Dynamic evolution of regulatory element ensembles in primate CD4+ T-cells

Charles G. Danko[1,2,*], Lauren A. Choate[1], Brooke A. Marks[1], Edward J. Rice[1], Zhong Wang[1], Tinyi Chu[1,3], Andre L. Martins[1,3], Noah Dukler[4], Scott A. Coonrod[1,2], Elia D. Tait Wojno[1,5], John T. Lis[6], W. Lee Kraus[7,8], & Adam Siepel[4,*]

Supplementary Notes

Supplementary Note 1: Comparison of dREG sites to other marks of genome function.

In general, dREG sites are highly concordant with other marks of active transcriptional regulatory elements in human CD4+ T-cells. dREG identified >83% of DNase-I hypersensitive sites (DHS) marked by H3K27 acetylation in human CD4+ T-cells, consistent with prior work suggesting that transcription patterns alone can recover the majority of active enhancers defined using independent criteria[1,2]. Furthermore, we identified 88% and 91% of DHSs marked, respectively, by H3K9ac and H4K16ac, two other markers of regulatory function. Yet despite a high degree of overlap, H3K27ac ChIP-seq peaks were 2-fold longer (848 bp on average) than the regions identified by dREG (380 bp for human dREG sites). Thus, dREG predictions are closer in size to the regulatory element core region, consisting of divergently opposing RNA polymerase initiation sites flanking TFBSs[2]. Histone modification data aligned to human dREG sites revealed good agreement with the center of the nucleosome free region (**Fig. 1e**), demonstrating a substantial improvement in both resolution and site localization accuracy compared to the histone modification ChIP-seq data used in previous evolutionary studies. Taken together, these data suggest that PRO-seq patterns reveal the locations of TREs with high sensitivity and spatial resolution.

Supplementary Note 2: Heuristics for the discovery of TREs that change between species.

We defined two types of changes in TRE activities between species. First, we used deSeq2 to identify species-specific changes in the abundance of Pol II at TREs that were active in all species. Second, we developed heuristic tests based on dREG scores to identify complete lineage-specific gains or losses of TREs. The rationale for, and validation of, this second set of tests is described in detail in this Supplementary Note.

A significant weakness of relying on DESeq2 alone is that this approach is overly conservative for identifying bona fide differences between species in TREs, particularly when levels of transcription are low (as with many eRNAs). Moreover, previous reports[3,4] suggest that evolutionary changes in TREs are common in mammals. Thus, filtering on false discovery rates from deSeq2 would result in very large numbers of false negatives. We therefore developed alternative criteria based on the dREG scores themselves, which are more sensitive to the transcriptional signatures of TREs than the raw read counts considered by deSeq2. This strategy can be considered analogous to the peak-calling strategy adopted in several previous studies, in that it depends on a customized, species-specific processing of the data rather than on raw read counts.

Our strategy selects putative differences between species using two thresholds in order to minimize errors where both species are near a single selected threshold. In particular, we select sites that both (1) exceed a stringent dREG score threshold (>0.3) in one or more of the species, indicating high-confidence presence of a TRE, and (2) fall below a permissive dREG threshold in at least one of the other species (<0.05), indicating high-confidence absence of a TRE. The stringent threshold (0.3) was selected because at this cutoff more than 93% of dREG sites identified in human CD4+ T-cells overlap another mark of active promoters and enhancers, including DNase-I, H3K27ac, H3K9ac, H4K14ac, H3K4me3, H3K4me1, or promoters and enhancers annotated using chromHMM[5]. The permissive threshold (0.05) was selected because at this threshold less than 6% of DNase-I hypersensitive sites that are also marked by H3K27ac remain to be identified, suggesting that relatively few true TREs exist below this threshold. By applying both of these thresholds in combination, we obtain a relatively stringent test for TREs that are present in at least one species and absent in at least one species, which require a gain or a loss event to explain.

Several lines of evidence support the idea that differences between species discovered using these two thresholds are highly enriched for bona fide evolutionary changes. First, differences between species on the basis of these criteria were highly enriched for statistically significant p-values estimated by DESeq2 based on the abundance of Pol II loading at these sites (**Supplementary Fig. 5**). Second, we used two independent statistical strategies[6–8] to estimate that, in not more than 10-15% of these cases, the null hypothesis of no evolutionary change is true, suggesting that 85-90% of these differences reflect bona fide gains and losses. Third, based on these criteria, gains and losses of TREs between treated and untreated samples were rare (though many TREs changed the abundance of transcription at TREs) (**Fig. 2a**) demonstrating that our criteria are robust when applied within a species. Fourth, the predicted gains and losses exhibit correlated changes in active histone modifications measured using orthogonal forms of genomic data in human CD4+ T-cells (**Fig. 2b**).

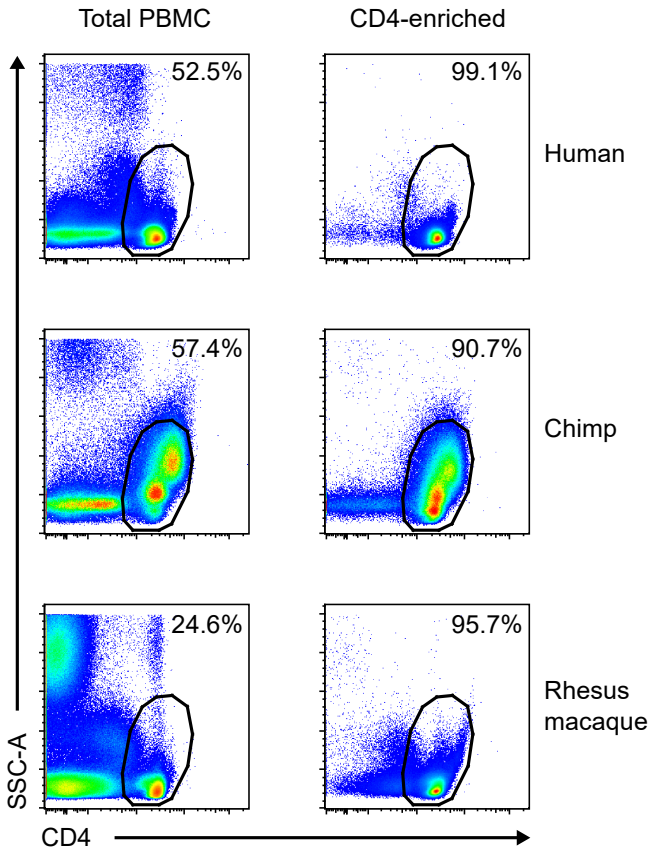
Fifth, several differences between species at the *SGPP2* locus were validated experimentally using reporter gene assays (**Fig. 3; Supplementary Fig. 10**). These results suggest that our heuristic criteria are largely comprised of bona-fide differences between species.

Lastly, we also tested the sensitivity of our major results to the specific values of these thresholds. All of the major covariates that correlate with rates of change in enhancer activity between species were robust to reasonable changes in the threshold, or even to whether or not we filter on the false discovery rate in Pol II loading based on p-values estimated by DESeq2 (see **Supplementary Table 2**).

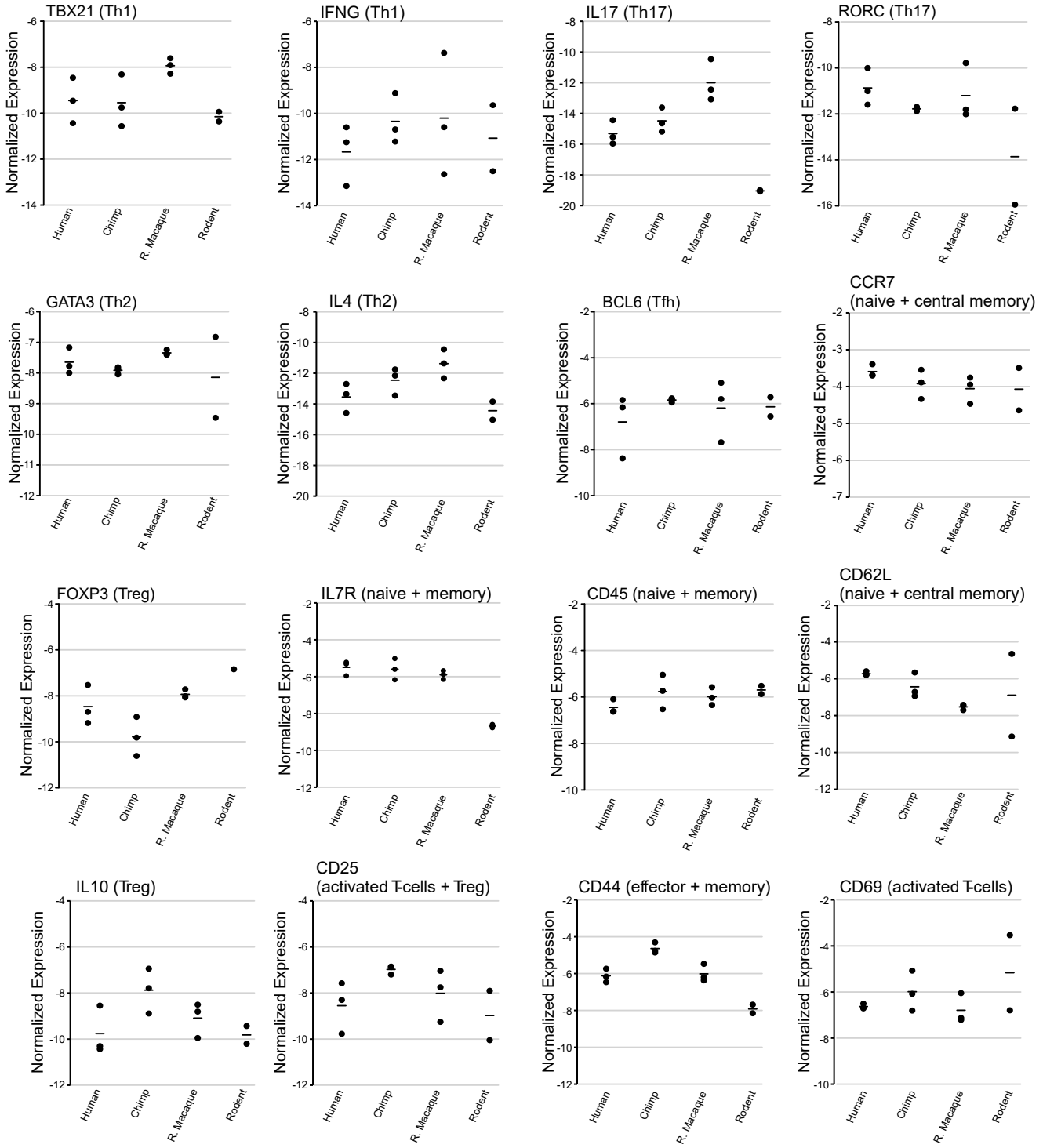
Supplementary Note 3: Correlation between protein-coding and non-coding transcription.

We noticed that evolutionary changes in protein-coding gene transcription frequently correlate with changes in non-coding transcription units (TU) located nearby. To examine this pattern more generally, we adapted our recently reported hidden Markov model (HMM)[9] to estimate the location of TUs genome-wide, based on patterns of aligned PRO-seq reads and the location of TREs. Using this method, we annotated 54,793 TUs active in CD4+ T-cells of at least one of the primate species, approximately half of which overlap existing GENCODE annotations or their associated upstream antisense RNAs (**Supplementary Fig. 13**). A cross-species comparison of the transcription levels for various TU classes (**Fig. 4a**) revealed that non-coding RNAs evolve in expression most rapidly and protein-coding genes evolve most slowly. GENCODE-annotated lincRNAs undergo evolutionary changes in expression about as frequently as the unannotated non-coding RNAs predicted by our HMM, which are likely enriched for bi-directionally unstable eRNA species.

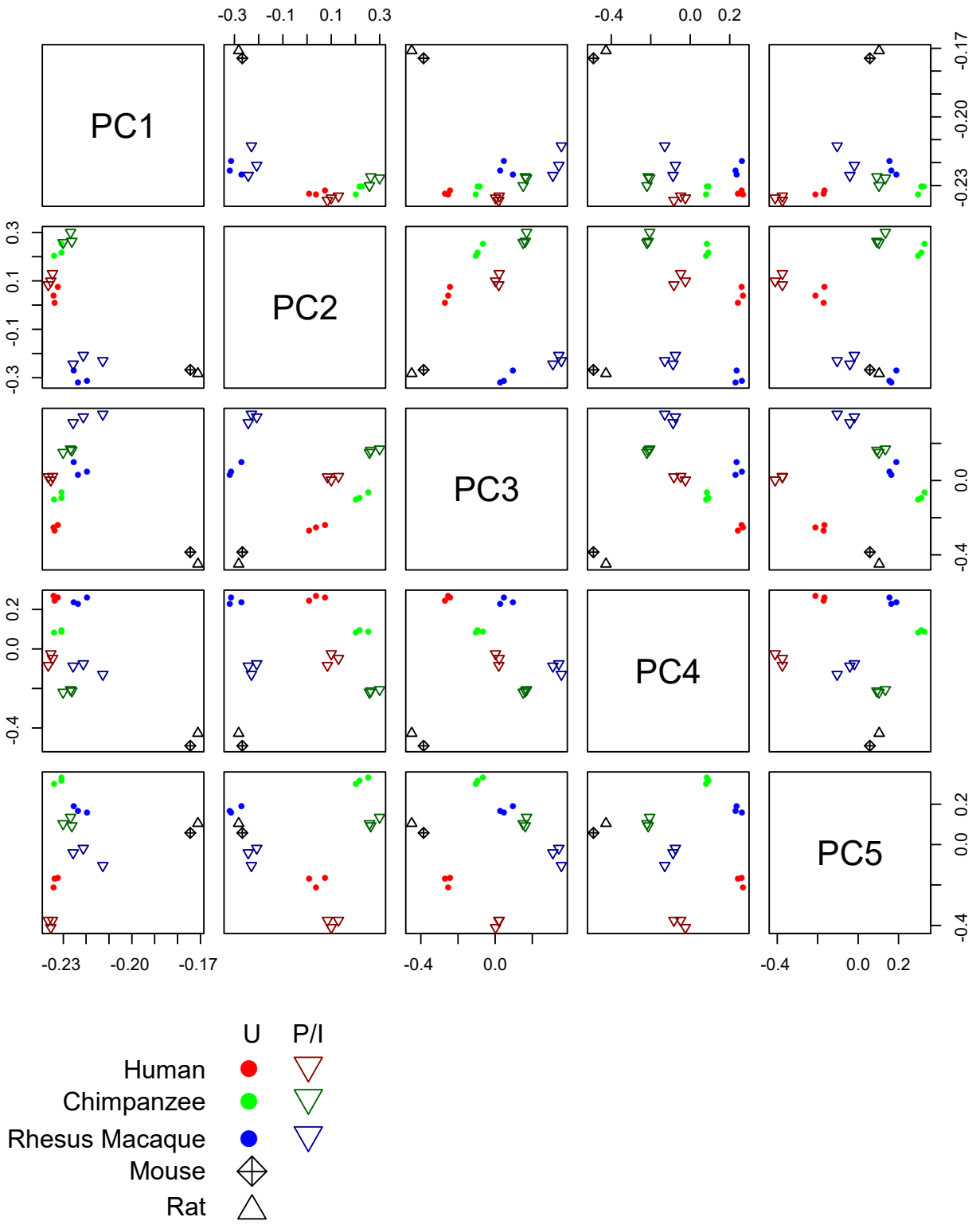
Reports to date have indicated a surprisingly limited correlation between evolutionary changes in gene expression and changes in the activity of distal enhancers[10–12]. We used PRO-seq read counts for non-coding RNAs as a proxy for TRE activity and measured the extent to which non-coding and protein-coding transcriptional activities are correlated through evolutionary time. Using a generalized linear model to integrate expression changes in multiple types of TUs, we were able to explain 74% of the variance in gene transcription levels when we observe differences between species (R^2 = 0.74 in a held-out set of sites, p < 2.2e-16; **Fig. 4b**) based on the activities of looped TREs, nearby TREs in the same topological associated domain, internal antisense TUs, and the upstream antisense TU. Thus, evolutionary changes that result in differences in Pol II recruitment to protein-coding genes are well correlated at the transcriptional level across interacting TREs



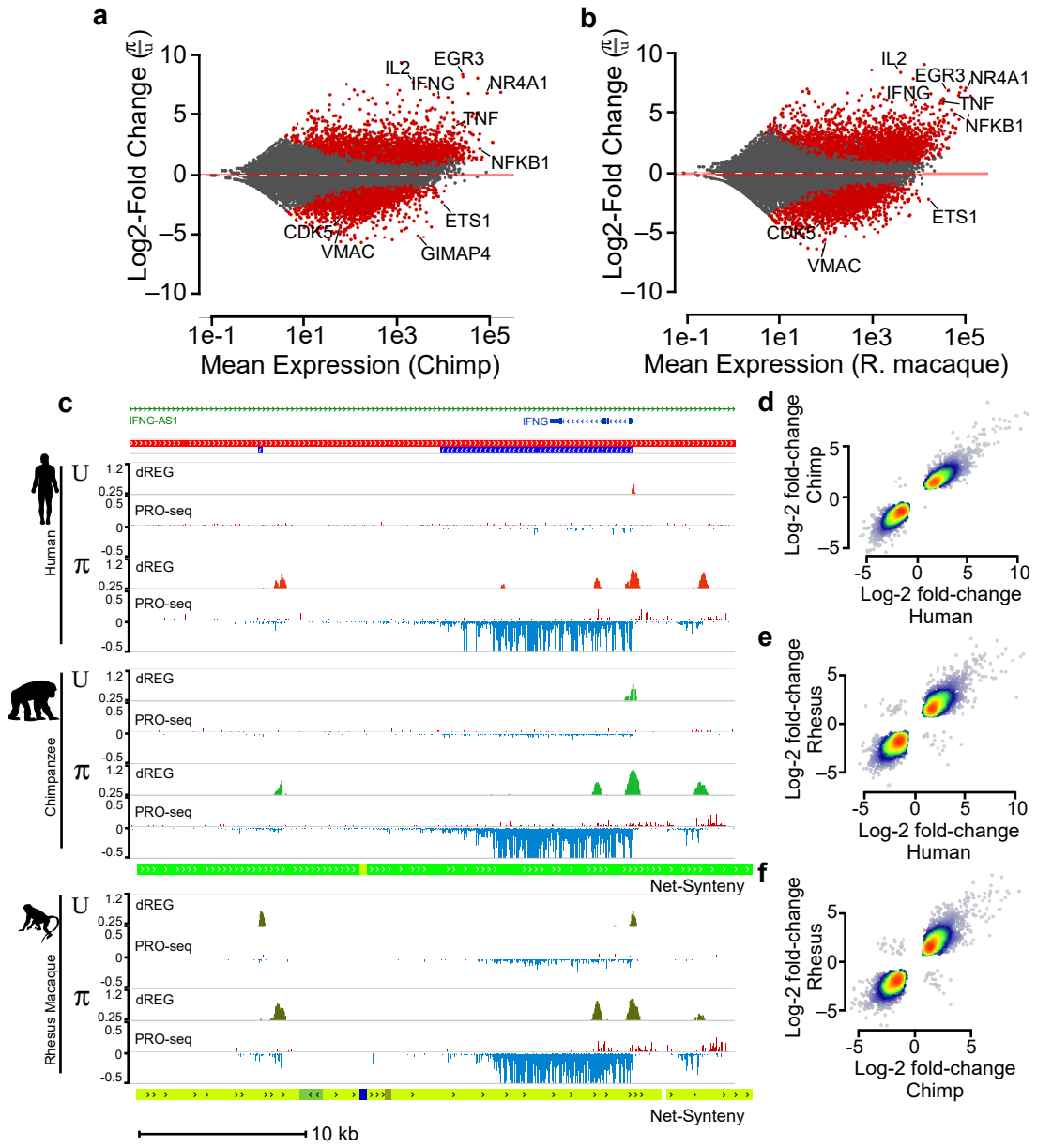
Supplementary Fig. 1 | Validation of CD4+ cell enrichment by flow cytometry. Representative plots of CD4 expression in human, chimpanzee, and rhesus macaque PBMC, before (left) and after (right) CD4 microbead enrichment. Percentage of total live lymphocytes shown.



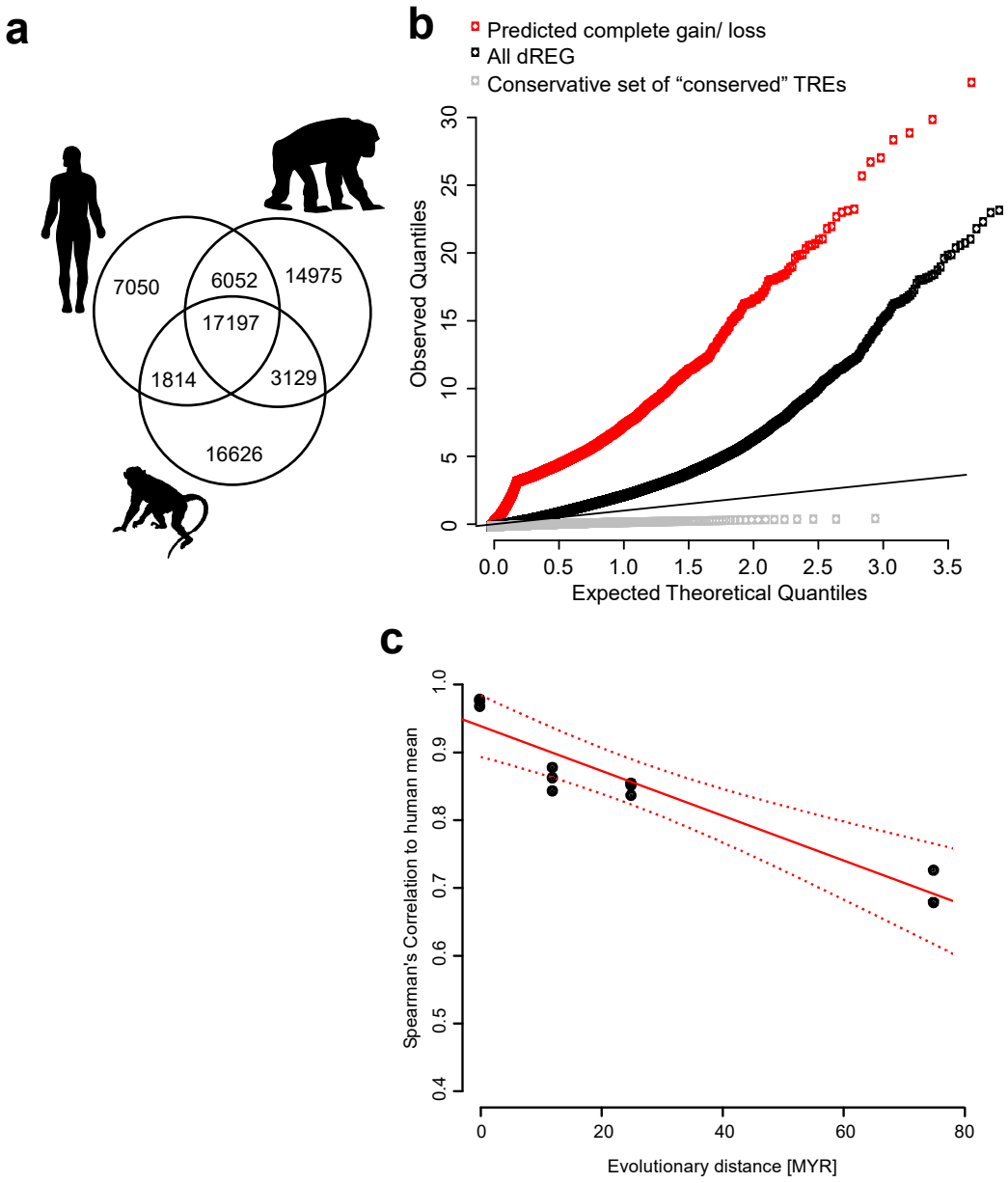
Supplementary Fig. 2 | Transcription abundance in the gene bodies of T-cell lineage specific markers. Plots show normalized expression (log2 scale) of transcription factors and cytokines that mark specific subsets of CD4+ T-cell population in the species indicated below the plot. Each point represents the transcription of the indicated gene in a different untreated T-cell sample. The bar indicates the mean in each species. In all cases read counts were limited to regions of orthology in the bodies of genes indicated on each plot.



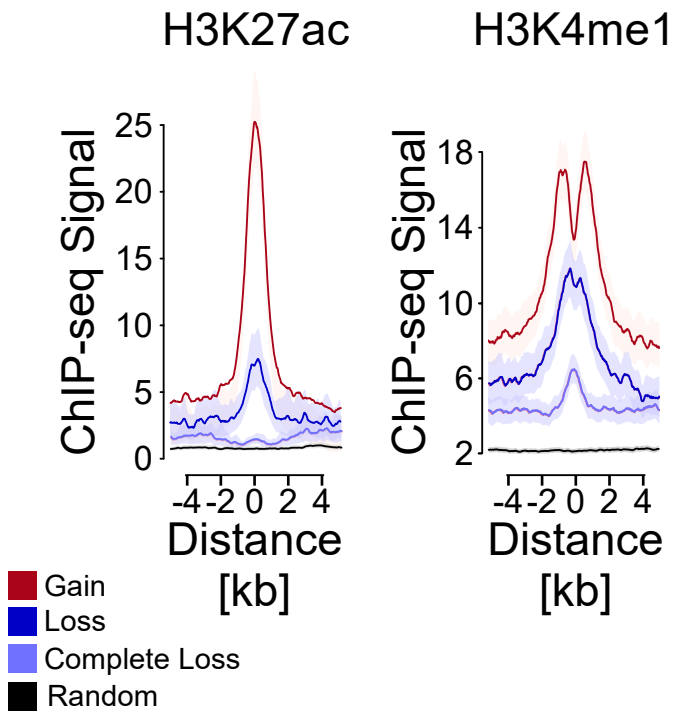
Supplementary Fig. 3 | Principal component analysis (PCA) of CD4+ T-cell PRO-seq libraries. Scatterplots show the first five principal components (PC) from CD4+ T-cell PRO-seq libraries. PCA was constructed using regions of orthology in all five species in the bodies of transcription units identified by a three state hidden Markov model. The key shown below the plot indicates the species and treatment condition of each point.



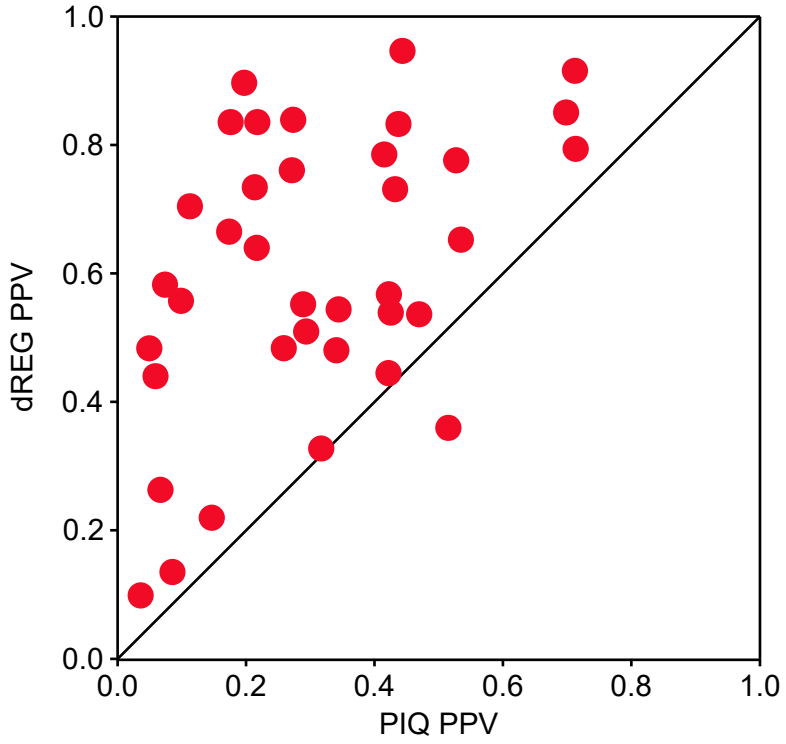
Supplementary Fig. 4 | Changes in gene transcription following PMA+Ionomycin treatment in chimpanzee and rhesus macaque CD4+ T-cells. (a-b) MA plot shows the log-2 fold change following π treatment (y-axis) as a function of the mean transcription level in GENCODE annotated genes (x-axis) in data from chimpanzee (left) and rhesus macaque (right) CD4+ T-cells. Red points indicate statistically significant changes ($p < 0.01$). Several classical response genes that undergo well-documented changes in transcript abundance following CD4+ T-cell activation (e.g., IL2, IFNG, TNF, and EGR3) are marked. **(c)** UCSC genome browser track shows transcription in the IFNG locus in untreated (U) and PMA+ionomycin (π) treated CD4+ T-cells isolated from the primate species indicated at left. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands. dREG tracks show the distribution of dREG signal. The net-synteny tracks show the fraction of the genomic area that is mappable in the indicated species. The location of transcription units inferred in the common ancestor of human and chimpanzee, and the location of RefSeq gene annotations, are shown at the top. **(d-f)** Scatterplots show the correlation between changes in gene expression (log-2 scale) following π treatment in the species indicated on the axes. Color scale indicates the density of points in the region.



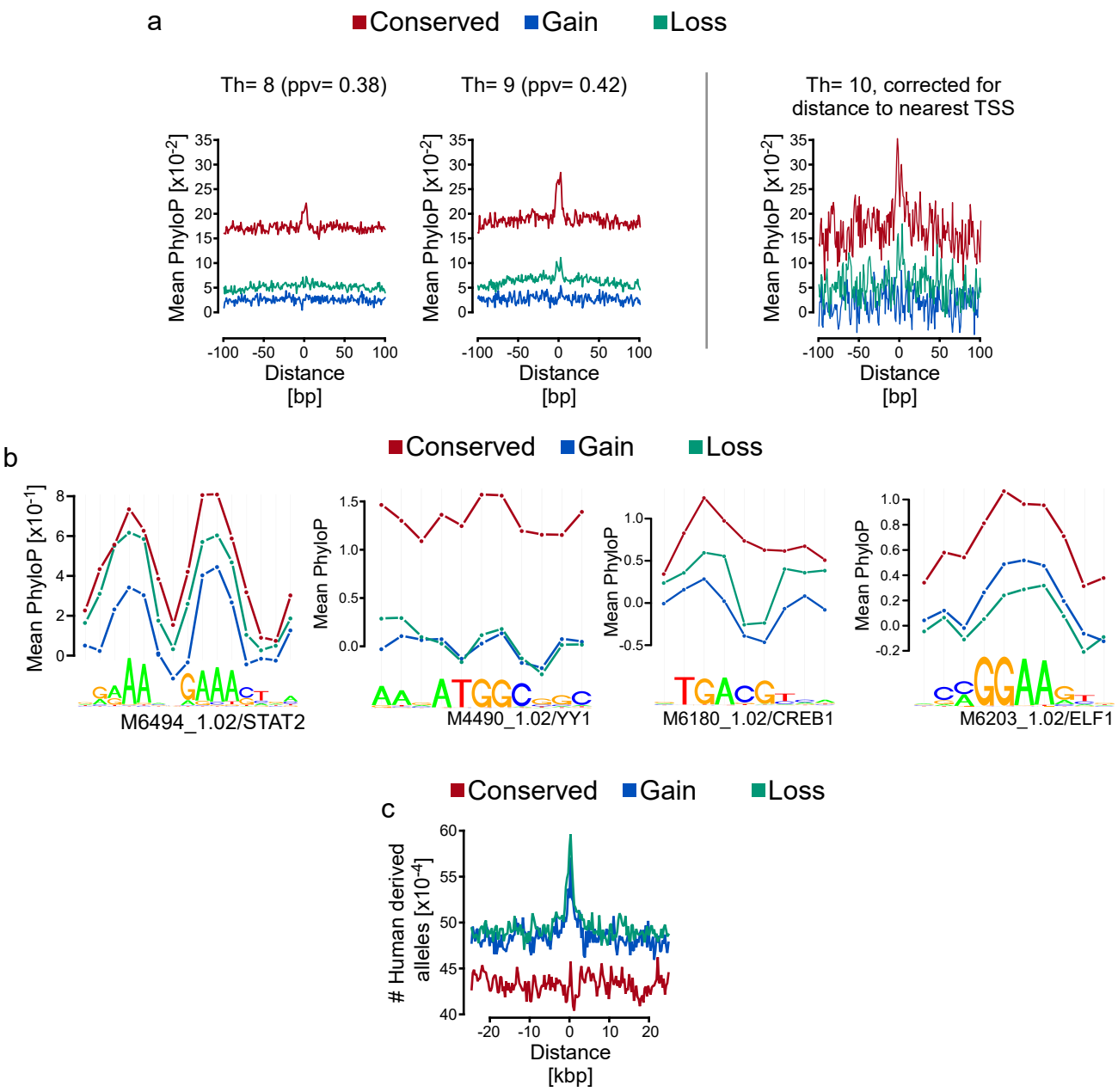
Supplementary Fig. 5 | Evolutionary changes in TREs. **(a)** Venn diagram illustrating raw changes in TREs among primate species. In all cases, TREs were discovered in untreated CD4+ T-cells using dREG (threshold > 0.3). **(b)** Q-Q plot showing observed p-values (deSeq2 in human compared to the other two primate species) among TREs that were not identified by dREG in at least one species (red), all TREs identified (black), and a set of conserved TREs (gray). **(c)** Scatterplot shows the evolutionary divergence time (X-axis) as a function of Spearman’s correlation in gene body transcription between each sample collected in the untreated condition and the mean gene expression in untreated human CD4+ T-cells (Y-axis). The red line shows the best linear fit and dotted lines indicate the 99% confidence interval. We assume the following evolutionary divergence estimates for each species pair with respect to human, 12 MYR for chimp-human [Moorjani et. al. (2016); ref[13]], 25 MYR for human-rhesus [Rogers (2013); ref[14]], and 75 MYR for human-rodent [Chinwalla et. al. (2002); ref[15]].



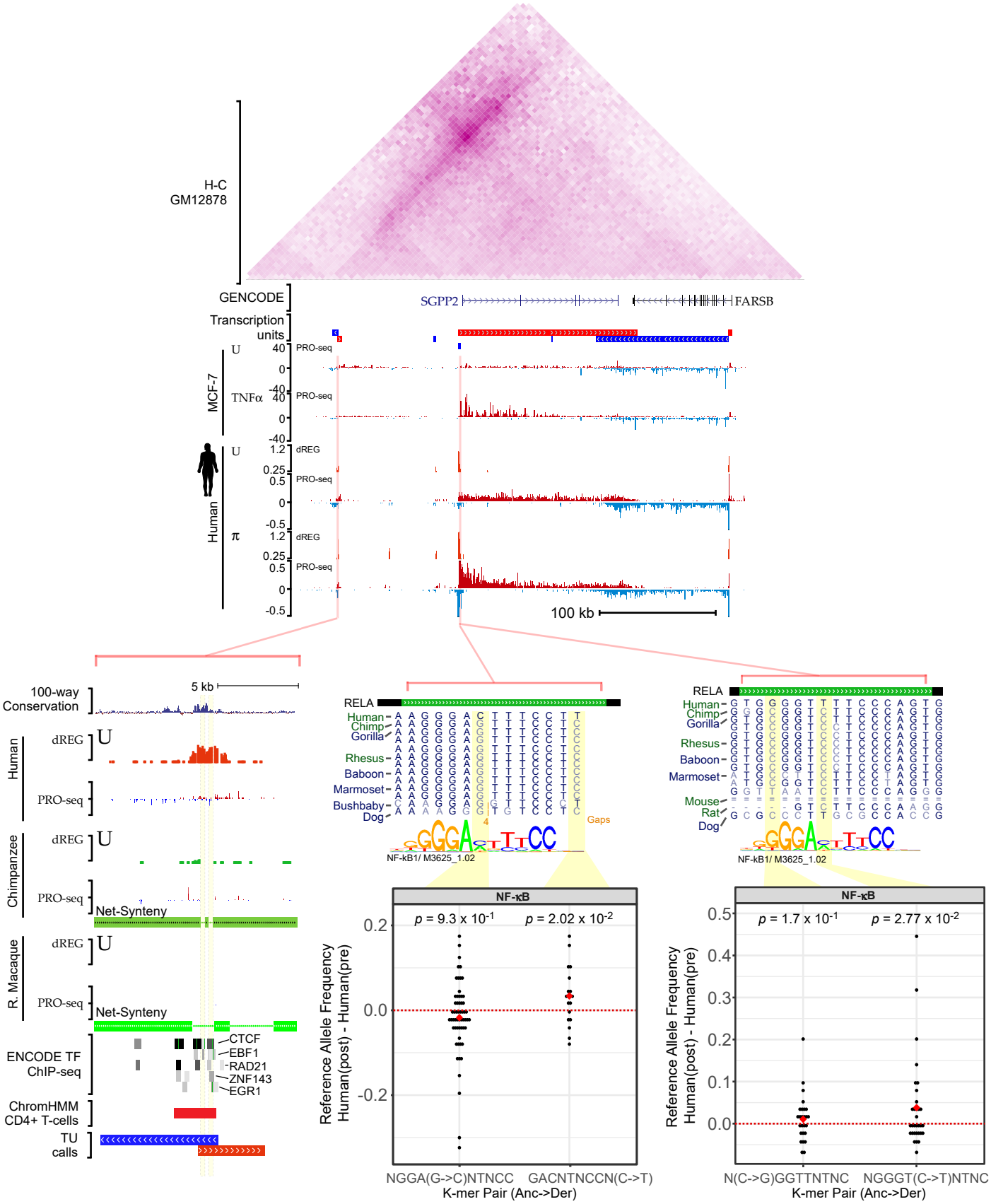
Supplementary Fig. 6 | Evolutionary changes in TREs correlate with chromatin and DNA modifications. ChIP-seq signal for H3K27ac and H3K4me1 near dREG sites classified as gains, losses, or complete losses of TRE signal (dREG score < 0.05) on the human branch.



Supplementary Fig. 7 | Accuracy of dREG and PIQ for detecting transcription factor binding motifs. Positive predictive values (PPV) for dREG (Y-axis) and PIQ (X-axis) for 37 transcription factors. Scores reflect the fraction of true positive motif matches (motif match score >10; see methods). True positive matches were defined by ChIP-seq data in K562 cells.

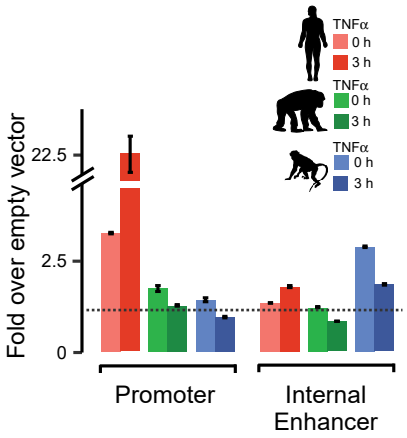


Supplementary Fig. 8 | PhyloP scores in transcription factor (TF) binding motifs. **(a)** Evolutionary conservation centered on matches to a TF binding motif at the indicated cut off score (left), or adjusted for distance to the nearest annotated transcription start site by subsampling (right). **(b)** PhyloP scores that fall within the binding motifs recognized by STAT2 (M6494_1.02), YY1 (M4490_1.02), CREB1 (M6180_1.02), and ELF1 (M6203_1.02). In all cases motifs fall in dREG-HD that are gained (blue) or lost (cyan) on the human branch, or are conserved among all primate species (red). **(c)** The distribution of human derived alleles near dREG sites that are gained (blue) or lost (cyan) on the human branch, or are conserved among all primate species (red).

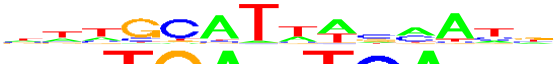
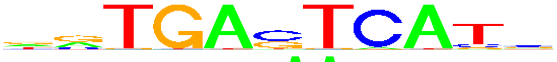
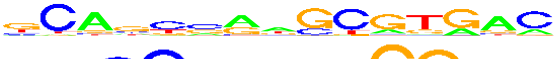
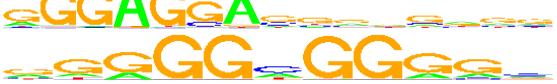
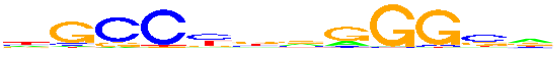


Supplementary Fig. 9 | Candidate causal DNA sequence differences underlying changes in SGPP2 transcription. (See next page for a full legend.)

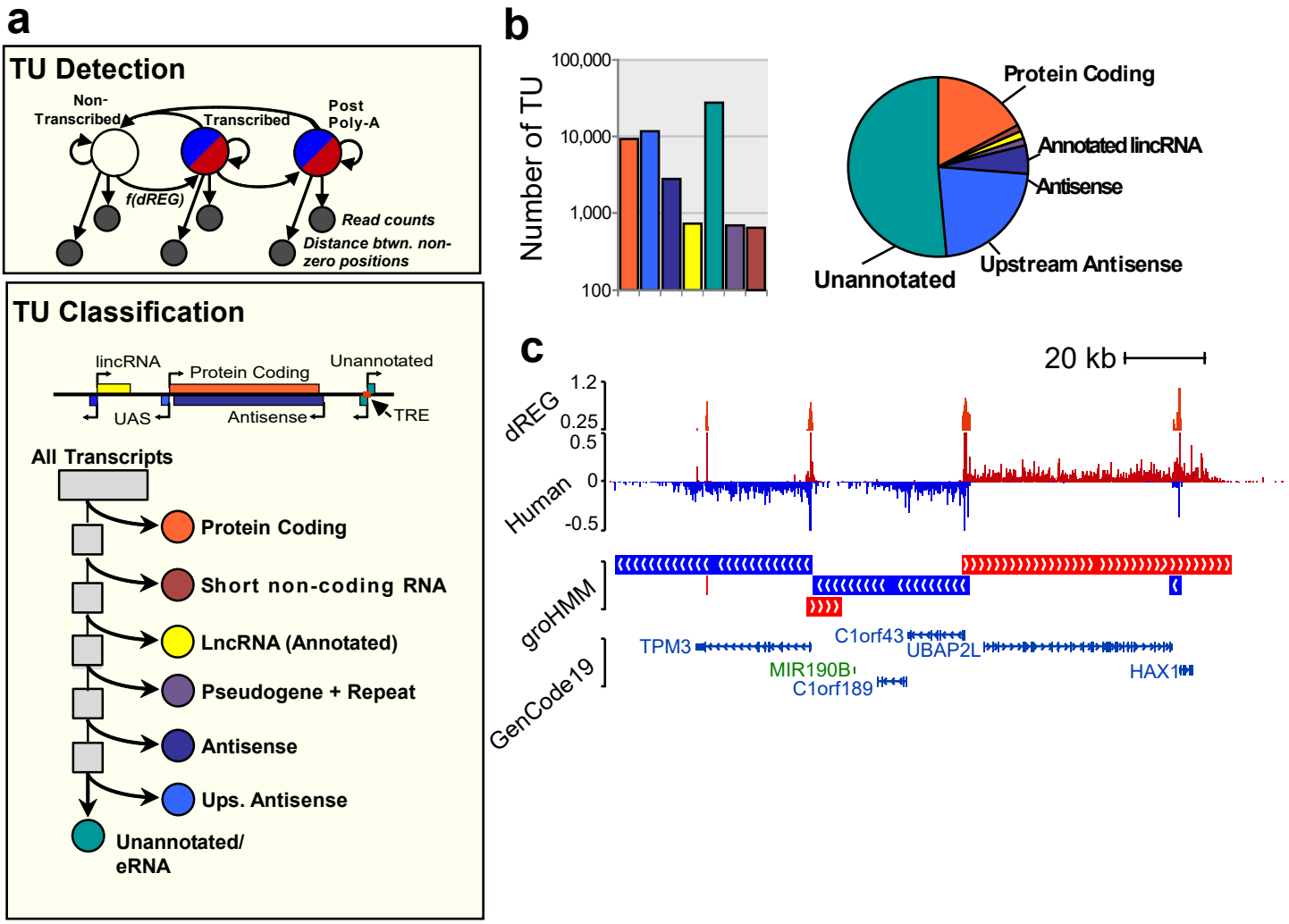
Supplementary Fig. 9 | Candidate causal DNA sequence differences underlying changes in SGPP2 transcription. UCSC genome browser track shows transcription near SGPP2 and FARSB in untreated (U) and PMA+ionomycin (π) treated human CD4+ T-cells or in human MCF-7 cells. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands. Axes for the PRO-seq data are in units of reads per kilobase per million mapped (RPKM) or in raw reads (MCF-7). dREG tracks show the distribution of dREG signal. Heatmap (top) shows Hi-C signal in GM12878 lymphoblastoid cell lines. Insert (bottom) shows lack of orthology in chimpanzee and rhesus macaque in an active TRE (human) that binds a number of TFs in ENCODE cell lines (left) and substitutions in NF-kB binding motifs near SGPP2. Two motif occurrences in the proximal promoter were bound by RELA, a subunit of NF-kB, based on human ChIP-seq data in ENCODE cell lines (green boxes). Positions where human carries a derived allele are indicated by yellow highlights. PRO-seq reads matched the human reference allele in all positions (15/ 15 reads match C and 26/ 26 match the reference T allele in the NF-kB binding site in the promoter; 11/ 11 reads match the G and 11/ 11 match the T reference allele in the NF-kB binding site in the promoter; and 24/ 24 reads support the TG human reference sequence in the internal enhancer). Scatterplots show the relative frequencies of the human allele in RELA (NF-kB) ChIP-seq data matching NF-kB binding QTLs that mimic the human and ancestral alleles, while controlling for the flanking sequence indicated below the plot. The red dot denotes the mean. All four human-specific DNA sequence changes in NF-kB binding motifs in the proximal promoter together show trend toward higher NF-kB binding in human ($p = 0.017$, using Fisher's method to combine p-values).



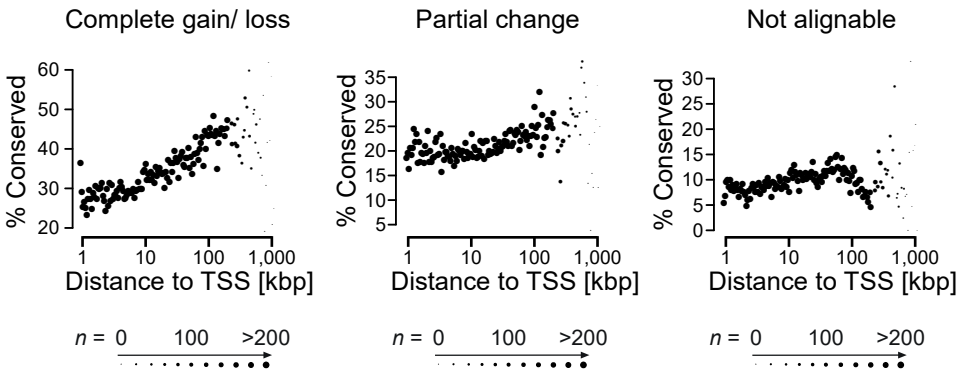
Supplementary Fig. 10 | Luciferase assays for TREs identified near SGPP2. The Y-axis shows the luciferase signal driven by the SGPP2 promoter or the internal enhancer in MCF-7 cells using DNA from each primate species following 3 hours of stimulation with TNFα or vehicle control. Bars show the mean luciferase activity in each species, over the empty vector and renilla controls. Error bars represent the standard error of the mean.

TF	Motif Logo	Sites	Bases	E[A]	# Adaptive Substitutions
FOXJ2		1137	10560	2.987 ± 1.208	32
ZBTB7A		1121	5802	2.901 ± 1.118	17
POU2F1		1392	16544	2.626 ± 0.918	43
FOXO1		1001	8501	2.505 ± 1.439	21
EHF		1338	8517	2.407 ± 0.874	21
ELF1		1148	8862	2.245 ± 0.953	20
SP4		1529	6790	2.172 ± 0.968	15
E2F1		980	2790	1.919 ± 1.479	
POU3F1		942	9106	1.707 ± 1.263	
JUND		943	10669	1.658 ± 0.81	
IRF4		930	11113	1.59 ± 0.81	
EGR1		1378	7413	1.423 ± - 1	
KLF5		1056	3468	1.35 ± 1.917	
FLI1		1064	8349	1.324 ± 0.808	
SP2		1280	5642	0.973 ± - 1	
PAX5		932	7146	0.965 ± - 1	
TFAP2A		943	5800	0.694 ± 1.245	
ZEB1		1100	7623	0.547 ± - 1	
SP3		1503	7211	0.528 ± - 1	
JUN		928	10223	0.481 ± 0.792	
ZNF263		1211	13064	0.417 ± - 1	
SP1		1880	9251	0.319 ± - 1	
GATA1		971	9420	0 ± 0	
TFAP2C		1010	6526	0 ± 0.769	

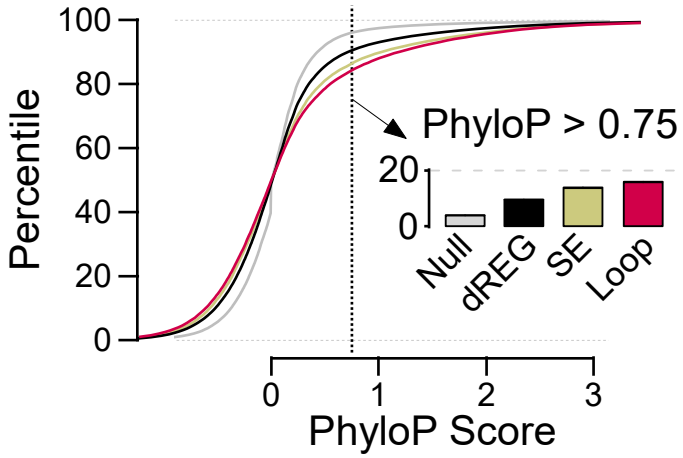
Supplementary Fig. 11 | Adaptive substitutions in specific TF binding motifs. Adaptive substations in TF binding motifs (TFBM) occurring commonly (>900 times) in human lineage-specific dREG-HD sites. Columns denote the TF name annotated in CisBP (TF), number of sites (Sites), the number of bases (Bases), the expected number of adaptive substitutes per kilobase (E[A]), the standard error in the expected substitutions per kilobase (E[A]_stderr), and the estimated number of adaptive substitutions (# Adaptive Substitutions). TFBSs may be bound by any TF that recognizes a similar motif. TFBM in which E[A] is significantly larger than 0 are highlighted in bold fold. The estimated number of adaptive substitutions for each of these sites is shown.



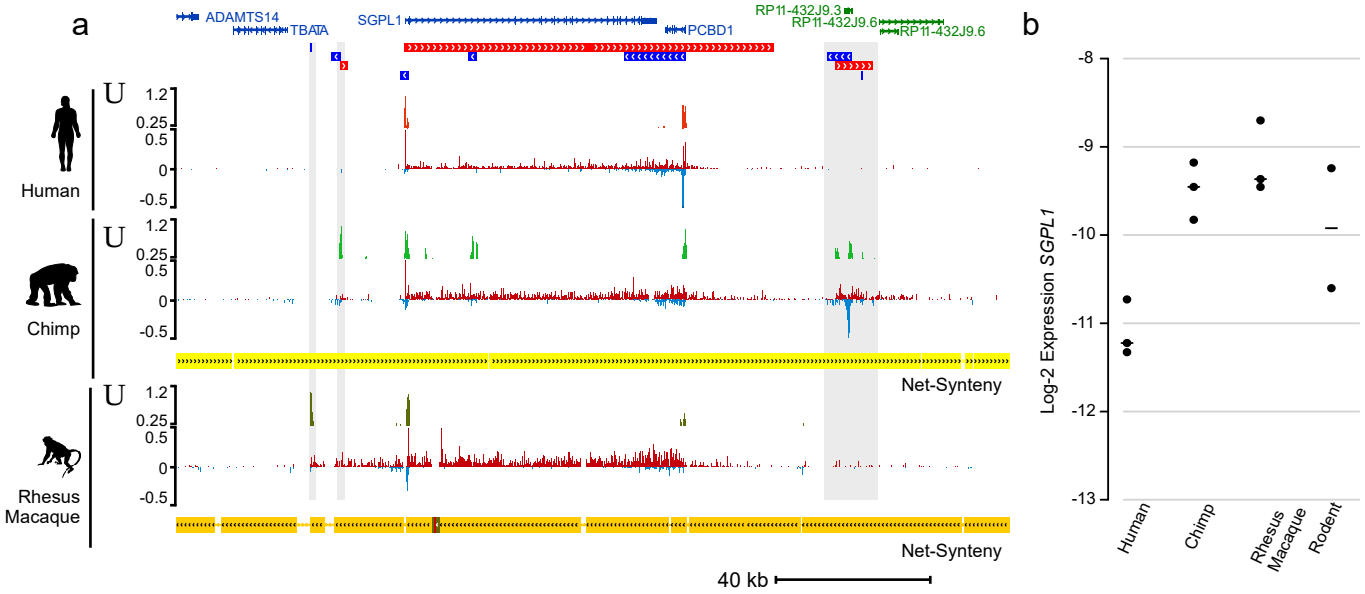
Supplementary Fig. 12 | Discovery of transcription units (TU) in primate T-cells. **(a)** A novel three-state hidden Markov model (HMM) was used to discover transcription units. States correspond to non-transcribed background sequence, transcribed sequence, and post polyA transcription. TUs were classified into one of seven classes as indicated in the cartoon. **(b)** The number and fraction of transcription units that fall into each TU classification. **(c)** Example of the hidden Markov model (HMM) in a typical region. TUs largely agree with RefSeq gene annotations when available.



Supplementary Fig. 13 | DNA sequence conservation as a function of genomic distance to the nearest start site. Scatterplot shows the percentage of TREs undergoing complete gains and losses (left), undergoing a partial change in the abundance of Pol II (center), or that are not alignable between species (right) as a function of distance from the nearest annotated transcription start site (x-axis). The size of each point represents the amount of data in the corresponding distance bin.



Supplementary Fig. 14 | Evolutionary conservation of DNA sequence mirrors functional conservation at looped- and un-looped enhancers. Cumulative distribution function of phyloP scores from the 100-way alignments in the indicated class of dREG site. The insert shows the fraction of sites in each class exceeding a phyloP score cutoff of 0.75.



Supplementary Fig. 15 | Changes at regulatory architecture in the SGLP1 locus. (a) UCSC genome browser track shows transcription near SGLP1 in untreated (U) CD4+ T-cells in the indicated species. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands. Axes for the PRO-seq data are in units of reads per kilobase per million mapped (RPKM). dREG tracks show the distribution of dREG signal. Whereas the chimpanzee has several lineage specific enhancers and a single promoter that is shared with human, the rhesus macaque transcribes a copy of the SGPL1 gene from two alternative promoters, one of which is not found in the other species. Notably, in the human, which lacks both the chimpanzee enhancers and the second promoter, transcription is 2-fold lower than in either of the two non-human primates. This example suggests that changes in enhancers and promoters spanning tens of kilobases can compensate for one another, and can lead to widely divergent transcriptional regulatory architectures in different species. **(b)** Plot shows normalized expression (log2 scale) of SGLP2 in the species indicated below the plot. Each point represents the transcription of the indicated gene in a different untreated T-cell sample. The bar indicates the median in each species.

Human	U	P/I 30 min.	Total (per species)
H1	38,014,810	39,747,353	
H2	38,940,161	72,458,965	
H4	34,191,759	32,460,837	
Total	**111,146,730**	**144,667,155**	**255,813,885**
Chimp			
C3	43,877,724	46,137,893	
C4	44,017,406	35,603,695	
C5	49,891,741	54,423,856	
Total	**137,786,871**	**136,165,444**	**273,952,315**
Rhesus Macaque			
M2	55,626,584	47,939,020	
M3	56,740,558	52,678,876	
M4	31,803,492	27,379,589	
Total	**144,170,634**	**127,997,485**	**272,168,119**
Mouse			
Mm1	82,584,705	-	
Total	**82,584,705**	-	**82,584,705**
Rat			
Rn1	77,799,614	-	
Total	**77,799,614**	-	**77,799,614**
Total=	**962,318,638**		

Supplementary Table 1 | Sequencing of PRO-seq data. PRO-seq data was collected from CD4+ T-cells isolated from five mammalian species in two biological conditions. CD4+ T-cells were analyzed from three individuals representing each of the primate species (human, chimpanzee, and rhesus macaque) in two biological conditions (U – “untreated” and P/I – 30 min. of PMA and Ionomycin). Data was also collected from CD4+ T-cells in the untreated condition isolated from the thymus of a mouse and a rat.

Result	dREG stringent > 0.25	dREG stringent > 0.3	dREG stringent > 0.35	dREG permissive < 0.1	DESeq2 q < 0.1
Precision (human T-cells)	89%	93%	96%	93%	93%
Recall (human T-cells)	85%	83%	80%	83%	83%
Correlation between distance from GENCODE TSS (correlation coef)	-0.73 (p < 2.2E-16)	-0.71 (p < 2.2E-16)	-0.68 (p < 2.2E-16)	-0.61 (p < 2.2e-16)	-0.72 (p < 2.2E-16)
High conservation of looped enhancers (Fisher's exact p-value)	p < 2.2E-16	3.31E-07	2.98E-04	6.39E-05	5.70E-03
Correlation with Capture Hi-C loop strength (correlation coef)	0.71 (p < 1E-3)	0.54 (p = 1E-3)	0.41 (p = 2E-3)	0.52 (p = 1E-3)	0.28 (p = 0.026)
Correlation between loops and number of interactions (correlation coef)	0.88 (p = 2E-3)	0.87 (p = 1E-3)	0.86 (p = 1E-3)	0.88 (p = 1E-3)	0.84 (p = 3E-3)

Supplementary Table 2 | Insensitivity of evolutionary change co-vitiates results to differences in enhancer threshold. We completed the enhancer analysis described in the manuscript at several different cutoff thresholds. Summary statistics for key results denoted in the text are shown in this table.

Assay:	TRE:	Species/ orientation/ notes:	Sequence:
Luciferase	Internal Enhancer	Human Left	GCGGTACCACTCTCCCACTTGTTGGTGC
Luciferase	Internal Enhancer	Chimp Left	GCGGTACCACTCTCCCACTTGTGGGTGC
Luciferase	Internal Enhancer	Rhesus Macaque Right	GCGGTACCACTCTCCCACTCGTTGGTGC
Luciferase	Internal Enhancer	Human Chimp Right	ACACGCGTCAGGTGGGCTTTTCAGTCCT
Luciferase	Internal Enhancer	Rhesus Macaque Right	ACACGCGTCAGGTGGTCTTTTCAGTCCT
Luciferase	Promoter	Human Chimp Left	ACGGTACCTGCAAGTGCCTTCACAGGAA
Luciferase	Promoter	Rhesus Macaque Left	ACGGTACCTGCCAGTGTCTTCACAGGAA
Luciferase	Promoter	Human Chimp Rhesus Macaque Right	GCACGCGTTGCCCTTCTGATTTGGCAAC
Luciferase	SNP1	Human Chimp Left	TAGGTACCAACCACAGGGTTTTGCCTCA
Luciferase	SNP1	Rhesus Left	TAGGTACCAACCACGGGGTTTTGCCTCA
Luciferase	SNP1	Human Chimp Rhesus Macaque Right	CCACGCGTTCTGATTTGGCAACTGGGGA
Luciferase	SNP2	Human Chimp Rhesus Macaque Left	CCGGTACCAAGTCAAAAGTGTGGCGCAG
Luciferase	SNP2	Human Chimp Rhesus Macaque Right	CAACGCGTAGCAACTCACGTGTCTACCTG
CRISPRi	Internal Enhancer	G1	TCCTCTTACTCGCTGCTCACTGG
CRISPRi	Internal Enhancer	G2	GGGTGTAGAATTTCCGTCTGTGG
CRISPRi	Internal Enhancer	G3, overlaps RELA binding site	GTGGAATTTCCTGACCGTGAAGG
CRISPRi	Upstream Enhancer	G4	GACCGTGGTGCACGCGCCGGGGG
CRISPRi	Upstream Enhancer	G5, overlaps RELA binding site	TGACTGCTGTCGCCTGGCGGGGG
CRISPRi	Upstream Enhancer	G6, overlaps ESR1 binding site	GGGTTCCCCCGCCGCCGTAGTGG
CRISPRi	Upstream Enhancer	G7, overlaps EGR1 binding site	CTTCTCGGTCGTTGTCACCTGGG
CRISPRi	Promoter	G11, overlaps RELA binding site	GTGGGAAGGAAAGTCCCTTCTGG
CRISPRi	Promoter	G12, overlaps RELA binding site	ATGTCCTTGGCACTTCCCCGGGG
qPCR	SGPP2, intron 1	SGPP2_In1_In1_F1_BM	ACAGAACTTGGGGCTCACAC
qPCR	SGPP2, intron 1	SGPP2_In1_In1_R1_BM	TCACTGGTGTCGGTCCATAA

Supplementary Table 3 | Sequences used during experimental validation. Table specifies the DNA sequences used to clone genomic DNA from each species in the luciferase assays, sequences of the gRNAs used in the CRISPRi experiments, and the primers in intron 1 of *SGPP2* used to measure transcription level in the RT-qPCR experiments.

Supplementary References

1. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).
2. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46,** 1311–1320 (2014).
3. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
4. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346,** 1007–1012 (2014).
5. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28,** 817–825 (2010).
6. Mosig, M. O. *et al.* A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics* **157,** 1683–1698 (2001).
7. Nettleton, D., Gene Hwang, J. T., Caldo, R. A. & Wise, R. P. Estimating the number of true null hypotheses from a histogram of p values. *JABES* **11,** 337 (2006).
8. Phipson, B. Empirical Bayes modelling of expression profiles and their associations. (2013).
9. Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16,** 222 (2015).
10. Paris, M. *et al.* Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genet.* **9,** e1003748 (2013).
11. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10,** e1004226 (2014).
12. Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* **25,** 167–178 (2015).
13. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proceedings of the National Academy of Sciences* **113,** 10607–10612 (2016).
14. Rogers, J. In transition: primate genomics at a time of rapid change. *ILAR J.* **54,** 224–233 (2013).
15. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).