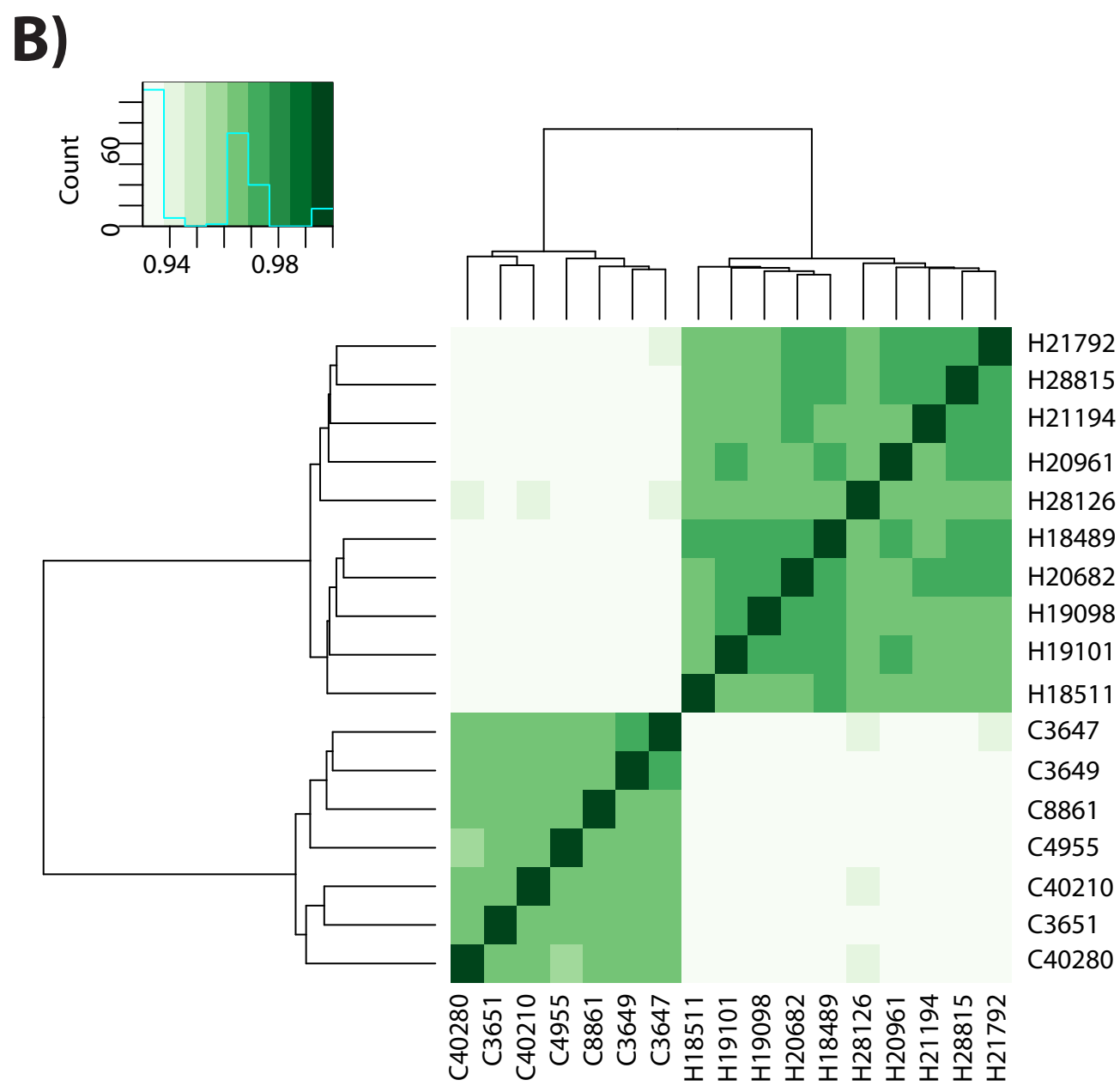
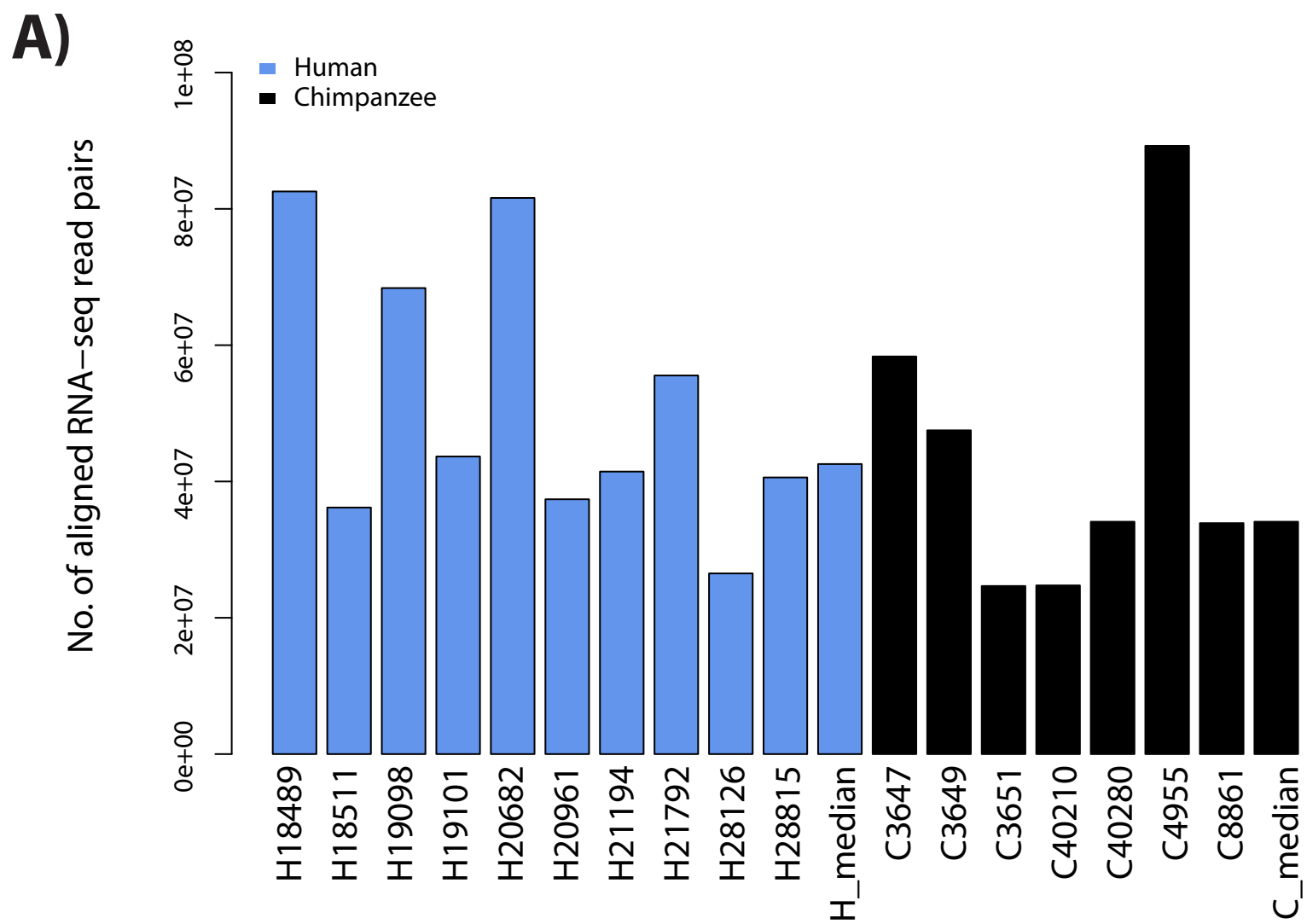


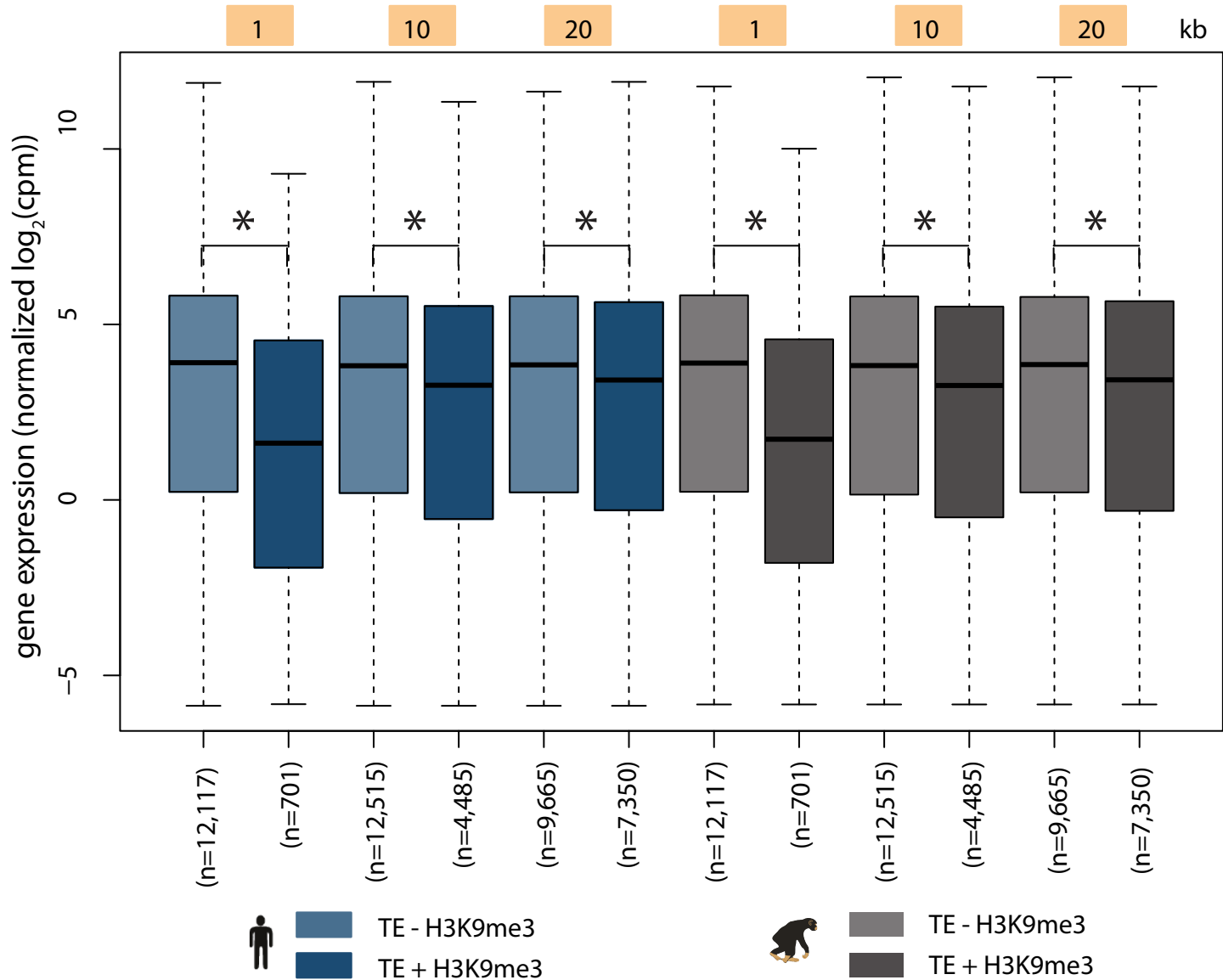
## Ubiquitous TFs



## Pluripotency TFs



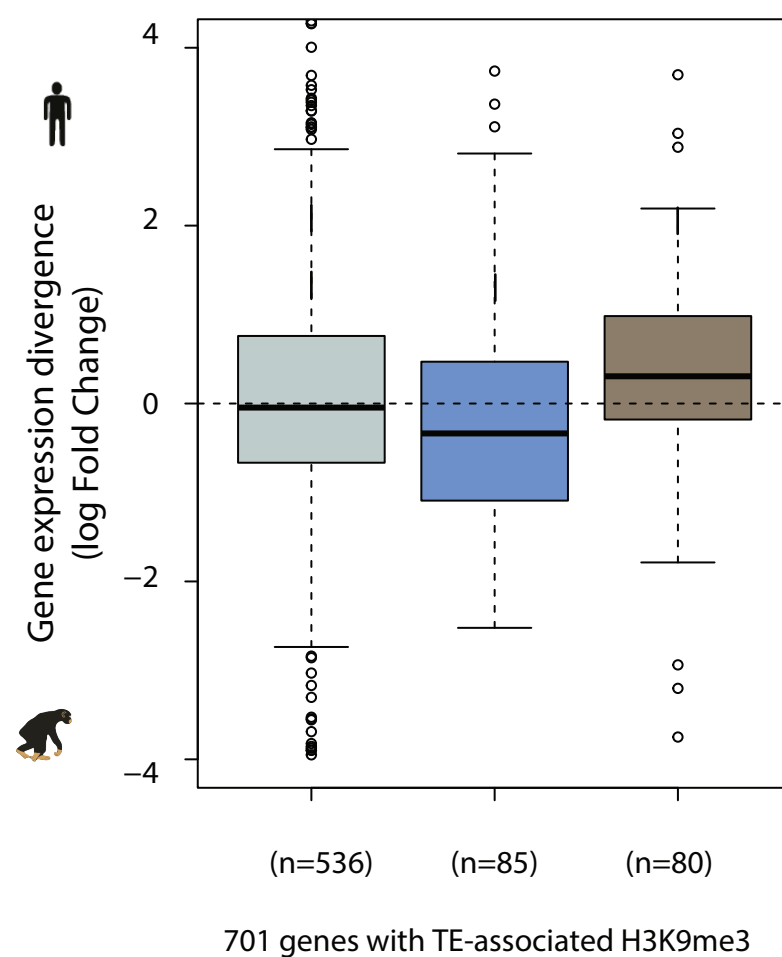
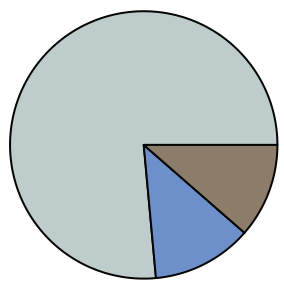
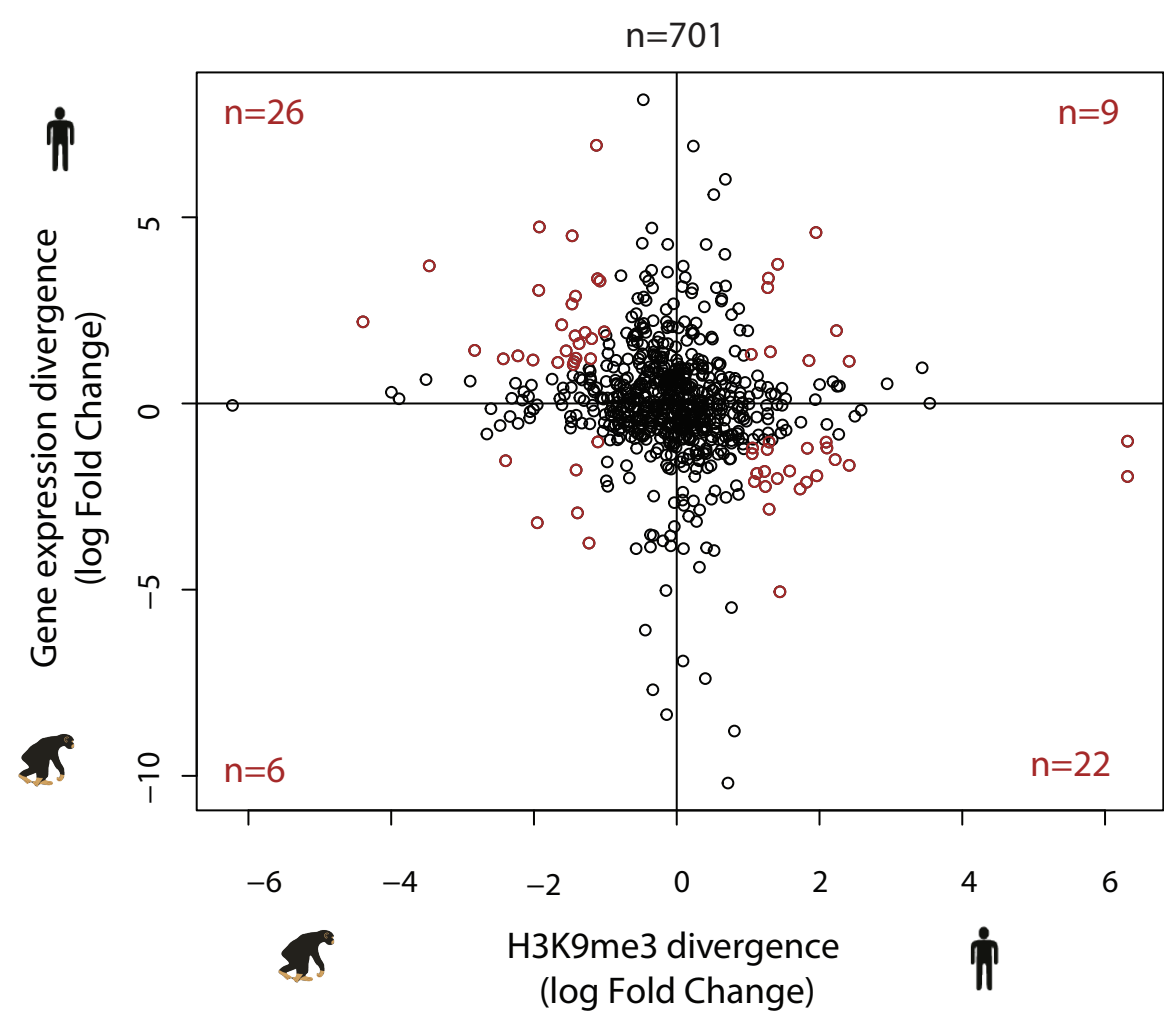




**A)**

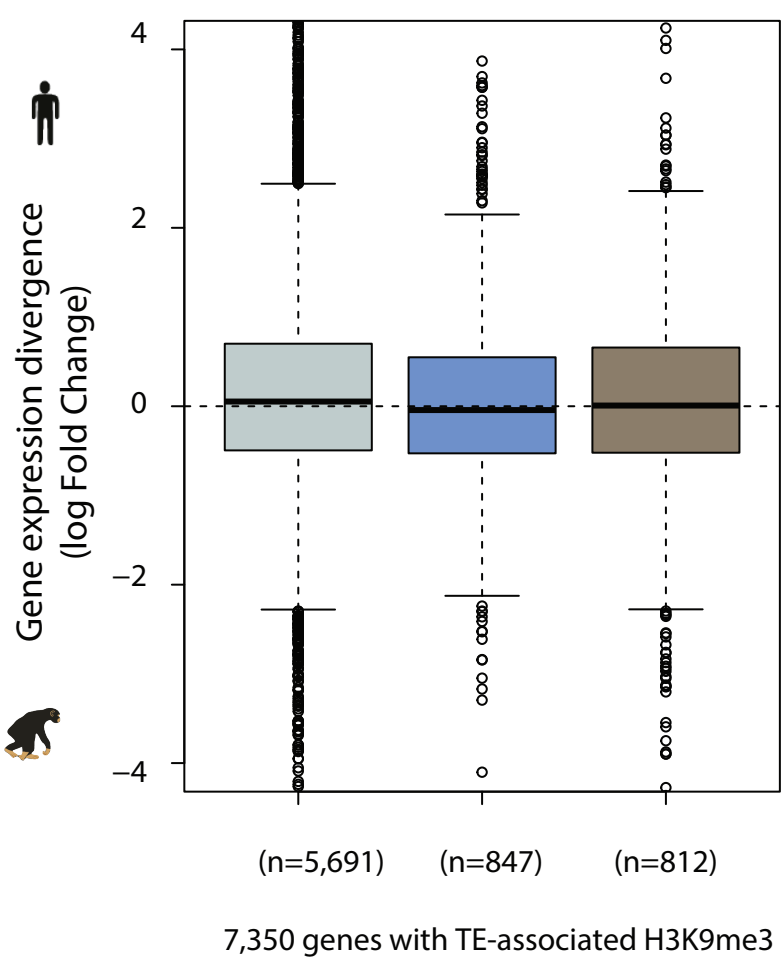
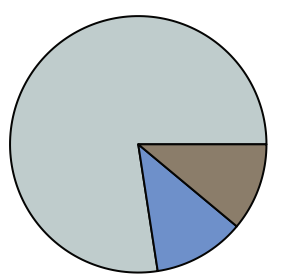
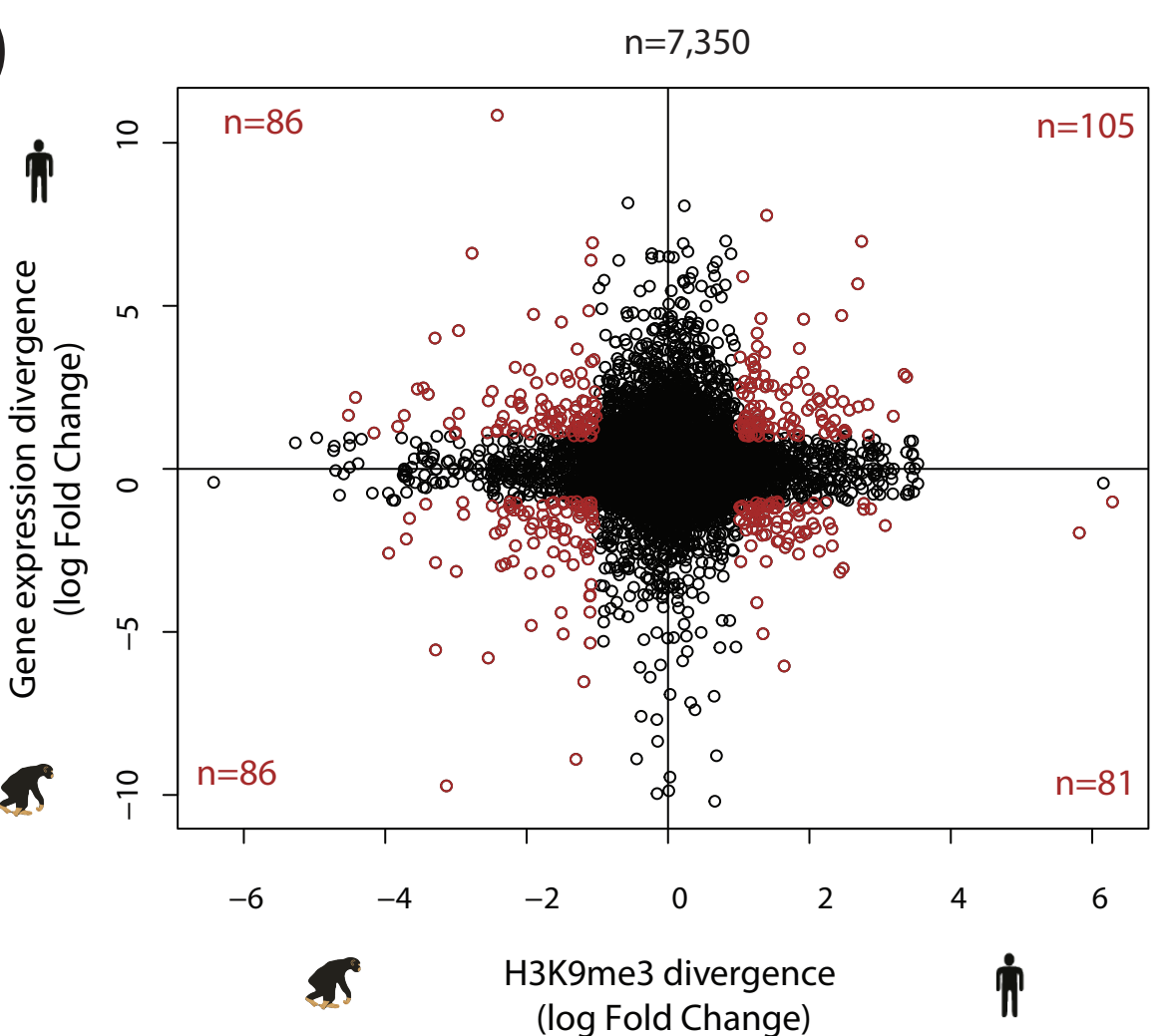
Sum H3K9me3 overlapping TEs  
in 1 kb window:

Shared  
Human-enriched  
Chimpanzee-enriched

**B)****C)**

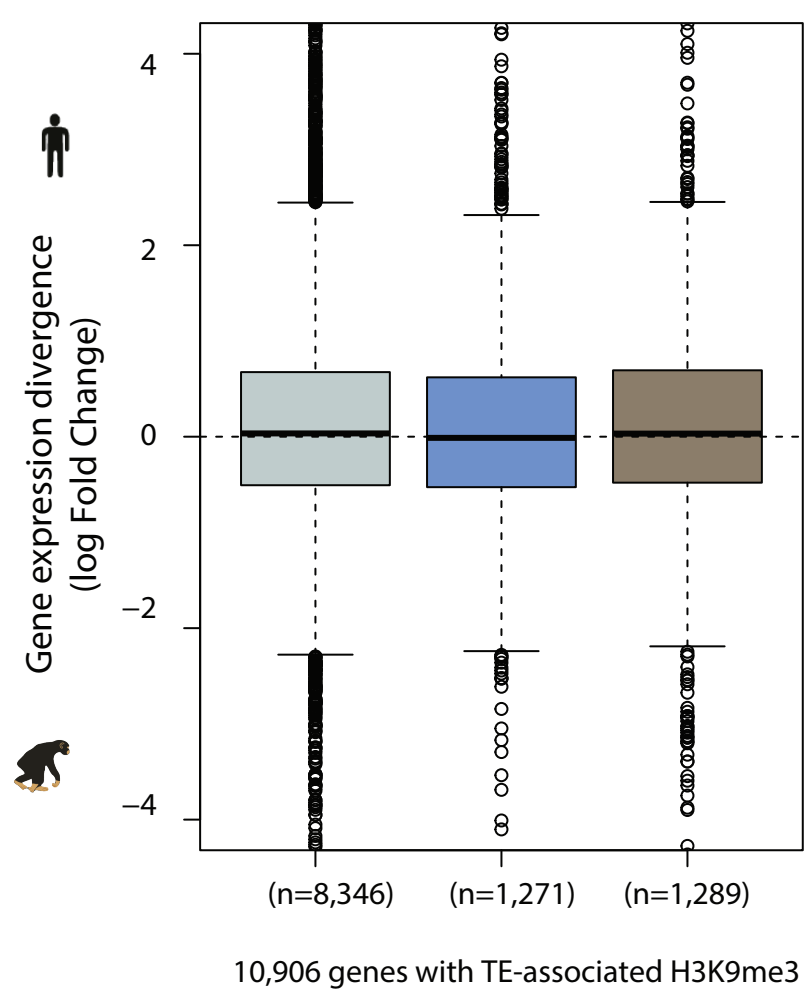
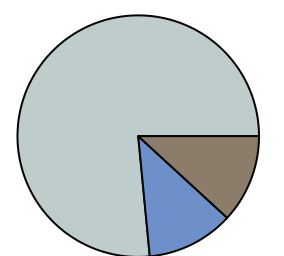
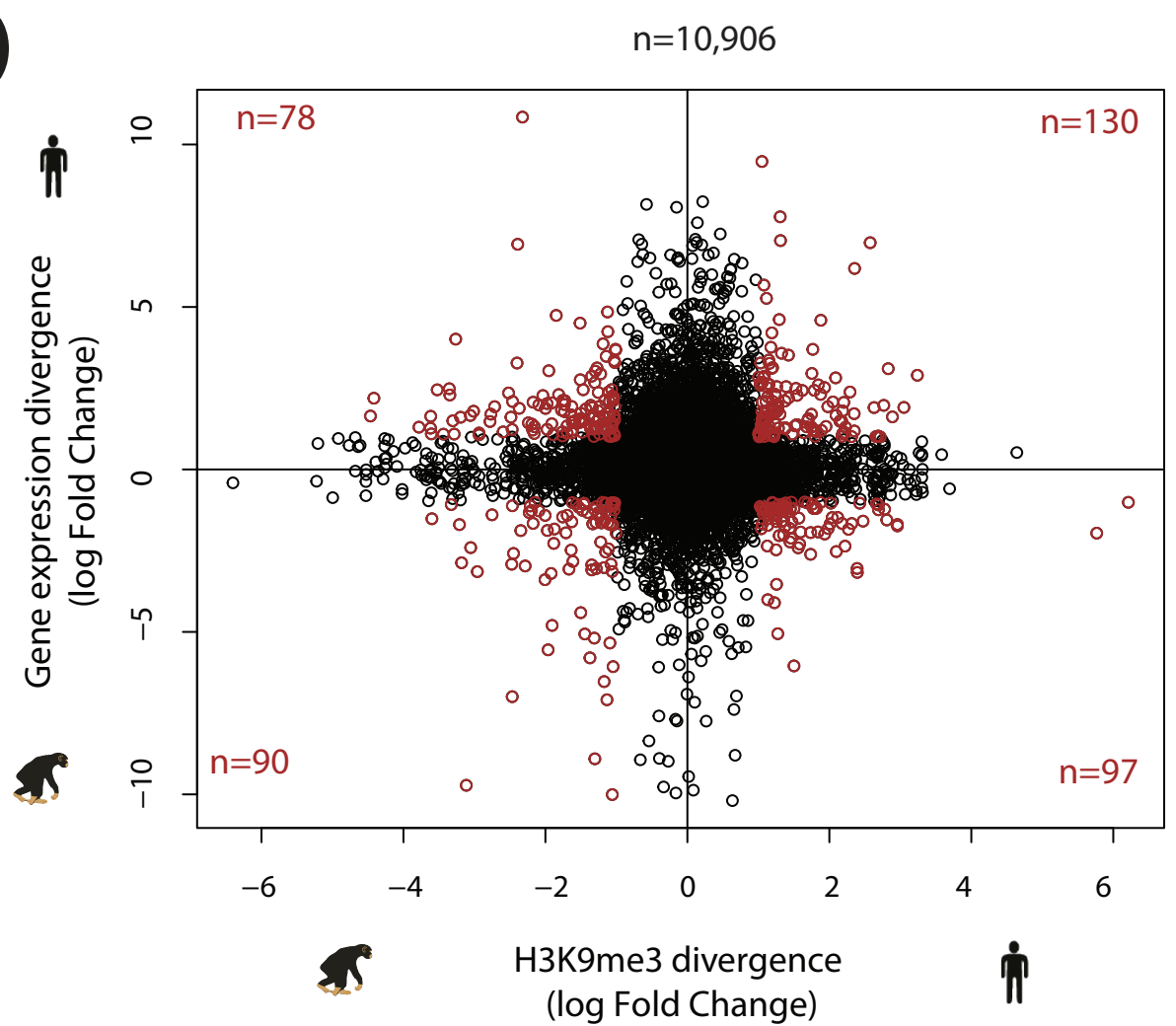
Sum H3K9me3 overlapping TEs  
in 20 kb window:

Shared  
Human-enriched  
Chimpanzee-enriched

**D)****E)**

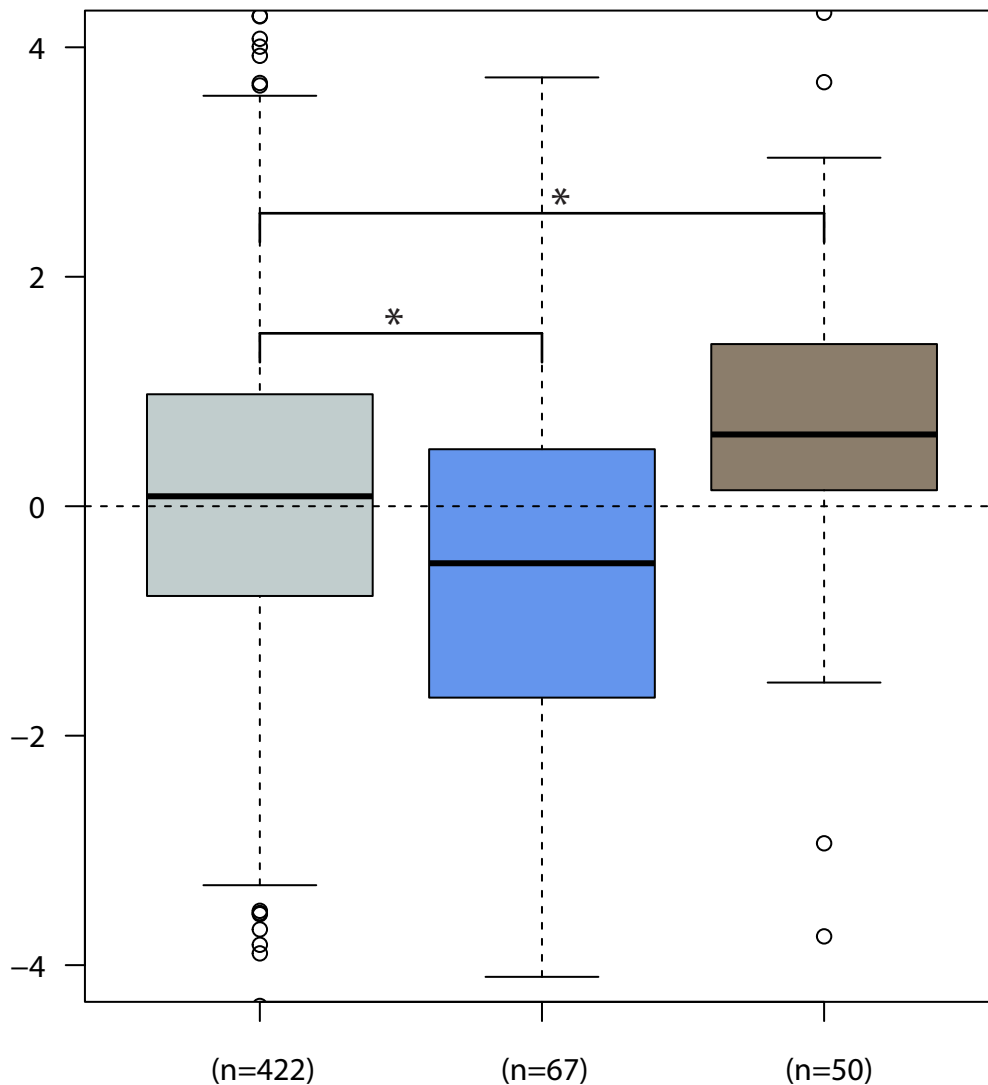
Sum H3K9me3 overlapping TEs  
in 40 kb window:

Shared  
Human-enriched  
Chimpanzee-enriched

**F)**



Gene expression divergence (log Fold Change)



H3K9me3 overlapping TSS:



Shared

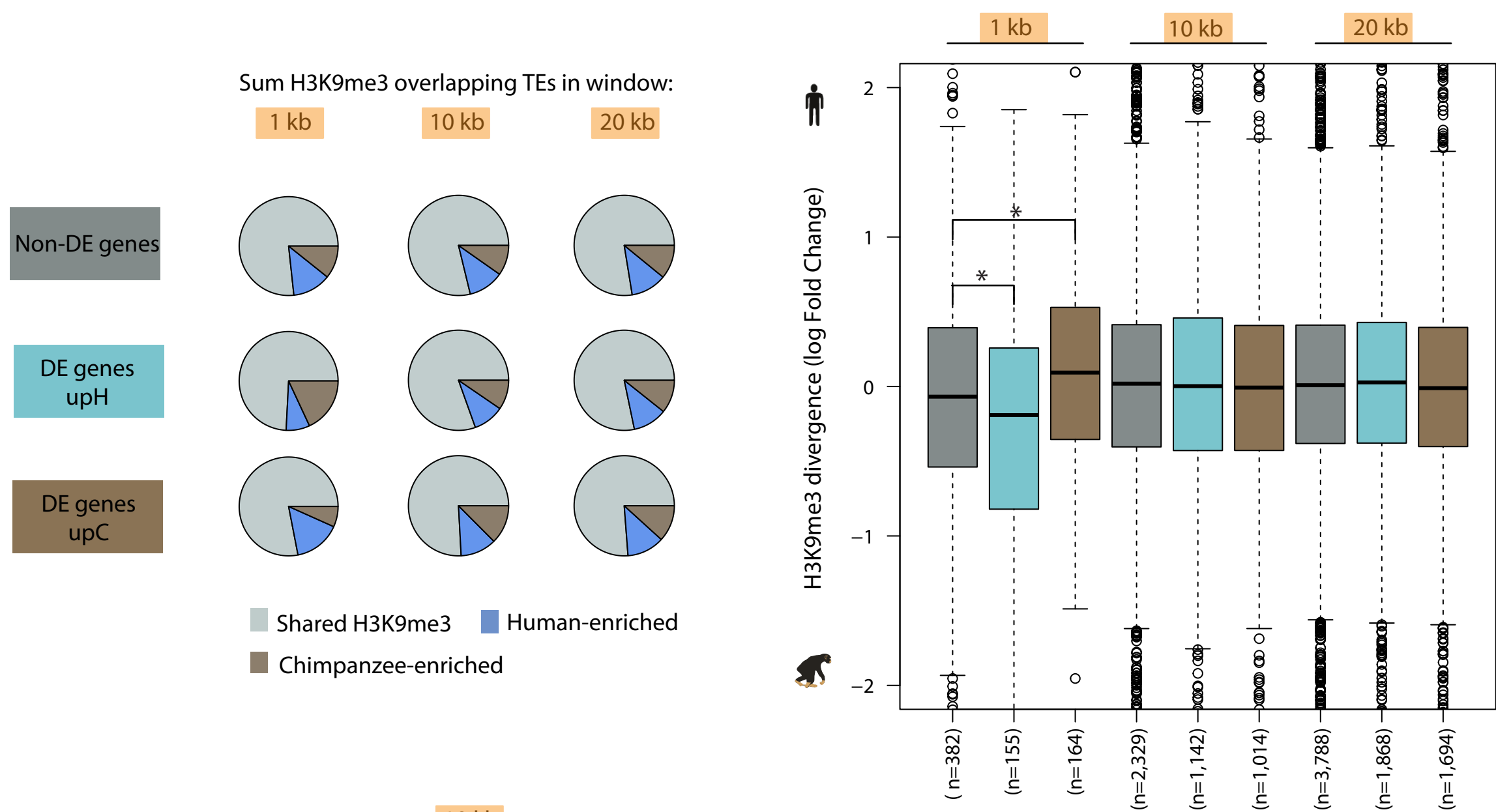


Human-enriched

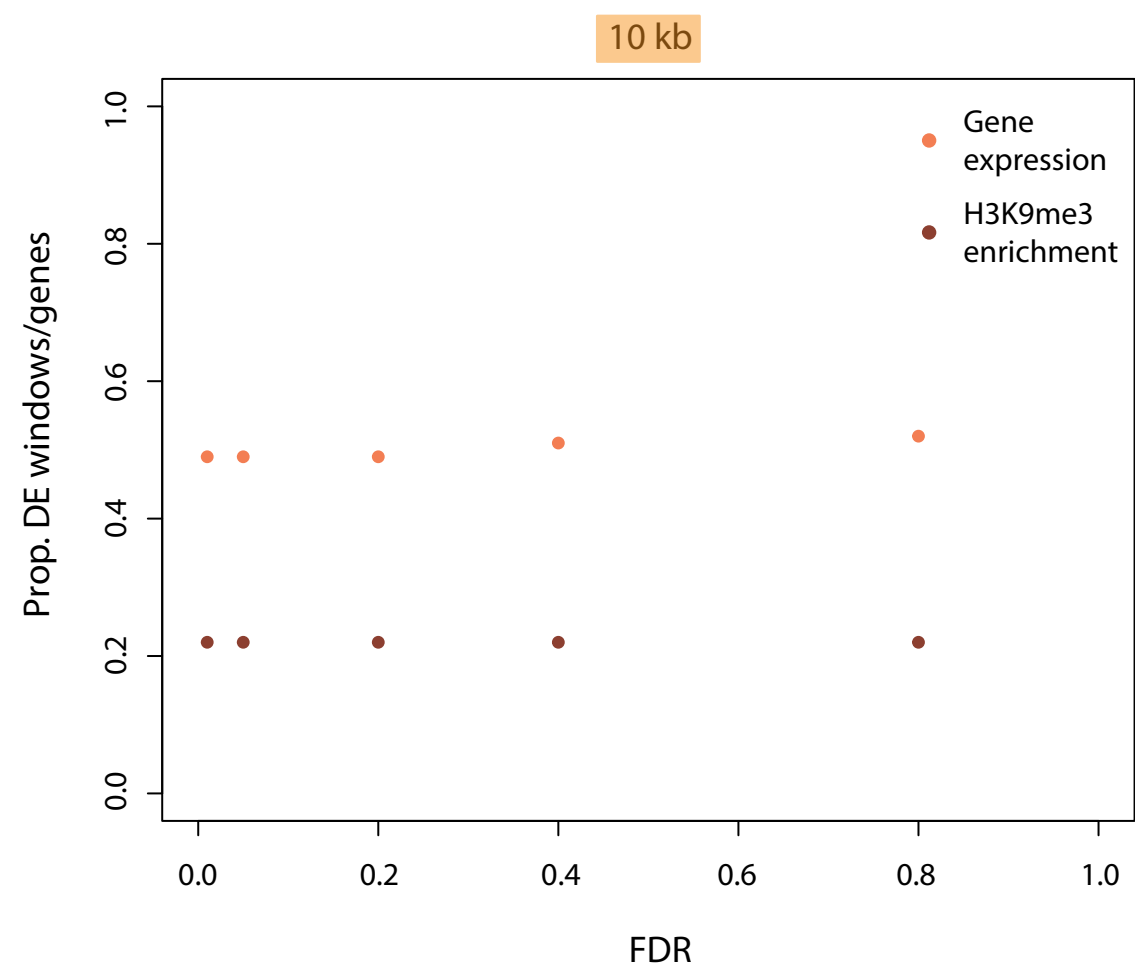


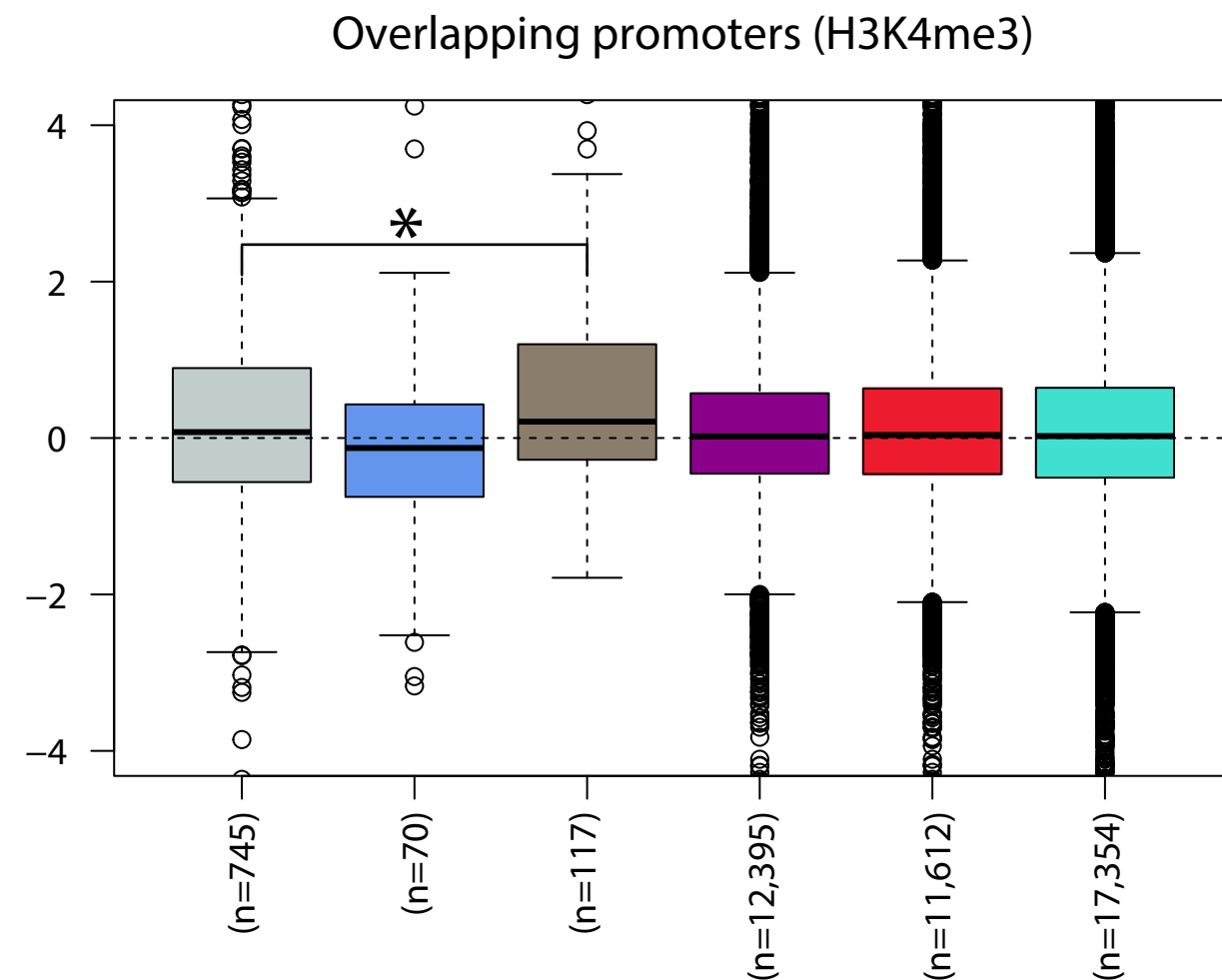
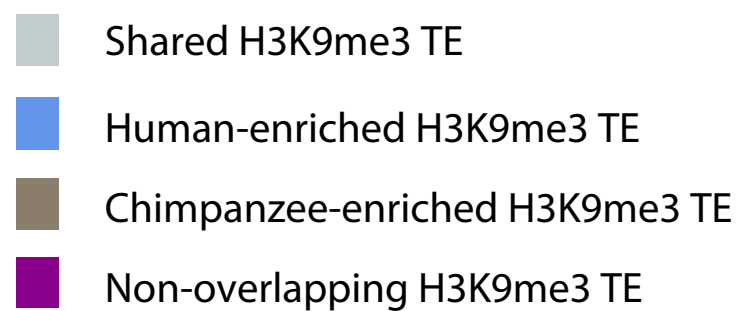
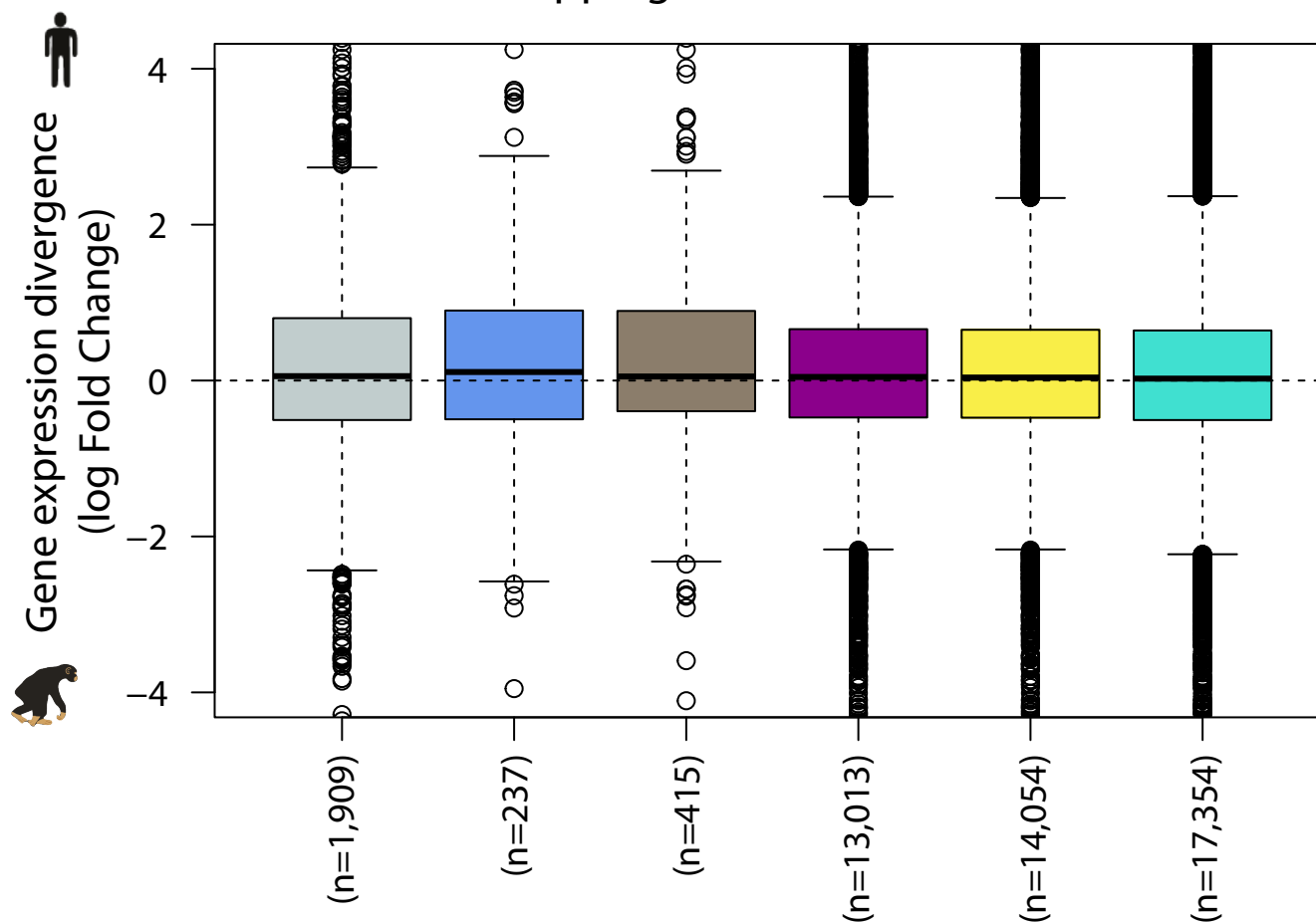
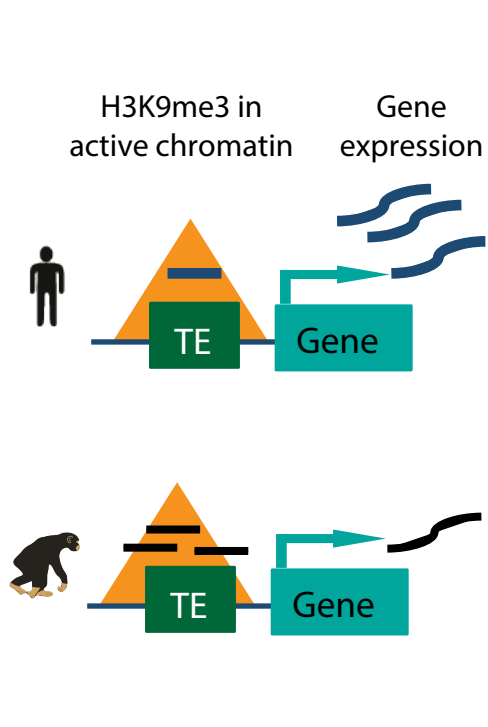
Chimpanzee-enriched

A)



B)

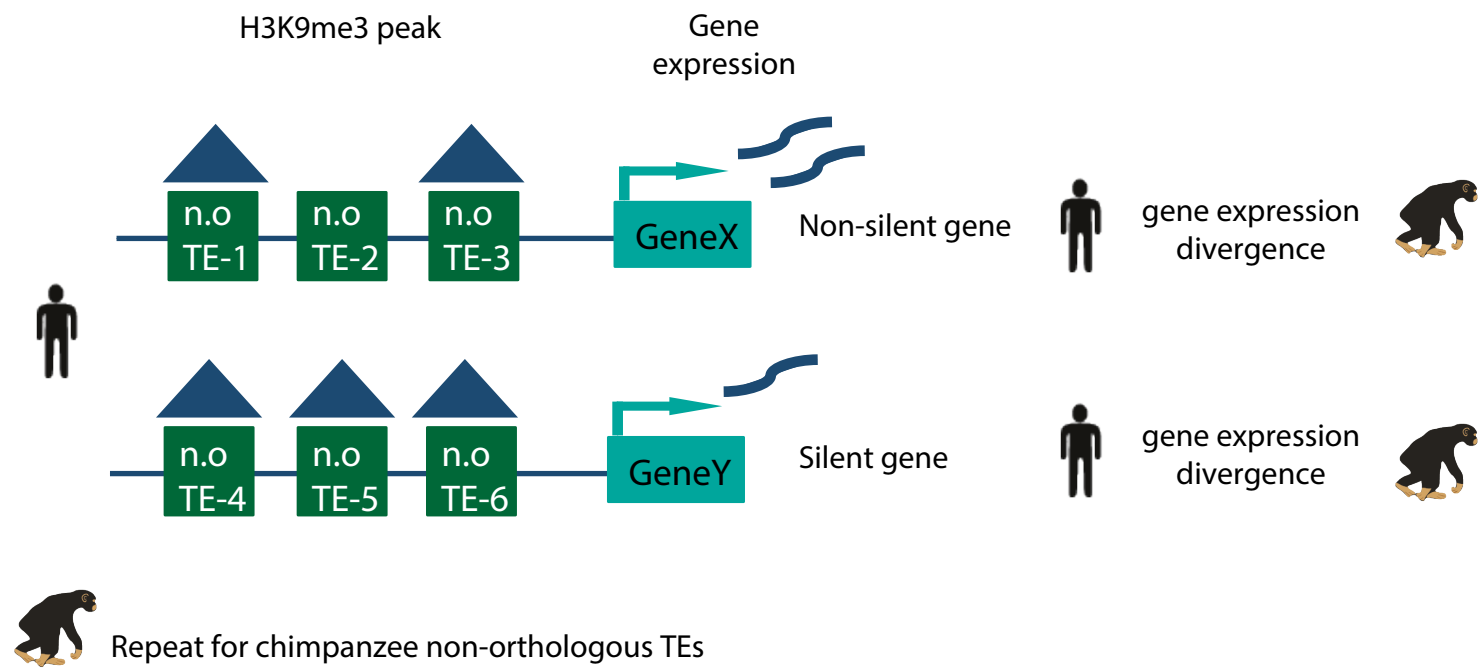




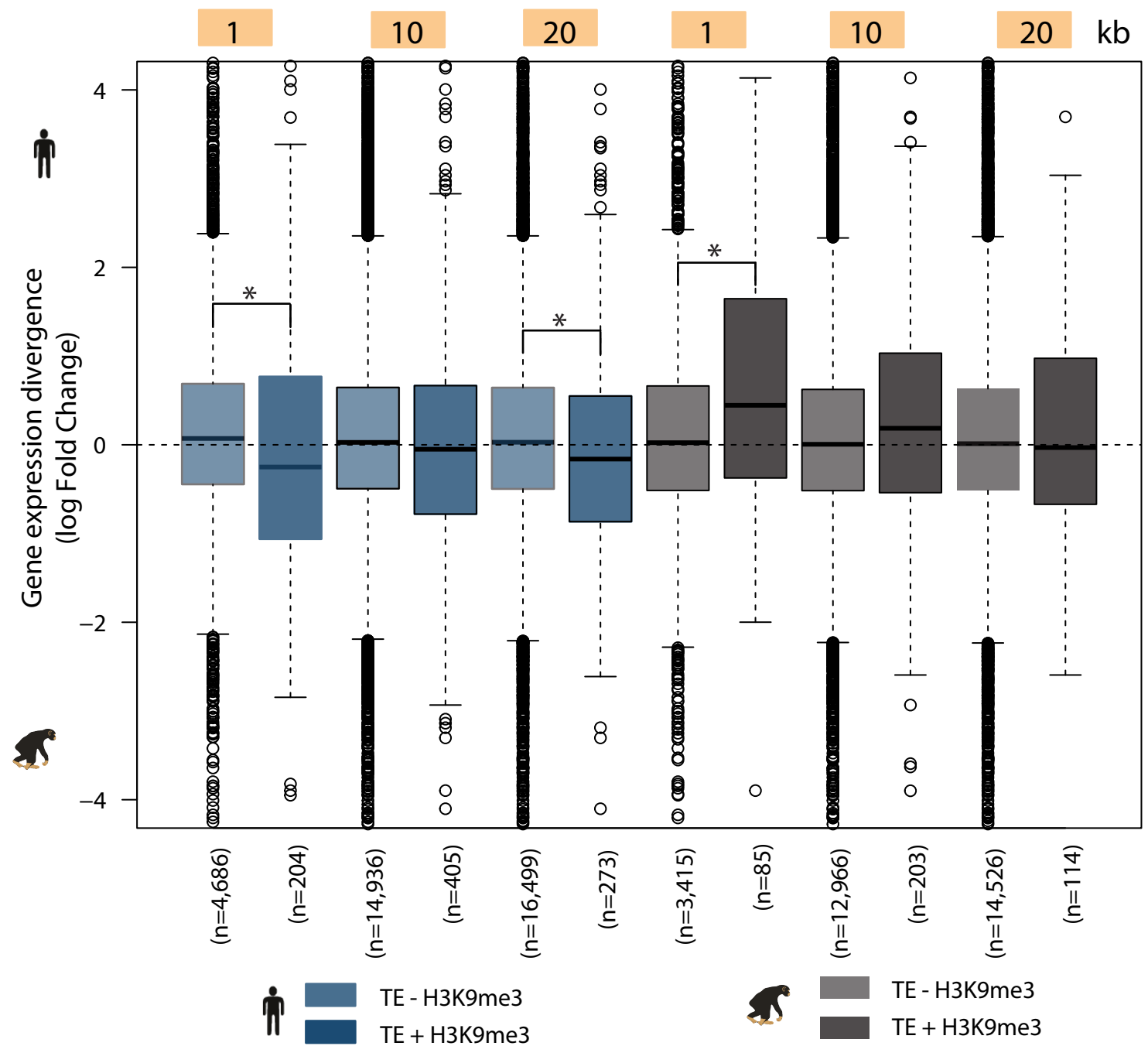


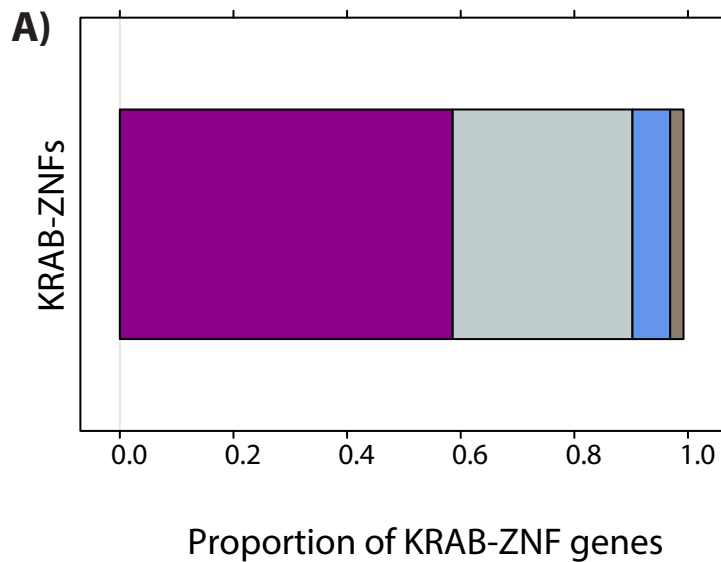
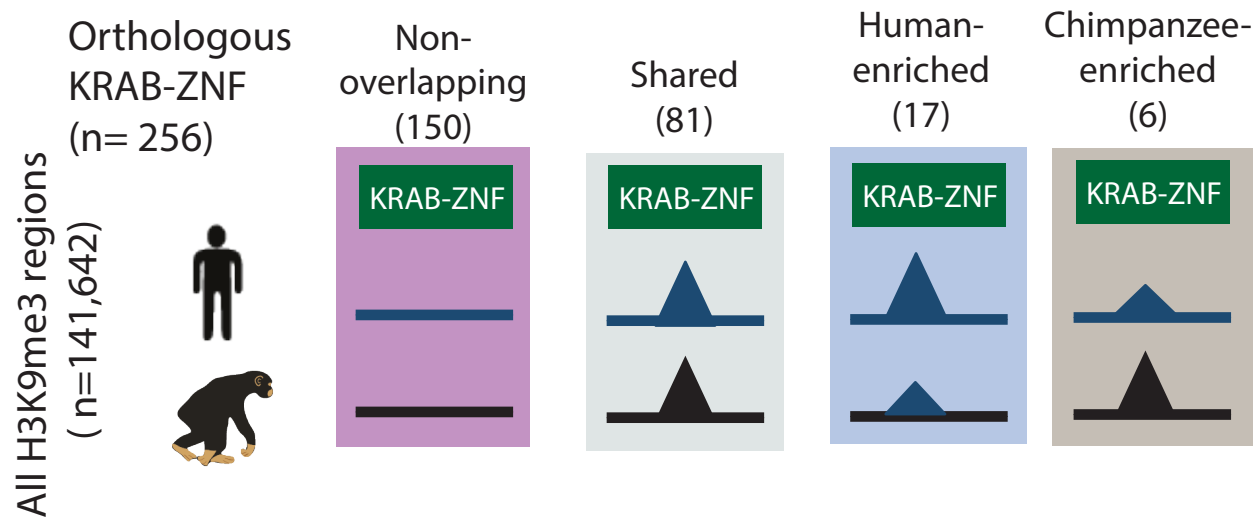
# A)

1/10/20 kb

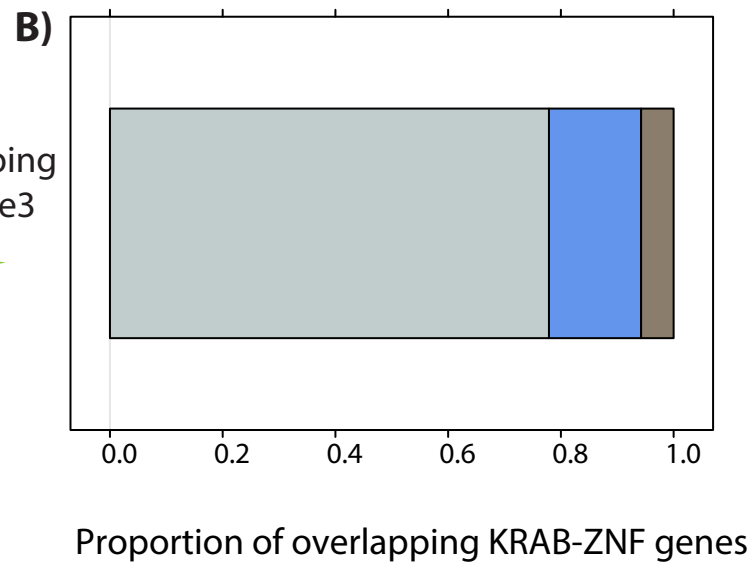


# B)



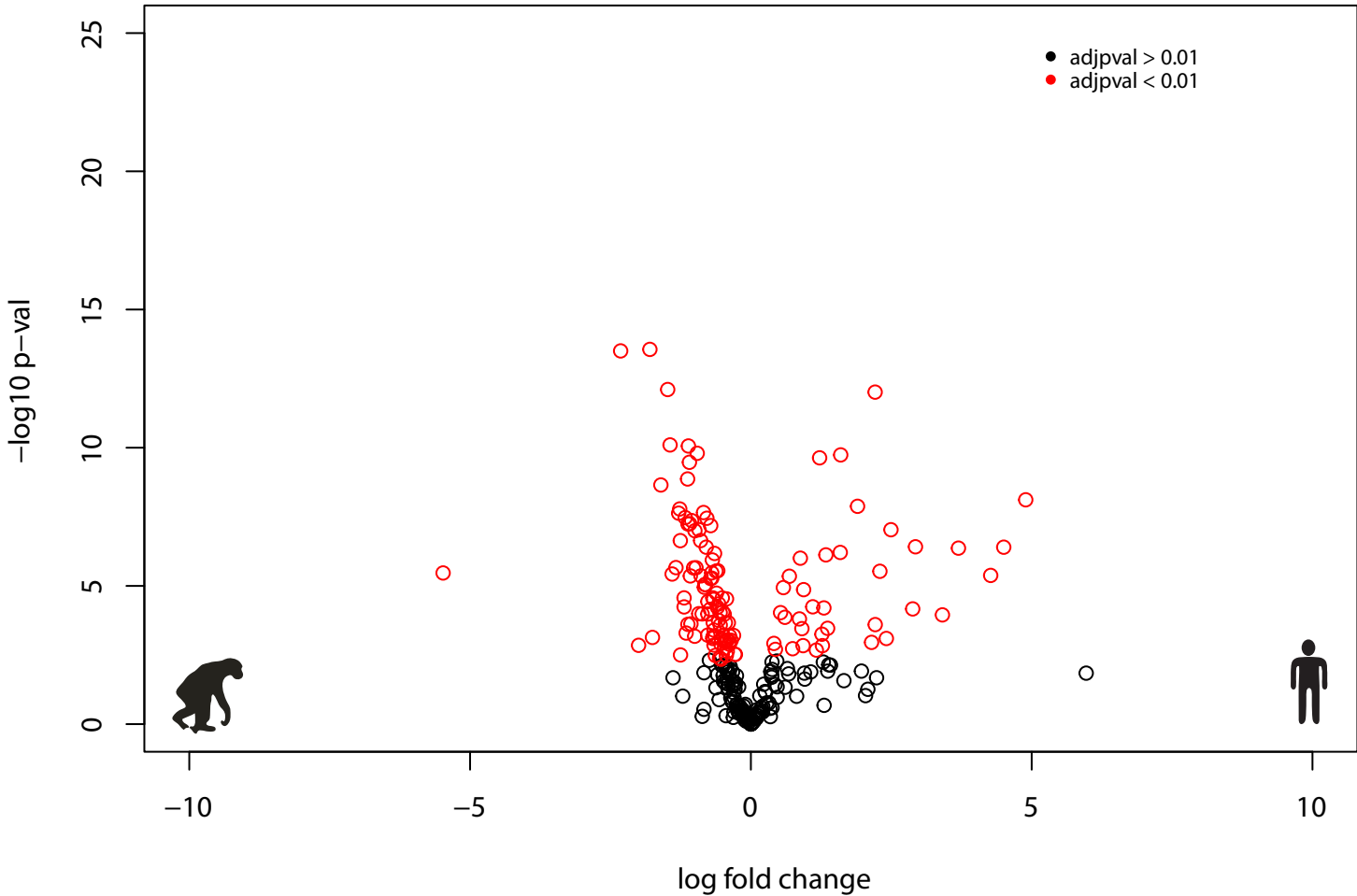


Overlapping H3K9me3

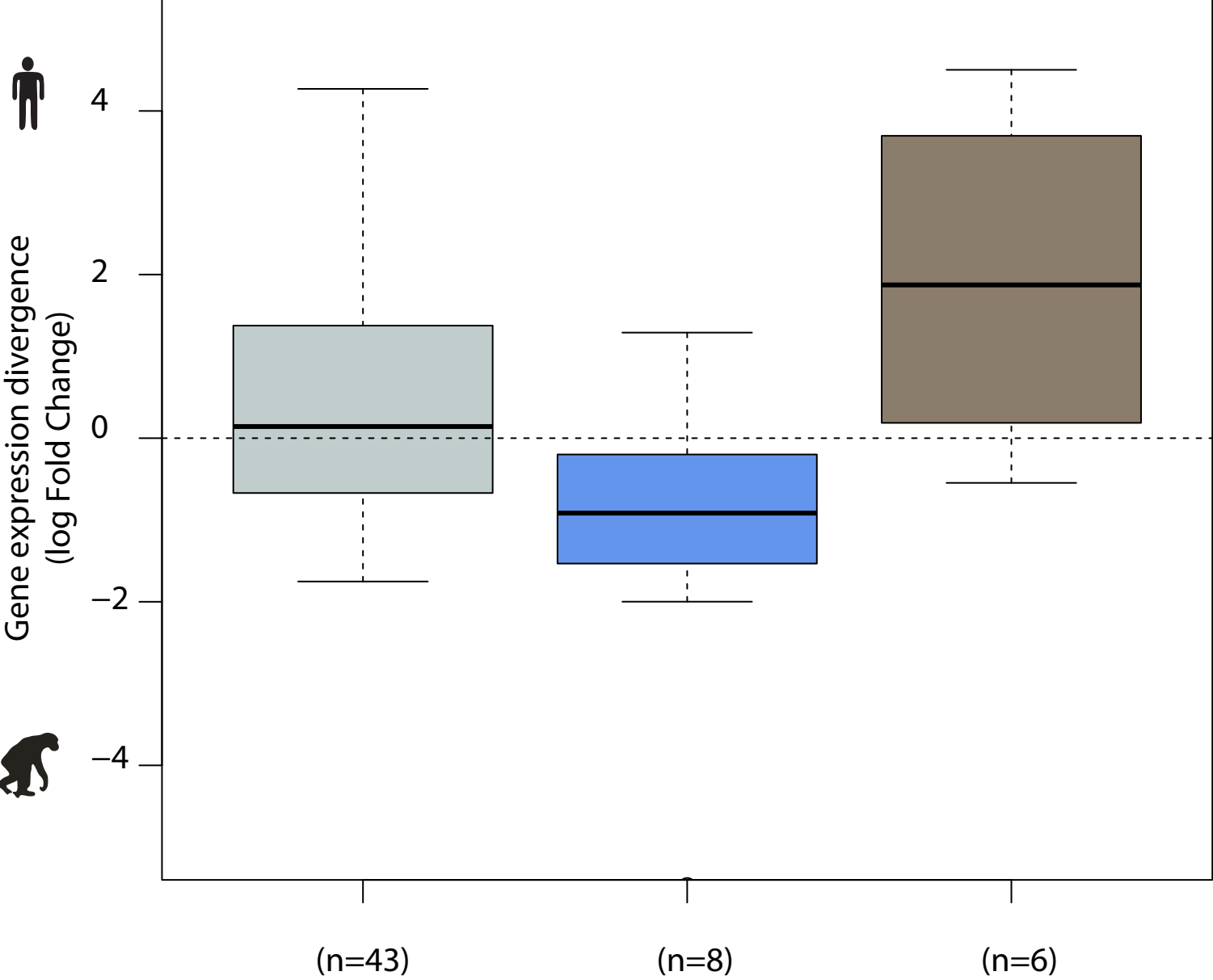


**Ai)**

**Orthologous KRAB-ZNF genes (256)**



**B)**



Shared  
Human-enriched  
Chimpanzee-enriched

**Aii)**

**All orthologous genes (17,354)**

