HHS Public Access

Author manuscript

Nat Ecol Evol. Author manuscript; available in PMC 2018 May 17.

Published in final edited form as:

Nat Ecol Evol. 2018 March ; 2(3): 537–548. doi:10.1038/s41559-017-0447-5.

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], and **Adam Siepel**[4,*]

[1]Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853

[2]Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853

[3]Graduate field of Computational Biology, Cornell University, Ithaca, NY 14853

[4]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

[5]Department of Microbiology & Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853

[6]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

[7]Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390

[8]Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical Center, Dallas, TX 75390

Abstract

How evolutionary changes at enhancers affect the transcription of target genes remains an important open question. Previous comparative studies of gene expression have largely measured the abundance of mRNA, which is affected by post-transcriptional regulatory processes, hence limiting inferences about the mechanisms underlying expression differences. Here we directly measured nascent transcription in primate species, allowing us to separate transcription from post-transcriptional regulation. We used PRO-seq to map RNA polymerases in resting and activated

[*]Address correspondence to: Charles G. Danko, Ph.D., Baker Institute for Animal Health, Cornell University, Hungerford Hill Rd., Ithaca, NY 14853, Phone: 607-256-5620, dankoc@gmail.com. Adam Siepel, Ph.D., Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Rd., Cold Spring Harbor, NY 11724, Phone: 516-367-6922, asiepel@cshl.edu.

Author Contributions

LAC, BAM, CGD, EJR, and ETW performed CD4+ T-cell extraction, validation, and PRO-seq experiments. CGD, ZW, TC, ALM, LAC, and ND analyzed the data. CGD, AS, JTL, WLK, and SAC supervised data collection and analysis. CGD and AS wrote the paper with input from the other authors.

Competing Financial Interests

The authors declare no competing financial interests.

Author Information

PRO-seq data was deposited into the Gene Expression Omnibus database under accession number GSE85337. All data analysis scripts and software are publicly available on GitHub: https://github.com/Danko-Lab/CD4-Cell-Evolution.

CD4+ T-cells in multiple human, chimpanzee, and rhesus macaque individuals, with rodents as outgroups. We observed general conservation in coding and non-coding transcription, punctuated by numerous differences between species, particularly at distal enhancers and non-coding RNAs. Genes regulated by larger numbers of enhancers are more frequently transcribed at evolutionarily stable levels, despite reduced conservation at individual enhancers. Adaptive nucleotide substitutions are associated with lineage-specific transcription, and at one locus, *SGPP2*, we predict and experimentally validate that multiple substitutions contribute to human-specific transcription. Collectively, our findings suggest a pervasive role for evolutionary compensation across ensembles of enhancers that jointly regulate target genes.

Introduction

Following decades of speculation that changes in the regulation of genes could be a potent force in the evolution of form and function[1–3], investigators have now empirically demonstrated the evolutionary importance of gene regulation across the tree of life[4–8]. Evolutionary changes in gene expression are primarily driven by mutations in non-coding DNA sequences, particularly those that bind transcription factors[9]. Accordingly, adaptive nucleotide substitutions at transcription factor binding sites (TFBSs)[4–8,10–12] and gains and losses of TFBSs[13–21] both appear to make major contributions to the evolution of gene expression. These events are believed to modify rate-limiting steps during transcriptional activation[22]. In addition, transcriptional activity is correlated with various epigenomic and structural features, including post-translational modifications to core histones, the locations of architectural proteins such as CTCF, and the organization of topological associated domains. Like TFBSs, these features display general conservation across species, yet do exhibit some variation, which correlates with differences in gene expression[12,20,23–25].

Nevertheless, many open questions remain about the roles of TFBSs, chromatin organization, and posttranscriptional regulation in the evolution of gene expression. Notably, the correlation between differences in TF binding and differences in mRNA abundance is surprisingly low[26–28]. Possible reasons for this discordance include non-functional TF binding[26,27,29], compensatory gains and losses of TFBSs[16,30–33], difficulties associating distal enhancers with target genes[34], and a dependency of TF function on chromatin or chromosomal organization[35]. In addition, some changes in mRNA expression appear to be “buffered” at the post-transcriptional level[36–38]. One reason why it has been difficult to disentangle these contributions to gene expression is that expression is typically measured in terms of the abundance of mRNA, which is subject to posttranscriptional processing[39] and therefore is an indirect measure of the transcription of genes by RNA polymerase II.

Here we use Precision Run-On and sequencing (PRO-seq)[40] to directly measure transcription and map the location of active regulatory elements, including distal enhancers[41], in CD4+ T-cells. We found that the rates at which regulatory elements are gained and lost over evolutionary time correlates with the distance to the nearest transcription start site and the number of chromatin interactions detected in chromatin conformation capture data. Surprisingly, transcription levels in genes are more stable over evolutionary time when the gene is regulated by larger numbers of distal enhancers, yet in

this case each enhancer exhibits more freedom to diverge. These findings suggest a critical role for redundancy and compensation in the evolution of ensembles of enhancers that jointly determine expression at target genes.

Results

Patterns of Transcription in Resting and Activated CD4+ T-cells

We developed nucleotide-resolution maps of RNA polymerase in CD4+ T-cells isolated from five mammalian species. Samples were collected under resting and activated conditions from three unrelated individuals representing each of three primate species—humans, chimpanzees, and rhesus macaques—spanning ~25–30 million years of evolution (MYR) (Fig. 1a). To compare with studies that focus on longer evolutionary branch lengths, we also collected resting samples from a single individual in each of two rodent species—mouse and rat—which together serve as an outgroup to the primates (~80 MYR divergence). We used flow cytometry to validate the purity of isolated CD4+ cells (Supplementary Fig. 1). In addition, we used measurements of transcriptional activity of T-cell subset markers for T-helper type 1 (Th1), Th2, Th17, T-regulatory, and T-follicular helper cells to demonstrate that the population of CD4+ T-cell subsets within the total CD4+ population was largely similar across species (Supplementary Fig. 2). PRO-seq[40,42] libraries were sequenced to a combined depth of ~1 billion uniquely mapped reads (~78–274 million per species) (Supplementary Table 1). Hierarchical clustering and principal component analysis of these data grouped the primate samples first by cell type or treatment condition and subsequently by species (Fig 1b; Supplementary Fig. 3).

To gain further insight into the evolution of the response to CD4+ T-cell stimulation, we compared transcriptional activity under resting and activated conditions within and between species. In humans, we found that PMA and ionomycin (π) significantly altered the transcription levels of 6,940 (13%) GENCODE-annotated transcription units (TUs) ($p < 0.01$, DESeq2[43]; Fig. 1c). Parallel analyses in chimpanzee and rhesus macaque revealed many similarities in transcriptional changes following π treatment (Supplementary Fig. 4). We identified a core set of 3,157 TUs that undergo evolutionarily conserved transcriptional changes in all three species following 30 min. of π-treatment, including many of the classical response genes (e.g., IFNG, TNFα, IL2, and IL2RA), as well as numerous novel genes and lincRNAs. Active transcriptional regulatory elements (TREs) undergoing changes following π-treatment were enriched for a similar set of transcription factor binding motifs across species, including those for NF-kB and the AP-1 heterodimers FOS and JUN, which are known to be activated by canonical T-cell receptor signaling (Fig. 1d). Thus, the core regulatory principles responsible for T-cell signaling and activation appear to remain broadly conserved across primate evolution.

Rapid Evolutionary Changes in Transcribed Enhancers

We used dREG to identify 30,357 active TREs in human CD4+ T-cells, based on patterns of enhancer-templated RNA (eRNA) or upstream antisense (uaRNA) transcription evident from PRO-seq data. We classified these predicted TREs as either protein-coding promoters or candidate enhancers based on their proximity to gene annotations. The predicted TREs in

each group were highly concordant with other marks of regulatory function in human CD4+ T-cells used previously to define groups of candidate enhancers, including acetylation of histone 3 lysine 27 (H3K27ac), mono- and trimethylation of histone 3 lysine 4 (H3K4me1 and H3K4me3), and DNase-I-seq signal (Fig. 1e; Supplementary Note 1).

Extending our dREG analysis to untreated CD4+ T-cells from additional species revealed 71,748 TREs that were active in untreated T-cells in at least one species (ranging between 27,581 and 39,387 TREs in each species). We defined two types of changes between species: (1) changes in the abundance of Pol II at TREs that were present across all species, and (2) complete gains or losses in at least one species (see Supplementary Note 2; Supplementary Fig. 5). We found that 52% of distal TREs (henceforth referred to simply as "enhancers") showed evidence of transcriptional changes in at least one of the three primate species and 81% showed changes at the longer evolutionary distance between primates and rodents (Fig. 2a), similar to recent observations in other systems[20,44].

Next we tested whether evolutionary changes in transcriptional activity correlate with the enrichment of other marks of active enhancers. Predicted lineage-specific human enhancers were enriched for both active and repressive enhancer marks (Fig. 2b; Supplementary Fig. 6). Whereas apparent human gains were enriched for high levels of the active enhancer marker H3K27ac, sites with reduced transcriptional activity in humans showed much lower enrichments of H3K27ac. Furthermore, locations at which the dREG signal was completely lost in a human-specific fashion displayed levels of H3K27ac approaching those of randomly selected background sites (Fig. 2b). Intriguingly, many of the losses on the human branch retained H3K4me1, which marks both active and inactive enhancers[45], and these losses displayed higher levels of chromatin marks associated with transcriptional repression than a random background (Fig. 2b), indicating that active ancestral primate enhancers often retain a 'poised' chromatin state in human, despite losing both transcriptional activity and H3K27ac.

Transcriptional Changes Correlate with DNA Sequence Differences

To investigate whether changes in TRE activity are accompanied by changes in DNA sequence, we compared phyloP sequence conservation scores at transcriptionally conserved TREs with phyloP scores at TREs that display evolutionary changes in transcription. We restricted our sequence conservation analyses to TFBSs and selected a TFBS match threshold at which 60% of binding sites were expected to be bona fide TFBSs, as measured by ChIP-seq (positive predictive value [PPV] = 0.60; Supplementary Fig. 7). TFBSs found in transcriptionally conserved dREG sites showed a marked enrichment for higher phyloP scores relative to surrounding regions, indicating local sequence conservation (Fig. 3a). By contrast, TFBSs in lineage-specific dREG sites had substantially reduced enrichments in phyloP scores (Fig. 3a, cyan/blue). Notably, TFBSs in dREG sites lost on the human lineage showed enhanced conservation compared with those in human-specific gains. This observation is consistent with losses evolving under conservation in other mammalian species (which contribute to the phyloP scores) and gains emerging relatively recently. Each of these patterns was robust to corrections for potentially confounding differences in the distribution of sites, as well to choices of motif score thresholds (Supplementary Fig. 8a).

Relaxing the motif score threshold to provide sensitivity for larger numbers of TFBSs at the expense of specificity, revealed patterns of conservation that correlate with the information content of positions within the DNA sequence motif (Supplementary Fig. 8b), further supporting TF binding as the functional property driving sequence conservation at these sites.

We searched for examples of DNA sequence differences that might be responsible for transcriptional changes following π treatment, hypothesizing that causal alleles might be characterized more easily than in the untreated condition (Fig. 1d). In one example, we found nucleotide substitutions in three apparent NF-kB binding sites in the proximal promoter and an internal enhancer of *SGPP2* that correlate with differences in *SGPP2* expression (Fig. 3b; Supplementary Fig. 9). Two of these TFBSs were bound by NF-kB in human cell lines according to ChIP-seq data from ENCODE. Moreover, substitutions were either located in core positions of the DNA sequence motif (Fig. 3b) or were found to disrupt the same position in the motif as NF-kB binding QTLs (Supplementary Fig. 9). All DNA sequence changes observed showed a trend toward higher predicted NF-kB binding affinity in human than non-human primates. To test the hypothesis that observed DNA sequence changes produced differential transcriptional activity, we cloned DNA from each primate species into a reporter vector driving luciferase activity in an MCF-7 cell model, which recapitulates the primary transcriptional features of the *SGPP2* locus (Supplementary Fig. 9). Differences in basal luciferase activity were generally concordant with those observed between species (Supplementary Fig. 10). Moreover, both the proximal promoter of *SGPP2* and the internal enhancer both activated luciferase expression more strongly following NF-kB activation with cloning of human DNA, but not with orthologous DNA from the other primates (Fig. 3c).

To determine whether these TREs affect the expression of *SGPP2* in its native genomic context, we silenced each TRE by using CRISPRi, which targets a catalytically dead CAS9 fused to the Krüppel-associated box repressor (dCAS9-KRAB), to specifically tri-methylate lysine 9 of histone 3 (H3K9me3). Three independent single-guide RNAs (sgRNAs) targeting the internal enhancer and two designed for the proximal promoter reduced *SGPP2* transcription to 50–60% of its resting level (p = 1.5e-3 and 2.6e-2, two-tailed t-test; Fig. 3d), consistent with these TREs directly contributing to *SGPP2* expression in MCF-7 cells. Three sgRNAs targeting the upstream enhancer also had a significant effect on *SGPP2* expression (p = 1.8e-4, two-tailed t-test). Notably, the genome assemblies for chimpanzee and rhesus macaque harbor deletions that affect multiple TFBSs in this upstream TRE (Supplementary Fig. 9). However, although silencing individual enhancers reduced the transcription level of *SGPP2* following NF-kB activation, it was insufficient to completely abolish induction of *SGPP2.* Taken together, our findings show that at least two of the three TREs regulating *SGPP2* drove expression patterns matching PRO-seq data in a reporter assay, but none completely explained *SGPP2* activation *in situ*. These observations suggest that that multiple causal substitutions in NF-kB binding sites may work in concert to achieve human-specific activation of *SGPP2*.

Human-Specific TREs Appear to be Evolving Under Positive Selection

In many cases, as with *SGPP2* (Fig. 3b), we observed numerous nucleotide substitutions within individual or clustered TFBSs. These clusters of substitutions are highly unlikely to occur by chance and suggest that positive selection may have driven adaptation of these binding sites. To more directly measure the impact of positive selection, we used INSIGHT[46] to compare patterns of within-species polymorphism and between-species sequence divergence in TREs that had undergone human lineage-specific transcriptional changes. This analysis indicated that although dREG sites overall are most strongly influenced by weak negative selection (Fig. 3e), TREs with lineage-specific transcriptional changes in human are strikingly enriched for adaptive nucleotide substitutions ($p < 0.01$ INSIGHT likelihood ratio test; Fig. 3e). We estimate a total of at least 121 adaptive substitutions since the human/chimpanzee divergence within TFBSs that undergo transcriptional changes in human CD4+ T-cells. Despite limited power to detect the specific contributions of many individual TFs at our stringent motif match score threshold, we did note significant excesses of putatively adaptive substitutions in the predicted binding sites of several TFs, including motifs recognized by forkhead box family, POU-domain containing, and ELF/ETS family (Supplementary Fig. 11; $p < 0.01$, INSIGHT likelihood ratio test).

Rates of Enhancer Evolution Vary with Evidence for Gene Interactions

Despite an overall positive correlation between transcription at distal TREs and genes (Fig. 4, Supplementary Fig. 12, discussed in Supplementary Note 3), transcription at enhancers evolves rapidly and is frequently unaccompanied by transcriptional changes at nearby protein-coding genes. For example, *CCR7* transcription is highly conserved among both primate and rodent species in spite of several apparent changes in enhancer activity within the same locus (Fig. 5a). One possible explanation for this disparity is that many predicted distal enhancers may actually have at most weak effects on the transcription of a target gene, and therefore be under reduced evolutionary constraint. If this hypothesis is true, it should be possible to identify subsets of predicted distal enhancers that have stronger effects on transcription than others, and therefore are more conserved over evolutionary time.

We searched for genomic features correlated with conservation of transcription at enhancers, focusing on untreated CD4+ T-cells in order to leverage the large amount of public data available for this cell type. Not surprisingly, one of the features most strongly correlated with transcriptional conservation at enhancers is the distance from the nearest transcription start site of a protein-coding gene (Fig. 5b). More than half of enhancers located within 10 kbp of an annotated TSS are shared across all three primate species, whereas for distal enhancers located between 100 kbp to 1 Mbp from a TSS that fraction drops to roughly a third. This relationship is driven by lineage-specific gains or losses of enhancer activity, and to a lesser extent by changes in TRE activity levels, rather than by differences in the alignability of orthologous DNA (Supplementary Fig. 13).

These simple distance-based observations, however, ignore the critical issue of chromatin interactions between enhancers and promoters. To account for such loop interactions, we extracted 6,520 putative TRE interactions from Chromatin Interaction Analysis with Paired End Tag sequencing (ChIA-PET) data recognizing loops marked with H3K4me2 in human

CD4+ T-cells. We found that 55% of enhancers that participate in these loops were conserved between primate species compared to only 47% of non-looped enhancers (Fig. 5c; p = 5.6e-4, Fisher's exact test). Moreover, higher transcriptional conservation at looped enhancers does not depend on the distance to the transcription start site. Parallel analysis of promoter-capture Hi-C data revealed that the strength of chromatin interaction was correlated with evolutionary conservation of distal TREs, corroborating the result obtained using ChIA-PET (p < 1e-3, bootstrap test). We observed similar levels of conservation at recently defined super-enhancers, although this conservation may simply reflect an enrichment for loop interactions (48% of TREs in super-enhancers loop according to ChIA-PET, compared to 15% of all TREs). Looped enhancers were also enriched for elevated phyloP scores relative to either non-looped enhancers or randomly selected DNA sequences (Supplementary Fig. 14; phyloP > 0.75; p < 2.2e-16, Wilcoxon Rank Sum Test).

Enhancer-Promoter Interactions Contribute to Constraint on Gene Transcription Levels

Another possible explanation for the differences in the rates of enhancer and promoter evolution is that stabilizing selection on transcription levels drives enhancers to compensate for one another as they undergo evolutionary flux. Examination of specific loci revealed several interesting examples where widely different strategies appeared to drive consistent levels of transcription in distinct combinations of species. In one example, *SGPL1* is transcribed at similar levels in chimpanzee and rhesus macaque, but both species activate transcription of *SGPL1* in a distinct manner (Supplementary Fig. 15). We more commonly observed species-specific changes in enhancer activities at densely populated loci, as in the case of *CCR7* (Fig. 5a).

We therefore hypothesized that redundancy in the set of enhancers associated with a target gene might enable compensation during enhancer evolution. Indeed, we found that evolutionary conservation of promoter TRE transcription is remarkably strongly correlated with the number of loop interactions a promoter has with distal sites (Fig. 6a, weighted Pearson's correlation = 0.87; p < 1e-3 by a bootstrap test). A similar trend was observed between the number of loop interactions made by a target promoter and DNA sequence conservation in putative TFBSs at the promoter, although the effect was weaker and did not meet our criteria for statistical significance (Fig. 6b, weighted Pearson's correlation = 0.65; p = 0.07 bootstrap test).

But how does redundancy in enhancers relate to the evolutionary conservation of the enhancers themselves? If redundant enhancers compensate for one another, perhaps each one will be less, rather than more, conserved when the associated promoters have larger numbers of loop interactions. To address this question, we examined the rate of conservation of looped enhancers as a function of the number of loops in which their gene-proximal partners participated. We found that DNA sequence conservation in putative TFBSs negatively correlates with the number of loops at the proximal end (Fig. 6d; weighted Pearson's correlation = −0.80; p = 2e-3 bootstrap test). We noted a similar trend toward a negative correlation between the conservation of distal TRE transcription and the number of loop interactions (weighted Pearson's correlation = −0.67; p = 0.059, bootstrap test, Fig. 6c). These results suggest that each associated distal TFBS is individually less essential at genes

having larger numbers of loop interactions with distal sites, and they are therefore consistent with a model in which such TFBSs are more freely gained and lost during evolution. Taken together, our results imply that distance, looping, and redundancy of enhancers all contribute to constraints on the evolutionary rates of changes in gene transcription.

Discussion

Observations made recently across a number of biological systems[13–21] have demonstrated that changes in distal TREs arise surprisingly rapidly during the course of evolution, at much faster rates than evolutionary changes in protein-coding genes. However, the available data have not allowed these discordances in evolutionary rate to be explained. Do these disparities reflect compensation at the level of RNA stability, as has been observed at the translational level[36], compensatory changes in transcriptional regulation, or are other factors involved? By making use of direct measurements of primary transcription and excluding the confounding effects of mRNA stability, our work strongly suggests that the effects of evolutionary changes in enhancers are buffered at the transcriptional level, most likely by compensatory changes at other enhancers.

Several lines of evidence in our study suggest that many apparent distal enhancers do not have a strong effect on gene expression, possibly explaining the rapid changes in TREs that are less likely to interact with protein-coding genes. In particular, we found that enhancer conservation is stratified by distance, in both one-dimensional genomic coordinates (Fig. 5b) and based on interactions in chromosome conformation capture data (Fig. 5c), relative to genes. It is unlikely that these results can be explained by false positive or false negative TRE calls, as they are also supported by a more conservative approach based on raw PRO-seq read counts (Supplementary Table 2). The higher levels of conservation observed for TREs that are found near genes suggest that these TREs have a disproportionately large effect on organism fitness, and therefore are likely to more directly regulate the transcription of critical protein-coding genes.

An alternative (but not mutually exclusive) explanation for differences in the rates of evolution between enhancers and promoters is that enhancers frequently compensate for one another as they undergo evolutionary flux, whereas promoters are less labile. Our observations suggest that such compensatory evolution within ensembles of enhancers is a surprisingly widespread feature of mammalian genomes. We find that such compensation frequently arises at the locus-level by changes in enhancers spread over genomic distances spanning tens to hundreds of kilobases and communicating through TRE-TRE loop interactions. This finding is supported by the strong correlation between the probability of conservation in the transcription of gene promoters (Fig. 6a) or their DNA sequences (Fig. 6b) and the number of loop interactions with distal sites. These observations are also compatible with those made very recently in liver tissue[47]. In addition, we find a strong correlation in the opposite direction at the distal end of loops, where promoters that form more loop interactions have enhancers that are less, rather than more, conserved at the level of both transcription (Fig. 6c) and DNA sequence (Fig. 6d). While this inverse correlation is at first counterintuitive, on further reflection it also supports the hypothesis of widespread enhancer compensatory turnover, because it suggests lower conservation for those TFBSs

that are part of large ensembles of regulatory elements than for those that act individually or in small numbers. This finding of widespread redundancy in the regulatory architecture of primate genes potentially has broad implications for our understanding of the general mode and tempo of regulatory evolution.

Online Methods

Multiple species PRO-seq library generation

Isolation of primate CD4+ T-cells—All human and animal experiments were done in compliance with Cornell University IRB and IACUC guidelines. We obtained peripheral blood samples (60–80 mL) from healthy adult male humans, chimpanzees, and rhesus macaques. Informed consent was obtained from all human subjects. To account for within-species variation in gene transcription we used three individuals to represent each primate species. Blood was collected into purple top EDTA tubes. Human samples were maintained overnight at 4C to mimic shipping non-human primate blood samples. Blood was mixed 50:50 with phosphate buffered saline (PBS). Peripheral blood mononuclear cells (PBMCs) were isolated by centrifugation (750× g) of 35 mL of blood:PBS over 15 mL Ficoll-Paque for 30 minutes at 20C. Cells were washed three times in ice cold PBS. CD4+ T-cells were isolated using CD4 microbeads (Miltenyi Biotech, 130-045-101 [human and chimp], 130-091-102 [rhesus macaque]). Up to 10^8 PBMCs were resuspended in binding buffer (PBS with 0.5% BSA and 2mM EDTA). Cells were bound to CD4 microbeads (20uL of microbeads/10^7 cells) for 15 minutes at 4C in the dark. Cells were washed with 1–2 mL of PBS/BSA solution, resuspended in 500uL of binding buffer, and passed over a MACS LS column (Miltenyi Biotech, 130-042-401) on a neodymium magnet. The MACS LS column was washed three times with 2mL PBS/BSA solution, before being eluted off the neodymium magnet. Cells were counted in a hemocytometer.

Isolation of CD4+ T-cells from mouse and rat—Spleen samples were collected from one male mouse (FVB) and one male rat (Albino Oxford) that had been sacrificed for IACUC-approved research not related to the present study. Dissected spleen was mashed through a cell strainer using a sterile glass pestle and suspended in 20 mL RPMI-1640. Cells were pelleted at 800xg for 3 minutes and resuspended in 1–5mL of ACK lysis buffer for 10 minutes at room temperature to lyse red blood cells. RPMI-1640 was added to a final volume 10 times that used for ACK lysis (10–40 mL). Cells were pelleted at 800xg for 3 minutes, counted in a hemocytometer, and resuspended in RPMI-1640 to a final concentration of 250,000 cells per ml. CD4+ T-cells were isolated from splenocytes using products specific for mouse and rat (Miltenyi Biotech, 130-104-453 [mouse], 130-090-319 [rat]) following instructions from Miltenyi Biotech, and as described above.

T-cell treatment and PRO-seq library generation—CD4+ T-cells were allowed to equilibrate in RPMI-1640 supplemented with 10% FBS for 2–4 hours before starting experiments. Primate CD4+ T-cells were stimulated with 25ng/mL PMA and 1mM Ionomycin (P/I or π) or vehicle control (2.5uL EtOH and 1.66uL DMSO in 10mL of culture media). We selected the minimum concentrations which saturate the production of IL2 and IFNG mRNA after 3 hours of treatment. A 30 min. treatment duration was selected after

observing a sharp increase in ChIP-qPCR signal for RNA Pol II phosphorylated at serine 5 on the C-terminal domain on the IFNG promoter at 30 min. To isolate nuclei, we resuspended cells in 1 mL lysis buffer (10 mM Tris-Cl, pH 8, 300 mM sucrose, 10 mM NaCl, 2 mM MgAc2, 3 mM CaCl2 and 0.1% NP-40). Nuclei were washed in 10 mL of wash buffer (10 mM Tris-Cl, pH 8, 300 mM sucrose, 10 mM NaCl and 2 mM MgAc2) to dilute free NTPs. Nuclei were washed in 1 mL, and subsequently resuspended in 50 μL, of storage buffer (50 mL Tris-Cl, pH 8.3, 40% glycerol, 5 mM MgCl2 and 0.1 mM EDTA), snap frozen in liquid nitrogen and kept for up to 6 months before making PRO-seq libraries. PRO-seq libraries were created exactly as described previously[40]. In most cases, we completed library preps with one member of each species (usually one human, chimpanzee, and rhesus macaque) to prevent batch effects from confounding differences between species. Samples were sequenced on an Illumina Hi-Seq 2000 or NextSeq500 at the Cornell University Biotechnology Resource Center.

Mapping PRO-seq reads—We mapped PRO-seq reads using standard informatics tools. Our PRO-seq mapping pipeline begins by removing reads that fail Illumina quality filters and trimming adapters using cutadapt with a 10% error rate[48]. Reads were mapped with BWA[49] to the appropriate reference genome (either hg19, panTro4, rheMac3, mm10, or rn6) and a single copy of the Pol I ribosomal RNA transcription unit (GenBank ID# U13369.1). Mapped reads were converted to bigWig format for analysis using BedTools[50] and the bedGraphToBigWig program in the Kent Source software package[51]. The location of the RNA polymerase active site was represented by the single base, the 3′ end of the nascent RNA, which is the position on the 5′ end of each sequenced read. After mapping reads to the reference genome, three samples (one human, U and PI, one chimpanzee, U and PI, and one rhesus macaque, U and PI) were identified as having poor data quality on the basis of the number of uniquely mapped reads, and were excluded from downstream analysis.

Mapping 1:1 orthologs between different species

During all comparative analyses, the genomic coordinates of mapped reads, dREG scores, and other parameters of interest were converted to the human assembly (hg19) using CrossMap[52]. We converted genomic coordinates between genome assemblies using reciprocal-best (rbest) nets[53]. Reciprocal-best nets have the advantage that comparisons between species are constrained to 1:1 orthologues. This constraint on mapping is enforced by requiring each position to map uniquely in a reciprocal alignment between the human reference and the other species in the comparison. We downloaded rbest nets for hg19-mm10, hg19-panTro4, hg19-rn6 from the UCSC Genome Browser. We created rbest nets for hg19-rheMac3 using the doRecipBets.pl script provided as part of the Kent Source software package.

Analysis of transcriptional regulatory elements

Defining a consensus set of transcriptional regulatory elements—We predicted TREs using dREG[41] separately in each species' reference genome. dREG uses a support vector regression model to score each site covered in a PRO-seq dataset based on its resemblance to features associated with transcription start sites in a reference training dataset. The dREG model was trained to recognize DNase-I-hypersensitive sites that also

show substantial evidence of GRO-cap data in six PRO-seq or GRO-seq datasets measuring transcription in resting K562 cells. dREG scores were computed in the reference genome of each species in order to provide as much information as possible on the native context of each locus. In all cases, we combined the reads from all individuals for each species in order to maximize power for the discovery of TREs. In the primate species, treated and untreated CD4+ T-cells were analyzed separately.

We then defined a consensus set of TRE annotations, each of which bore the signature of an active TRE in at least one species and treatment condition. To define such a set, dREG scores were first converted to human reference genome (hg19) coordinates using CrossMap and the reciprocal-best nets. The advantage of converting dREG scores between the reference genome is that individual bases transfer more completely than genomic intervals using CrossMap and related tools. We then identified TREs in each species separately by thresholding the dREG scores. In all analyses, we selected a threshold of 0.3, which corresponds to a predicted false discovery rate of <7% compared with other sources of genomic data in human CD4+ T-cells. In addition, parallel analyses at separate thresholds (0.25 and 0.35) provided results that were in all cases consistent with those reported in the main manuscript (Supplementary Table 2). The set of overlapping TREs from each species were reduced to a single element containing the union of all positions covered by the set using bedops, and sites within 500 bp of each other were further merged. We assigned each putative TRE the maximum dREG score for each species and for each treatment condition.

Identifying differences in TREs between species—Differences in TRE transcription in 3-way (human-chimp-rhesus macaque) or 5-way (human-chimp-rhesus macaque-mouse-rat) species comparisons were identified using a combination of heuristics and statistical tests. Starting with the consensus set of TREs in hg19 coordinates, we first excluded potential one-to-many orthologs, by eliminating TREs that overlapped gaps in the reciprocal-best nets that were not classified as gaps in the standard nets. The remaining TREs were classified as unmappable when no orthologous position was defined in the rbest nets. Complete gains and losses were defined as TREs that were mappable in all species and for which the dREG score was less than 0.05 in at least one species and greater than 0.30 in at least one other species (see Supplementary Note 1). Gains and losses were assigned to a lineage based on an assumption of maximum parsimony under the known species phylogeny. We defined a set of TREs that displayed high-confidence changes in activity by comparing differences in PRO-seq read counts between species using deSeq2 and thresholding at a 1% false discovery rate (as described below). Changes in TRE activities were compared to histone modification ChIP-seq, DNase-I-seq, and DNA methyl immunoprecipitation data from the Epigenome Roadmap project[54].

TRE classification—For some analyses, TREs were classified as likely promoters or enhancers on the basis of their distance from known protein-coding gene annotations (GENCODE v.19). TRE classes of primary interest include (see also Supplementary Fig. 7): (1) Promoters: near an annotated transcription start site (<100 bp); (2) Enhancers: distal to an annotated transcription start site (>5,000 bp)

Covariates that correlate with TRE changes—We compared the frequency at which evolutionary changes in transcription occur at TREs in a variety of different genomic contexts. We examined the rate of change as a function of distance from the nearest annotated transcription start site in GenCode v.19. TREs were binned by distance in increments of 0.02 on a log10 scale and we evaluated the mean rate at which evolutionary changes in TRE transcription arise in each bin. We also compared the rate of changes in TRE transcription across a variety of functional associations, including loop interactions, within the same topological associated domain, and in super-enhancers. H3K4me2 ChIA-PET data describing loop interactions were downloaded from the Gene Expression Omnibus (GEO) database (GSE32677) (ref.[55]) and the genomic locations of loops were converted from hg18 to hg19 coordinates using the liftOver tool. We also analyzed a separate dataset profiling loop interactions based on promoter capture Hi-C data in human CD4+ T-cells taken from the supplementary materials of ref.[56]. Topological associated domains (TADs) based on Hi-C data for GM12878 cells were also downloaded from GEO (GSE63525) (ref.[57]). Super-enhancers in CD4+ T-cells were taken from the supplementary data for ref.[58]. In all cases we excluded sites with potential one-to-many orthology in any of the species included in the comparison (typically just the three primates). Potential one-to-many orthologs were defined based on differences in the standard and reciprocal-best nets for each species pair.

Refining the location of active TREs using dREG-HD—During analyses of transcription factor binding motifs we further refined the location of TREs to the region between divergent paused RNA polymerase using a strategy that we call dREG-HD (manuscript in preparation, preliminary version available at https://github.com/Danko-Lab/dREG.HD). Briefly, we used an epsilon-support vector regression (SVR) with a Gaussian kernel to map the distribution of PRO-seq reads to smoothed DNase-I signal intensities. Training was conducted on randomly chosen positions within dREG peaks extended by 200bp on either side. Selection of feature vectors was optimized based on Pearson correlation coefficients between the imputed and experimental DNase-I score over the validation set. PRO-seq data was normalized by sequencing depth and further scaled such that the maximum value of any prediction dataset is within 90 percentile of the training examples. We chose a step size to be 60bp and extending 30 steps on each direction. The final model was trained using matched DNase-I and PRO-seq data in K562 cells.

Next we identified peaks in the imputed DNase-I hypersensitivity profile by fitting the imputed DNase-I signal using a cubic spline and identifying local maxima. We optimized two free parameters that control the (1) smoothness of spline curve fitting, and (2) threshold on the imputed DNase-I signal intensity. Parameters were optimized to achieve an appropriate trade-off between FDR and sensitivity on the testing K562 dataset. Parameters were tuned using a grid optimization over free parameters.

DNA sequence analysis

Finding candidate transcription factor binding motifs—All motif analyses focused on 1,964 human TF binding motifs from the CisBP database[59] clustered using an affinity propagation algorithm into 567 maximally distinct DNA binding specificities (see ref[60]).

Scores, which reflect a $\log_e$-odds ratio comparing each candidate motif model to a third-order Markov background model, were calculated using the RTFBSDB package[60].

We selected two separate motif thresholds for different analyses. Scores >10 were used in analyses which mix multiple TF binding motifs, and strike a tradeoff that focuses on minimizing false positives at the expense of sensitivity. We dropped the cutoff score to motifs >8 in analyses that use individual motifs in order to increase statistical power. For each of these thresholds, we estimated the mean genome-wide positive predictive values to be 0.60 and 0.38, respectively, for motif cutoffs of 10 and 8, by comparing motifs to ChIP-seq peak calls in K562 cells. This represents a 2-fold improvement in performance over publicly available tools, even those using high-resolution DNase-I-seq data[61] (Supplementary Fig. 8). Thus, our use of high-resolution PRO-seq data and our improved computational toolkit achieves a highly respectable level of specificity on the difficult problem of TFBS prediction.

During comparative analyses we scanned each primate reference genome separately with each motif to allow the detection of a putative binding site in any of the species included in the analysis, and then moved scores to a human (hg19) reference genome using the CrossMap tool. We chose this strategy because changes in TRE activity may reflect changes in binding in any of the primate species. For example, human gains may be explained by either a new binding site for a transcriptional activator in the human genome, or a loss in binding of a transcriptional repressor that was bound in both primate species.

Motif enrichment in TREs that change during CD4+ T-cell activation—Motifs enriched in up- or down-regulated dREG-HD TREs during CD4+ T-cell activation ($p < 0.01$) were selected using Fisher's exact test with a Bonferroni correction for multiple hypothesis testing. Up- or down-regulated TREs were compared to a background set of >2,500 GC-content matched TREs that do not change transcription levels following π treatment (log-2 fold change <0.5-fold in magnitude and $p > 0.25$) using the *enrichmentTest* function in RTFBSDB[60]. To test for motif robustness, the background resampling was repeated 100 times and motifs were selected that achieve a significant result in >90%.

DNA sequence conservation analysis—For our evolutionary conservation analysis, we used phyloP scores[62] based on the 100-way genome alignments available in the UCSC Genome Browser (hg19). In all cases, bigWig files were obtained from the UCSC Genome Browser and processed using the bigWig package in R. We represented evolutionary conservation as the mean phyloP score in each identified TFBS in the indicated set of dREG-HD sites.

Enrichment of DNA sequence changes in motifs—We identified single-nucleotide DNA sequence differences at sites at which two of three primate species share one base and the third species diverges. We intersected these species-specific divergences with matches to transcription factor binding motifs found within dREG-HD sites that undergo transcriptional changes between primate species. Because many motifs in Cis-BP are similar to one another, we first partitioned the motifs using clustering (as described above), and examined enrichments at the level of these clusters. Motifs were ranked by the Fisher's exact test p-

value of the enrichment of species divergences in dREG-HD sites that change transcription status (where changes in DNA sequence and transcription occur on the same branch) to dREG-HD sites that do not change. We also compute the enrichment ratio, which we define as the number of species divergences in each TF binding motif in dREG-HD sites that change on the same branch normalized to the same statistic in sites that do not change.

INSIGHT analysis—We examined the modes by which DNA sequences evolve in human lineage-specific dREG-HD sites or DHSs using INSIGHT[46]. We passed INSIGHT either complete DHSs, dREG-HD sites, or TFBS within dREG-HD sites that undergo the changes (see *Identifying differences in TREs between species*) indicated in the comparison. Human gains and losses, for example, were comprised of 4,384 dREG-HD sites with 9,924 separate regions (median length of 16 bp) after merging overlapping TFBSs with a log-odds score greater than 10. We also analyzed 24 transcription factors each of which has more than 900 occurrences in dREG-HD sites that change on the human branch (log-odds score >8). All analyses were conducted using the INSIGHT web server (http://compgen.cshl.edu/INSIGHT/) with the default settings enabled.

bQTL analysis—Frequency shift estimates for all variants in Tehranchi et al. (2016) (ref. [63]) were provided by the authors and converted to a queryable database filtered to include only variants with coverage by 25 reads (75^{th} percentile) or more to avoid noise at low read counts. For each sequence/variant query, a set of four equivalent sequences/alternate allele pairs was constructed by swapping which allele was the reference and getting the reverse complement for both alleles. For example, given a sequence:variant:position combination of AATCGAA:C:3, the other queries produced were AACCGAA:T:3 (allele swap), TTCGATT:G:5 (reverse complement), and TTCGGTT:A:5 (reverse complement allele swap). Frequency shifts were computed by taking the post-ChIP frequency minus the pre-ChIP frequency for the human reference allele. Since k-mers longer than 7 had few hits, we allowed for wildcards (N) in longer sequences that would match any base. Wildcards were introduced into a k-mer by matching the k-mer sequence to the NF-kB motif and replacing the 3 lowest information content positions with N(s). Systematic shifts from 0 were tested using a one-tailed t-test. P-values for systematic differences at multiple sites were combined using Fisher's method

De novo discovery of transcription units

Identification of transcription units (TU) using a three-state hidden Markov model—We inferred transcription units (TU) using a three-state hidden Markov model (HMM) similar to those we have recently published[64,65]. Each TU begins at a TRE identified using dREG and continues through the entire region inferred to be transcribed, which can covers tens- to hundreds- of kilobases. Three states were used to represent background (i.e., outside of a transcription unit), the TU body, and a post-polyA decay region. The HMM transition structure is shown in Supplementary Fig. 13a. We allow skipping over the post-polyA state, as unstable transcripts do not have these two-phase profiles. We took advantage of dREG as a potential signal for transcription initiation by incorporating the dREG score (maximum value in the interval from a given positive read-count position until the next, clamped to the zero-one interval) as a transition probability

from the background to the transcription body state. PRO-seq data is generally sparse, so we applied a transformation that encoded only non-zero positions and the distance between such consecutive positions (Supplementary Fig. 13a). Our model described this transformed data using emissions distributions based on two types of variables. The first type of emission variable defines the PRO-seq read counts in non-zero positions. These counts were modeled using Poisson distributions in the background and post-polyA states, and using a Negative Binomial distribution in the transcription body state. The negative binomial distribution can be seen as a mixture of Poisson distributions with gamma-distributed rates and therefore allows for variation in TU expression levels across the genome. The second type of emission variable describes the distribution of distances in base pairs between positions having non-zero read counts. This distribution was modeled using a separate geometric distribution for each of the three states. Maximum likelihood estimates of all free parameters were obtained via Expectation Maximization, on a per-chromosome basis. TU predictions were then obtained using the Viterbi algorithm with parameters fixed at their maximum-likelihood values. Finally these predictions were mapped from the transformed coordinates back to genomic coordinates. Source code for our implementation is publicly available on GitHub: https://github.com/andrelmartins/tunits.nhp.

Inferring TU boundaries in the common great ape ancestor—We identified the most likely TU locations in the great ape ancestor by maximum parsimony. TUs were identified and compared in human reference coordinates (hg19) for all species. We used the bedops package to find the intersection between the predicted TU intervals in each pair of species (i.e., human-chimp, human-rhesus macaque, and chimp-rhesus macaque). Intersections (>= 1bp) between pairs of species were merged, resulting in a collection of TUs shared by any two pairs of species, and therefore likely to be a TU in the human-chimp ancestor. All steps were applied independently on the plus and minus strands. These steps identified 37,626 putative TUs active in CD4+ T-cells of the primate ancestor. We added 17,167 TUs that did not overlap ancestral TUs but were found in any one of the three primate species.

Transcription unit classification—TUs were classified by annotation type using a pipeline similar to ones that we have described recently[64–66]. Before classifying TUs we applied a heuristic to refine TUs on the basis of known annotations. TUs that completely overlap multiple gene annotations were broken at the transcription start site provided that a dREG site overlapped that transcription start site. Classification was completed using a set of rules to iteratively refine existing annotations, as shown in Supplementary ig. 13a. Unless otherwise stated, overlap between a TU and a transcript annotation was defined such that >50% of a TU matched a gene annotation and covers at least 50% of the same annotation. TUs overlapping GENCODE annotations (>50% overlap, defined as above) were classified using the biotype in the GENCODE database into protein coding, lincRNA (lincRNA or processed transcript), or pseudogene. The remaining transcripts were classified as annotated RNA genes using GENCODE annotations, the rnaGenes UCSC Genome Browser track (converted from hg18 to hg19 coordinates), and miRBase v20[67]. As many RNA genes are processed from much longer TUs, we required no specific degree of overlap for RNA genes. Upstream antisense (i.e., divergent) TUs were classified as those within 500bp of the

transcription start site of any GENCODE or higher level TU annotation (including lincRNAs). Antisense transcripts were defined as those with a high degree of overlap (>50%) with annotated protein coding genes in the opposite orientation. The remaining transcripts with a high degree of overlap (>50%) to annotated repeats in the repeatmasker database (rmsk) were classified as repeat transcription. Finally, any TUs still remaining were classified as unannotated, and were further divided into those which are intergenic or that partially overlapping existing annotations.

Comparing transcription between conditions and species

Comparing transcription before and after CD4+ T-cell activation—We compared π treated and untreated CD4+ T-cells within each of the primate species using gene annotations (GENCODE v19). We focused on 42,556 GENCODE-annotated transcription units (TUs) best supported by PRO-seq data for human CD4+ T-cells using tuSelector[68]. We counted reads in the interval between 500 bp downstream of the annotated transcription start site and either the end of the gene or 60,000 bp into the gene body (whichever was shorter). This window was selected to avoid (1) counting reads in the pause peak near the transcription start site, and (2) to focus on the 5′ end of the gene body affected by changes in transcription during 30 minutes of π treatment assuming a median elongation rate of 2 kb/minute[64,69]. We limited analyses to gene annotations longer than 500 bp in length. To quantify transcription at enhancers, we counted reads in the window covered by each dREG-HD site plus an additional 250 bp on each end. Differential expression analysis was conducted using DESeq2[43].

Comparing transcription between species—Read counts were compared between different species in hg19 coordinates. In all analyses, reads were transferred to the hg19 reference genome using CrossMap with rbest nets. Our analysis focused on transcription units or on the union of dREG sites across species. We focused our analysis of transcription units on the interval between 250 bp downstream of the annotated transcription start site and either the end of the gene or 60,000 bp into the gene body (whichever was shorter). We limited our analyses to TUs longer than 500 bp in length. Reads counts were obtained within each transcription unit, gene annotation, or enhancer, abbreviated here as a ‘region of interest’ (ROI), that has confident one-to-one orthology in all species examined in the analysis. This strategy of focusing on blocks of one-to-one orthology avoids errors caused by systematic differences in mappability or repeat content of species-specific genomic segments. We broke each each ROI into segments that have conserved orthology between hg19 and all species examined in the analysis, which included either a three-way (human-chimp-rhesus macaque) or five-way (human-chimp-rhesus macaque-mouse-rat) species comparison. We defined intervals of one-to-one orthology as those represented in levels 1, 3, and 5 of the reciprocal best nets (with gaps defined in levels 2, 4, and 6)[53]. Reads that map to regions that have orthology defined in all species were counted using the bigWig package in R using reads mapped to hg19 coordinates. Final counts for each ROI were defined as the sum of read counts within the regions of orthology that intersect that ROI. ROIs without confident one-to-one orthologs in all species analyzed were discarded. Our pipeline makes extensive use of the bigWig R package, Kent source tools, as well as the bedops and

bedtools software packages[50,70]. Differential expression was conducted between species using the deSeq2 package for R, as described above.

MCF-7 G11 cell culture

Analysis of PRO-seq data in MCF7 cells[71] revealed a similar pattern of transcription at the SGPP2 locus (Supplementary Fig. 10). MCF7 G11 tamoxifen resistant cells, were a gift from Dr. Joshua LaBaer. Cells were maintained in DMEM with 5% FBS, antibiotics, and 1uM tamoxifen. MCF-7 G11 dCas9-KRAB stable cell lines were made (as described below) and were maintained in DMEM with 5% FBS, antibiotics, and 1uM tamoxifen. MCF-7 G11 dCas9-KRAB sgRNA stable cell lines were maintained in DMEM with 5% FBS, antibiotics, 1uM tamoxifen, and 150ug/ul Hygromycin B.

Luciferase assays

Genomic DNA was isolated from human, chimp, and rhesus macaque PBMCs depleted for CD4+ cells using a Quick-DNA Miniprep Plus Kit (#D4068S; Zymo research) following the manufacturer's instructions. Putative enhancer regions were amplified from the genomic DNA, restriction digested with KpnI and MluI, and cloned into the pGL3-promoter vector (Promega). The same orthologous regions were amplified from all three species with identical primers where possible or species-specific primers covering orthologous DNA in diverged regions. Vectors were co-transfected with pRL-SV40 Renilla (Promega) in a 10:1 ratio (500ng pGL3 to 50ng pRL-SV40) in MCF7 G11 cells cultured in 1uM tamoxifen. Transfected cells were treated with either 25ng/ml TNFa or water 21 hours after transfection. 24 hours post-transfection, luminescence was measured in triplicate using the Dual-Luciferase® Reporter Assay System (Promega).

Silencing endogenous TREs using dCAS9-KRAB

Cloning signle-guide RNAs (sgRNAs)—We used CRISPR interference (CRISPRi) to silence enhancers near *SGPP2*[72]. Single- guide RNAs (sgRNAs) were designed using the CRISPR design tool (http://crispr.mit.edu) and sequences are shown in Supplementary Table 3. Forward and reverse sgRNAs were synthesized separately by IDT and annealed. T4 Polynucleotide Kinase (NEB) was used to phosphorylate the forward and reverse sgRNA during the annealing. 10× T4 DNA Ligase Buffer, which contains 1mM ATP, was incubated for 30 minutes at 37°C and then at 95C for 5 minutes, decreasing by 5°C every 1 minute until 25°C. Oligos were diluted 1:200 using Molecular grade water. sgRNAs were inserted into the pLenti SpBsmBI sgRNA Hygro plasmid from addgene (#62205) by following the authors protocol[73]. The plasmid was linearized using BsmBI digestion (NEB) and purified using gel extraction (QIAquick Gel Extraction Kit). The purified linear plasmid was then dephosphorylated using Alkaline Phosphatase Calf Intestinal (CIP) (NEB) to ensure the linear plasmid did not ligate with itself. A second gel extraction was used as before to purify the linearized plasmid. The purified dephosphorylated linear plasmid and phosphorylated annealed oligos were ligated together using the Quick Ligation Kit (NEB). The ligated product was transformed into One Shot Stbl3 Chemically Competent E. coli (ThermoFisher Scientific). 100ul of the transformed bacteria were plated on Ampicillin (200ug/ml) plates.

Single colonies were picked, sequenced, and the plasmid was isolated using endo free midi-preps from Omega.

Transfection of MCF-7 G11 cell lines—We used lentivirus to transfect MCF-7 cells. Lentivirus was made using lipofectamine 3000 from Invitrogen. Phoenix Hek cells (grown in DMEM with 10% FBS and antibiotics) were seeded in a 6-well plate at 400,000 cells/plate. Cells were grown until ~90% confluent. 1ug of pHAGE_EF1a_dCas9-KRAB plasmid from addgene (#50919) plasmid or the pLenti SpBsmBI sgRNA Hygro (addgene #62205) containing each sgRNA, 0.5ug of psPAX (addgene #12260), and 0.25ug pMD2.G (addgene #12259) were mixed.

MCF-7 G11 cells were plated at ~200,000 cells/well in a 6-well plate. 24 hours later 3ml/well of virus was mixed with 10ug/ml polybrene and incubated for 5 minutes at room temperature. This mix was added to the cells and centrifuged for 40 minutes at 800g at 32C. 12–24 hours later the virus was removed and fresh media was added. 24–48 hours later the cells were selected with 2ug/ml puromycin for 2 weeks. The MCF-7 G11 dCas9-KRAB stable cell lines was grown and maintained in puromycin. A second lentiviral infection was done using the stable MCF-7 G11 dCas9-KRAB cells. The same protocol was used. 24–48 hours later the cells were selected with 150ug/ml Hygromycin B. New stable cell lines were grown and maintained in hygromycin B.

TNFa treatment—Prior to TNFa treatment, cells were grown for 3 days in DMEM with 5% FBS, antibiotics, tamoxifen and hygromycin. Cells were then left untreated or treated for 40 min with 25ng/ml TNFa. RNA was extracted using TRIzol Reagent (Invitrogen). We reverse transcribed 1ug of RNA and used this as input for real-time quantitative PCR (RT-PCR) to analyze *SGPP2* expression. Primers for *SGPP2* were designed targeting a sequence in intron 1, upstream of the intronic enhancer. Raw *Cp* values were transferred to units of expression using a standard dilution curve comprised of a mixture of cDNA from each sample within the biological replicate. We included four serial dilutions, each of which covered a two-fold difference in expression. Each sample was further normalized for differences in RNA content by primers recognizing the 18S rRNA control. The ratio between normalized *SGPP2* expression in each sgRNA-transfected MCF-7 cell line and the empty vector control was log-2 transformed and tested for differences from 0 using a two-sided t-test.

Data availability

PRO-seq data was deposited into the Gene Expression Omnibus database under accession number GSE85337.

Code availability

All data analysis scripts and software are publicly available on GitHub: https://github.com/Danko-Lab/CD4-Cell-Evolution.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank M. Jin for assistance in establishing the magnetic separation of CD4+ T-cells, J. Rogers for help establishing contacts with primate centers, L. Core, H. Kwak, N. Fuda, and I. Jonkers for assistance troubleshooting the PRO-seq library prep, and A. Wetterau for preparing nuclei for mouse and rat CD4+ T-cells. Work in this publication was supported by generous seed grants from the Cornell University Center for Vertebrate Genomics (CVG), the Center for Comparative and Population Genetics (3CPG), NHGRI (National Human Genome Research Institute) grant HG009309 to CGD, NHLBI (National Heart, Lung, and Blood Institute) grant UHL129958A to CGD and JTL, NIGMS (National Institute of General Medical Sciences) grant GM102192 to AS, NHGRI (National Human Genome Research Institute) grant HG0070707 to AS and JTL, NIH/NIDDK DK058110 to WLK, and CPRIT RP160319 to WLK. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health. Finally, we would like to thank the anonymous human and non-human primate donors who gave blood in support of this study.

References

1. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961; 3:318–356. [PubMed: 13718526]
2. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969; 165:349–357. [PubMed: 5789433]
3. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188:107–116. [PubMed: 1090005]
4. Rockman MV, et al. Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol. 2005; 3:e387. [PubMed: 16274263]
5. Prabhakar S, et al. Human-specific gain of function in a developmental enhancer. Science. 2008; 321:1346–1350. [PubMed: 18772437]
6. Capra JA, Erwin GD, McKinsey G, Rubenstein JLR, Pollard KS. Many human accelerated regions are developmental enhancers. Philos Trans R Soc Lond B Biol Sci. 2013; 368:20130025. [PubMed: 24218637]
7. McLean CY, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature. 2011; 471:216–219. [PubMed: 21390129]
8. Arbiza L, et al. Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet. 2013; 45:723–729. [PubMed: 23749186]
9. Wilson MD, et al. Species-specific transcription in mice carrying human chromosome 21. Science. 2008; 322:434–438. [PubMed: 18787134]
10. Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nat Genet. 2007; 39:1140–1144. [PubMed: 17694055]
11. Torgerson DG, et al. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet. 2009; 5:e1000592. [PubMed: 19662163]
12. Cotney J, et al. The evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013; 154:185–196. [PubMed: 23827682]
13. Schmidt D, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science. 2010; 328:1036–1040. [PubMed: 20378774]
14. Ballester B, et al. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. Elife. 2014; 3:e02626. [PubMed: 25279814]
15. Vierstra J, et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014; 346:1007–1012. [PubMed: 25411453]
16. Arnold CD, et al. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat Genet. 2014; 46:685–692. [PubMed: 24908250]
17. Doniger SW, Fay JC. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 2007; 3:e99. [PubMed: 17530920]

18. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M. Genetic analysis of variation in transcription factor binding in yeast. Nature. 2010; 464:1187–1191. [PubMed: 20237471]
19. Bradley RK, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS Biol. 2010; 8:e1000343. [PubMed: 20351773]
20. Villar D, et al. Enhancer evolution across 20 mammalian species. Cell. 2015; 160:554–566. [PubMed: 25635462]
21. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016; 351:1083–1087. [PubMed: 26941318]
22. Fuda NJ, Ardehali MB, Lis JT. Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature. 2009; 461:186–192. [PubMed: 19741698]
23. Cain CE, Blekhman R, Marioni JC, Gilad Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics. 2011; 187:1225–1234. [PubMed: 21321133]
24. Xiao S, et al. Comparative epigenomic annotation of regulatory DNA. Cell. 2012; 149:1381–1392. [PubMed: 22682255]
25. Zhou X, et al. Epigenetic modifications are associated with inter-species gene expression variation in primates. Genome Biol. 2014; 15:547. [PubMed: 25468404]
26. Paris M, et al. Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. PLoS Genet. 2013; 9:e1003748. [PubMed: 24068946]
27. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014; 10:e1004226. [PubMed: 24603674]
28. Wong ES, et al. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. Genome Res. 2015; 25:167–178. [PubMed: 25394363]
29. Hah N, Murakami S, Nagari A, Danko CG, Kraus WL. Enhancer transcripts mark active estrogen receptor binding sites. Genome Res. 2013; 23:1210–1223. [PubMed: 23636943]
30. Domené S, et al. Enhancer turnover and conserved regulatory function in vertebrate evolution. Philos Trans R Soc Lond B Biol Sci. 2013; 368:20130027. [PubMed: 24218639]
31. Wunderlich Z, et al. Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. Cell Rep. 2015; 12:1740–1747. [PubMed: 26344774]
32. Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 2000; 403:564–567. [PubMed: 10676967]
33. Cannavò E, et al. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr Biol. 2016; 26:38–51. [PubMed: 26687625]
34. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012; 489:109–113. [PubMed: 22955621]
35. Vietri Rudan M, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015; 10:1297–1309. [PubMed: 25732821]
36. Khan Z, et al. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013; 342:1100–1104. [PubMed: 24136357]
37. Battle A, et al. Genomic variation. Impact of regulatory variation from RNA to protein. Science. 2015; 347:664–667. [PubMed: 25657249]
38. Bauernfeind AL, et al. Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. Genome Biol Evol. 2015; 7:2276–2288. [PubMed: 26163674]
39. Pai AA, et al. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. 2012; 8:e1003000. [PubMed: 23071454]
40. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science. 2013; 339:950–953. [PubMed: 23430654]
41. Danko CG, et al. Identification of active transcriptional regulatory elements from GRO-seq data. Nat Methods. 2015; 12:433–438. [PubMed: 25799441]
42. Mahat DB, et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016; 11:1455–1476. [PubMed: 27442863]

43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15:550. [PubMed: 25516281]
44. Prescott SL, et al. Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell. 2015; 163:68–83. [PubMed: 26365491]
45. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. 2011; 21:1273–1283. [PubMed: 21632746]
46. Gronau I, Arbiza L, Mohammed J, Siepel A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol Biol Evol. 2013; 30:1159–1171. [PubMed: 23386628]
47. Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. Nature Ecology & Evolution. 2017; doi: 10.1038/s41559-017-0377-2
48. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011; 17:10–12.
49. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
50. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
51. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013; 14:144–161. [PubMed: 22908213]
52. Zhao H, et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014; 30:1006–1007. [PubMed: 24351709]
53. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003; 100:11484–11489. [PubMed: 14500911]
54. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
55. Chepelev I, Wei G, Wangsa D, Tang Q, Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. 2012; 22:490–503. [PubMed: 22270183]
56. Javierre BM, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016; 167:1369–1384.e19. [PubMed: 27863249]
57. Rao SSP, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159:1665–1680. [PubMed: 25497547]
58. Hnisz D, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–947. [PubMed: 24119843]
59. Weirauch MT, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158:1431–1443. [PubMed: 25215497]
60. Wang Z, Martins AL, Danko CG. RTFBSDB: an integrated framework for transcription factor binding site analysis. Bioinformatics. 2016; doi: 10.1093/bioinformatics/btw338
61. Sherwood RI, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol. 2014; 32:171–178. [PubMed: 24441470]
62. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20:110–121. [PubMed: 19858363]
63. Tehranchi AK, et al. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell. 2016; 165:730–741. [PubMed: 27087447]
64. Hah N, et al. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. Cell. 2011; 145:622–634. [PubMed: 21549415]
65. Chae M, Danko CG, Kraus WL. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. BMC Bioinformatics. 2015; 16:222. [PubMed: 26173492]

66. Luo X, Chae M, Krishnakumar R, Danko CG, Kraus WL. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. BMC Genomics. 2014; 15:155. [PubMed: 24564208]

67. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014; 42:D68–73. [PubMed: 24275495]

68. Dukler N, et al. Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. Genome Res. 2017; doi: 10.1101/gr.222935.117

69. Danko CG, et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol Cell. 2013; 50:212–222. [PubMed: 23523369]

70. Neph S, et al. BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012; 28:1919–1920. [PubMed: 22576172]

71. Franco HL, Nagari A, Kraus WL. TNFα signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. Mol Cell. 2015; 58:21–34. [PubMed: 25752574]

72. Thakore PI, et al. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nat Methods. 2015; 12:1143–1149. [PubMed: 26501517]

73. Pham H, Kearns NA, Maehr R. Transcriptional Regulation with CRISPR/Cas9 Effectors in Mammalian Cells. Methods Mol Biol. 2016; 1358:43–57. [PubMed: 26463376]

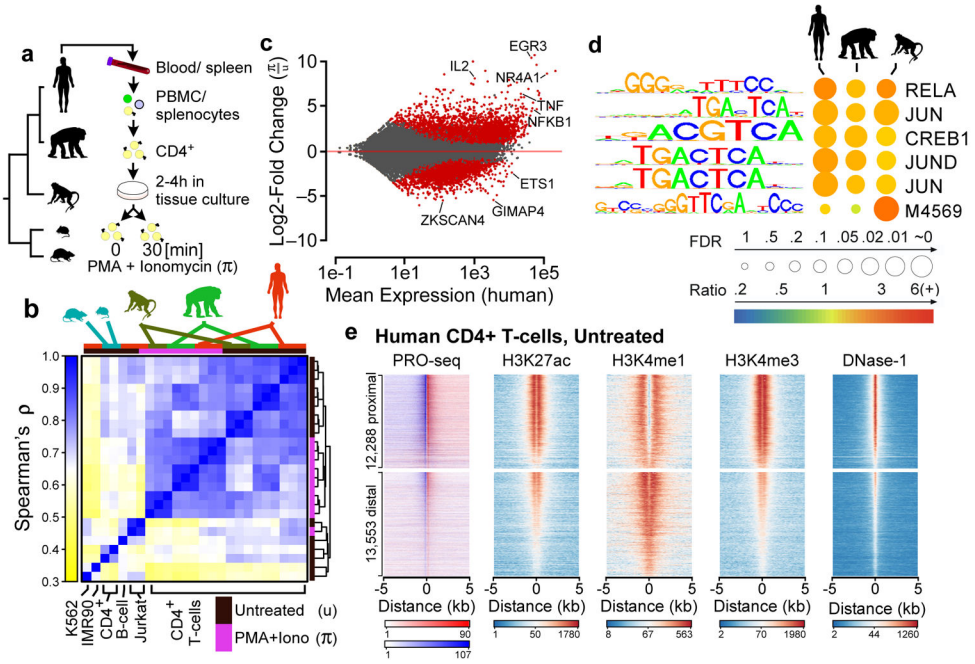

Fig. 1. Maps of primary transcription in CD4+ T-cells

(a) CD4+ T-cells were isolated from the blood or spleen of individuals from five vertebrate species, including human, chimpanzee, rhesus macaque, mouse, and rat. **(b)** Hierarchical clustering of PRO-seq signal intensities in gene bodies groups CD4+ T-cell samples first by treatment condition and second by species. The color scale represents Spearman's rank correlation between normalized transcription levels in active gene bodies. Colored boxes (top) represents the species and treatment condition of each sample. **(c)** MA plot shows the $\log_2$ fold-change following **π** treatment in human CD4+ T-cells (y-axis) as a function of the mean transcription level in GENCODE annotated genes (x-axis). 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*, *TNFα*, and *EGR3*) are marked. **(d)** Enrichment of TF binding motifs in TREs that increase transcription levels following **π** treatment in the indicated species (n = 8,030 [human], 7,258 [chimpanzee], 7,967 [rhesus macaque]) compared to TREs whose transcription abundance does not change. Table shows the Bonferroni corrected p-value based on a Fisher's exact test (circle size), and the fold-enrichment over a group of unchanged background sequences (color scale). Motif p-values were calculated based on 100 distinct samples of the background distribution each with >2,500 sites after correcting for differences in GC content (see **Online Methods**). Motif logos and the candidate transcription factor or Cis-BP motif ID are shown. **(e)** Heatmaps show the distribution of PRO-seq (red and blue indicate transcription on the plus and minus strand, respectively), ChIP-seq for H3K27ac, H3K4me1, and H3K4me3, and DNase-I-seq signal intensity. Plots are centered on transcriptional regulatory elements (TREs) predicted in untreated human CD4+ T-cells using dREG-HD (see Online Methods). All plots are ordered based on the maximum dREG score in the window.

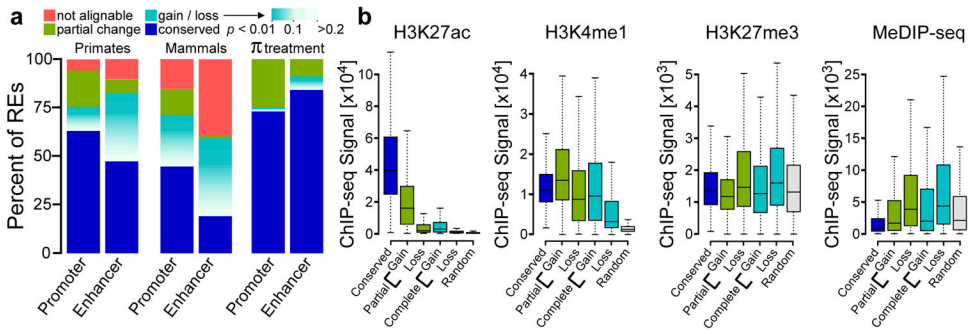

Fig. 2. Frequency of changes in TRE transcription

(a) The fractions of TREs active in untreated CD4+ T-cells that are present in the human reference genome and are conserved across all species (blue), are not detectable and are therefore inferred as gains or losses (teal-white) or undergo significant changes (green) in at least one species, or fall in regions for which no ortholog occurs in at least one of the indicated genomes (pink). Inferred gains or losses are colored according to the FDR corrected p-value associated with changes in RNA polymerase abundance (DESeq2). Plots labeled “Primate” illustrate frequency of changes in a three-way comparison of human, chimpanzee, and rhesus macaque focusing on the untreated condition, whereas those labeled “Mammal” summarize a five-way comparison also including rat and mouse. **π** treatment denotes a comparison between human untreated and PMA+Ionomycin treated CD4+ T-cell samples. **(b)** Boxplots show the ChIP-seq signal near dREG sites classified as conserved (n = 2,887), gain (n = 1,002), complete gain (n = 1,938), loss (n = 854), or complete losses (n = 1,430) for the indicated chromatin or DNA modification in units of reads per kilobase. The box represents the 25th and 75th percentile. Whiskers represent 1.5 times the interquartile range, and points outside of this range are not shown.

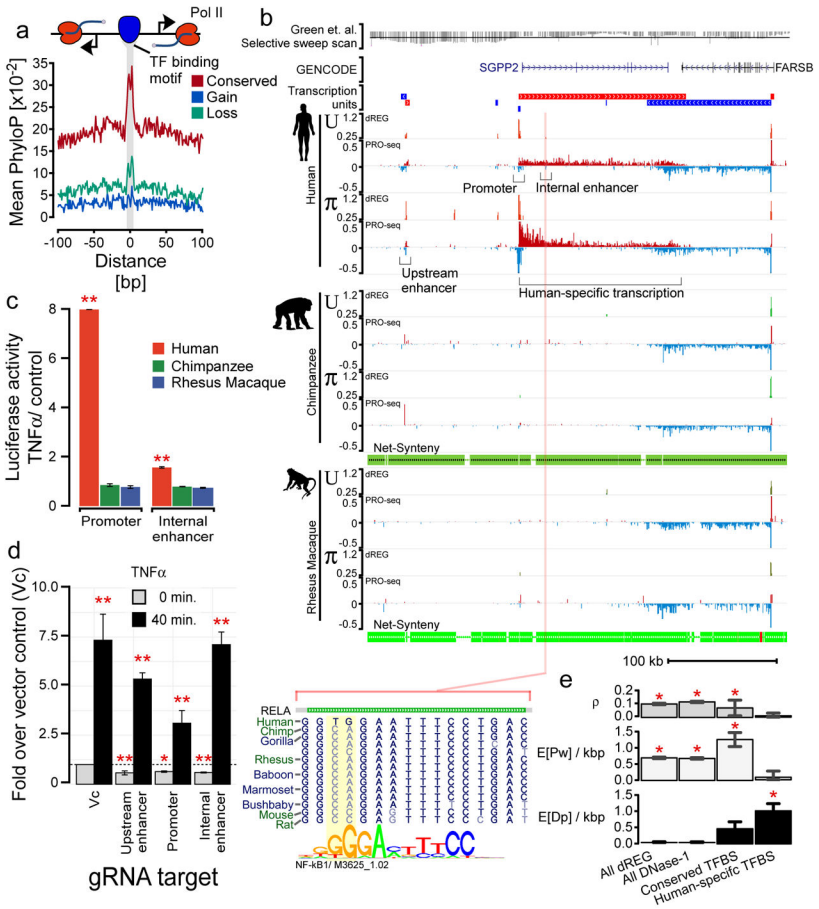

Fig. 3. Evolutionary changes in TRE transcription correlate with DNA sequence conservation

(a) Mean phyloP scores near TFBSs that are conserved (red, $n = 8,271$), gained (blue, $n = 9,642$), or lost (cyan, $n = 11,632$) on the human branch. Motifs (score > 10) are at least 100 bp from the nearest annotated exon. **(b)** UCSC Genome Browser track shows transcription near *SGPP2* and *FARSB* in untreated (U) and PMA+ionomycin (π) treated CD4+ T-cells isolated from the indicated primate species. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands in units of reads per kilobase per million mapped (RPKM). Transcription units inferred from the PRO-seq data are shown above the plot. The Green et. al. selective sweep scan track (top) represents the enrichment of derived alleles in modern human where Neanderthal has the ancestral allele. Points below the line represent a statistically significant number of derived alleles in modern human (line indicates a Z-score of −2). Net synteny tracks show the position of regions that have one-to-one orthologs in the chimpanzee and rhesus macaque genomes. **(c)** Luciferase signal driven by the *SGPP2* promoter or the internal enhancer in MCF-7 cells using DNA from each primate species. Bars show the mean fold-induction following 3 hours of stimulation with TNFα ($n = 3$). Error bars represent the standard error of the mean. Red ** denotes $p < 1e\text{-}3$ by a two-tailed t-test. **(d)** Transcription of *SGPP2* using primers targeting intron 1 following 0 or 40 min. of TNFα treatment after silencing the indicated TRE using dCAS9-KRAB. Bars represent the median of three independent biological replicates of two gRNAs targeting the promoter, three targeting the internal enhancer, and four targeting the upstream enhancer. Error bars

represent the standard error. Red * denotes $p < 5e\text{-}2$ and ** $p < 5e\text{-}3$ by a two-tailed t-test. **(e)** INSIGHT estimates of the fraction of nucleotides under selection (ρ), segregating polymorphisms under weak negative selection (E[Pw]/kbp), or human nucleotide substitutions driven by positive selection (E[Dp]/kbp) in human populations in the indicated class of sites. * denotes significant enrichment over background ($p < 0.01$; two-tailed X^2-test).

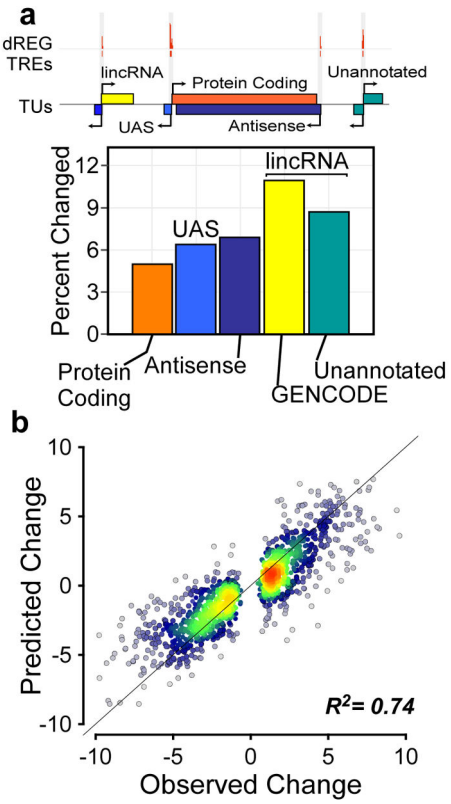

Fig. 4. Changes in non-coding RNA transcription predict changes in gene transcription

(a) The fraction of each indicated class of RNAs that undergo changes in transcription in human CD4+ T-cells (see Online Methods). The relationships among the indicated classes of transcription units are depicted at top. **(b)** Scatterplot shows the magnitude of changes in transcription predicted for protein-coding genes using changes in the transcription of nearby non-coding RNAs (y-axis) as a function of changes observed (x-axis). The line has a slope of 1 and an intercept of 0.

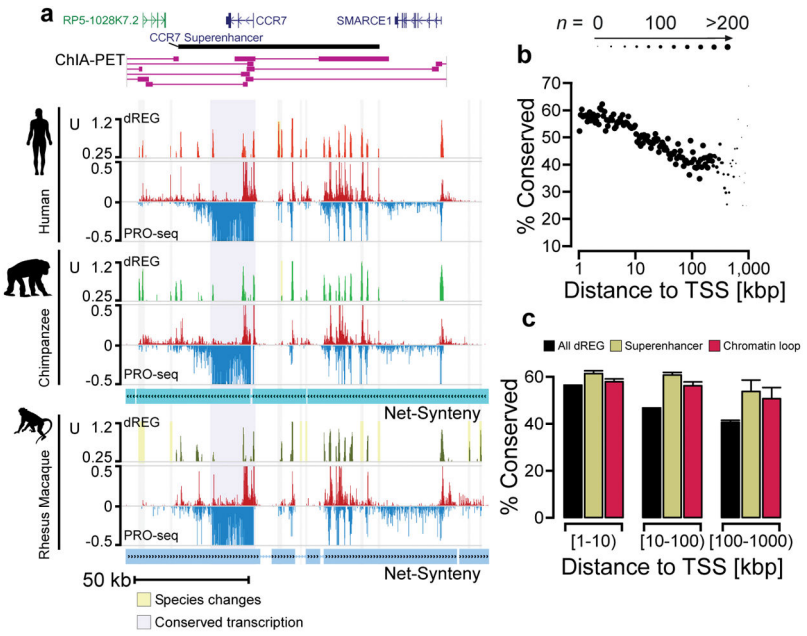

Fig. 5. TRE conservation correlates with loop interactions and distance to gene promoters

(a) UCSC Genome Browser tracks show transcription, dREG signal, and ChIA-PET loop interactions near the *CCR7* superenhancer in the human genome. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands in units of RPKM. Net synteny tracks show regions of one-to-one orthology with the chimpanzee and rhesus macaque genomes. **(b)** Scatterplot shows the percentage of TREs conserved among all three primate species (y-axis) as a function of distance, either upstream or downstream, from the nearest annotated protein-coding transcription start site (x-axis). The size of each point represents the amount of data in the corresponding distance bin. **(c)** The percentage of all dREG sites that are conserved in each indicated class of TRE. TREs are separated into three bins based on the distance relative to the nearest transcription start site. Error bars reflect a 1,000-sample bootstrap.

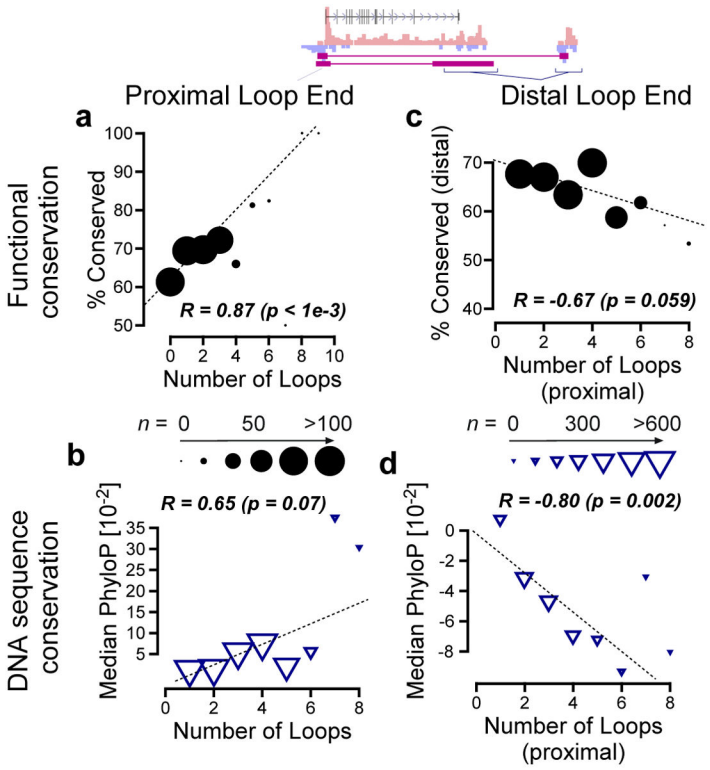

Fig. 6. Stabilizing selection on protein coding gene transcription

(a–b) Scatterplot shows promoter conservation **(a)** or DNA sequence conservation **(b)** as a function of the number of loop interactions made by that site to distal sites across the genome (x-axis). **(c–d)** TRE conservation **(c)** or DNA sequence conservation **(d)** as a function of the number of loop interactions made by the sequence at the distal end of the loop interaction (x-axis). In all panels the size of each point is proportional to the number of examples in the corresponding bin, following the scale shown in the center.