Natural Selection has Shaped Coding and Non-coding Transcription in Primate CD4+ T-cells

Charles G. Danko[1,2,*], Zhong Wang[1], Edward J. Rice[1], Tinyi Chu[1,3], Andre L. Martins[1], Elia Tait Wojno[1,4], John T. Lis[5], W. Lee Kraus[6,7], & Adam Siepel[8,*]

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

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

[6] 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.

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

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

[*] **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
E-mail: 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
E-mail: asiepel@cshl.edu

Abstract:

Transcriptional regulatory changes have been shown to contribute to phenotypic differences between species, but many questions remain about how gene expression evolves. Here we report the first comparative study of nascent transcription in primates. We used PRO-seq to map actively transcribing RNA polymerases in resting and activated CD4+ T-cells in multiple human, chimpanzee, and rhesus macaque individuals, with rodents as outgroups. This approach allowed us to directly measure active transcription separately from post-transcriptional processes. We observed general conservation in coding and non-coding transcription, punctuated by numerous differences between species, particularly at distal enhancers and non-coding RNAs. Transcription factor binding sites are a primary determinant of transcriptional differences between species. We found evidence for stabilizing selection on gene expression levels and adaptive substitutions associated with lineage-specific transcription. Finally, rates of evolutionary change are strongly correlated with long-range chromatin interactions. These observations clarify the role of primary transcription in regulatory evolution.

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[4–12]. Changes in gene expression are primarily driven by mutations to non-coding DNA sequences, particularly those that bind sequence-specific transcription factors[13]. Accordingly, adaptive nucleotide substitutions at transcription factor binding sites (TFBSs)[9,10,14–16] and gains and losses of TFBSs[17–25] both appear to make major contributions to the evolution of gene expression. These events are believed to modify a variety of rate-limiting steps early in transcriptional activation[26]. Transcriptional activity is generally correlated with epigenomic and structural features such as 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[8,24,27–29].

Nevertheless, many open questions remain about the relative roles of TFBSs, chromatin organization, and posttranscriptional regulation in the evolution of gene expression. There is a surprisingly limited correlation between differences in binding events and differences in mRNA expression levels[30–32]. Possible reasons for this discordance include non-functional TF binding[30,31,33], compensatory gains and losses of TFBSs[20,34–37], difficulties associating distal enhancers with target genes[38], and a dependency of TF function on chromatin or chromosomal organization[39]. In addition, it remains unclear to what degree epigenomic differences between species are causes and to what degree they are effects of differences in gene expression. Finally, some changes in mRNA expression appear to be "buffered" at the post-transcriptional level[40–42].

One reason why it has been difficult to disentangle these contributions is that gene expression is typically measured in terms of the abundance of mRNA, which is subject to post-transcriptional processing[43] and therefore is an indirect measure of the transcription of genes by RNA polymerase II. An alternative and complementary approach is to measure the production of nascent RNAs using Precision Run-On and sequencing (PRO-seq) and related technologies[44–48]. Nascent RNA sequencing methods directly measure active transcription and are highly sensitive to immediate and transient transcriptional responses to stimuli[49]. They can detect active regulatory elements as well as target genes, because these elements themselves display distinctive patterns of transcription, which are obscured in RNA-seq data owing to rapid degradation[33,50,51]. Indeed, the latest nascent RNA sequencing methods, such as PRO-seq[45], in combination with new computational tools for regulatory element prediction[52], serve as powerful single-assay systems for both identifying regulatory elements and measuring transcription levels.

With these advantages in mind, we undertook a genome-wide comparative analysis of transcription in primates using PRO-seq. Our comparison of PRO-seq data across species revealed overall conservation in the transcription of both coding and non-coding elements, but also uncovered numerous differences between species. Together, our observations provide new insights into the evolution of transcription in primates.

Patterns of transcription in resting and activated CD4+ T-cells

We developed nucleotide-resolution maps of RNA polymerase for 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**). Resting samples were also collected 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). PRO-seq[45,48] libraries were sequenced to a combined depth of 873 million uniquely mapped reads (~78-274 million per species) (**Supplementary Table 1**). Flow cytometry was used to validate the purity of isolated CD4+ cells (**Supplementary Fig. 1**), and measurements of transcriptional activity of T-cell subset markers for T-helper type 1 (Th1), Th2, Th17, T-regulatory, and T-follicular helper cells were used to demonstrate that the population of CD4+ T-cell subsets within the total CD4+ population is largely similar among these mammalian species (**Supplementary Fig. 2**). Principal component analysis ranked the rodent vs. primate species, variation within the primate species, and the treatment condition as the first, second, and third sources of variation, respectively, in the complete dataset (**Supplementary Fig. 3**). Similarly, hierarchical clustering of these data together with data from other cell types grouped primate samples first by cell type or treatment condition and subsequently by species (**Fig. 1b**). These results demonstrate that genome-wide patterns of transcription remain generally concordant in CD4+ T-cells in the species we examined, especially within the primates.

Before comparing species, we evaluated differences in transcription between resting and activated conditions. Human CD4+ T-cells activated using PMA and ionomycin (π) underwent significant changes in transcription levels at 5,945 genes ($p < 0.01$, deSeq2[53]), including classical early response genes such as *IL2*, *IL2RA*, and *EGR3* (**Fig. 1c**). Parallel analyses in chimpanzee and rhesus macaque revealed comparable transcription programs both before and following π treatment (**Supplementary Fig. 4a-b**). We identified a core set of 2,953 genes that undergo evolutionarily conserved transcriptional changes following 30 min. of π-treatment in all three species, including many of the classical response genes (e.g., IFNG, TNFα, IL2, and IL2RA), as well as numerous novel genes and lincRNAs (**Supplementary Fig. 4a-c**). Fold-changes of these core genes following π treatment are highly correlated and similar in magnitude among all three primate species ($R > 0.92$; **Supplementary Fig. 4d-f**), suggesting that species-specific differences in T-cell activation magnitude[54] may occur primarily during the initial signaling events and are largely bypassed by π-treatment.

To shed light on the mechanisms underlying transcriptional changes immediately following π treatment, we used dREG[52] to identify 25,513 active transcriptional regulatory elements (TREs), including promoters and enhancers, in human CD4+ T-cells, based on patterns of enhancer-templated RNA (eRNA) transcription evident from PRO-seq data (**Fig. 1d**). These predicted TREs displayed many of the expected marks of regulatory function in human CD4+ T-cells, including acetylation of histone 3 lysine 27 (H3K27ac), and mono- and trimethylation of histone 3 lysine 4 (H3K4me1 and H3K4me3)[55], consistent with previous observations[33,50,52]. We identified 7,340 TREs with π-dependent changes in RNA polymerase abundance ($p < 0.01$, deSeq2). TREs with transcriptional increases following π-treatment were enriched for DNA sequence motifs recognized by NF-kB, NFAT, and the AP-1 heterodimers FOS and JUN (**Fig. 1e**), all transcription factors activated by canonical T-cell receptor signaling[56]. Motifs for other calcium-responsive factors such as MEF2A and CREB were also enriched in up-regulated TREs, consistent with the activation of calcium signaling by Ionomycin, a calcium ionophore. Downregulated TREs were enriched for

different motifs, including those for FOXO1, ELK/ELF/ETS, and SMARCC2 (**Supplementary Fig. 5**). The enrichment of the forkhead box (FOXO1) binding motif is concordant with findings that FOXO1 is required for naive T-cell homeostasis[57], and suggests that this TF is down-regulated as naive T-cells differentiate. Motif analysis in the chimpanzee and rhesus macaque genomes revealed a similar composition of motifs putatively involved in mediating transcriptional changes following π treatment (**Supplementary Fig. 6**). 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 focused an initial comparative analysis on evolutionary changes in active enhancers. We used dREG to identify 53,476 TREs that are active in either treated or untreated CD4+ T-cells of at least one primate species (ranging between 25,269-35,343 TREs in each species). TREs were classified as promoters (n = 7,114) or enhancers (n = 21,121) based on their proximity to gene annotations and the stability of the associated transcription units predicted using the order and orientation of U1 and polyadenylation site motifs[50] (**Supplementary Fig. 7**, see Methods). We found that 62% of enhancers changed Pol II transcriptional activity in at least one of the three primate species and nearly 80% at the longer evolutionary distance between primates and rodents (**Fig. 2a**). Conservation was highest between human and chimpanzee (10-20% of TREs change), roughly consistent with recent estimates based on the distribution of H3K27ac[10]. Enhancers are completely gained or lost at nearly eight times the rate of promoters (38% of enhancers; 7% of promoters; $p < 2.2e\text{-}16$ by Fisher's exact test), consistent with recent observations based on H3K27ac and H3K4me3[24]. By contrast, π treatment is rarely associated with complete differences in the location of transcribed TREs and the rate of differences is similar at promoters and enhancers (**Fig. 2a**).

Next we tested whether evolutionary changes in transcriptional activity correlate with the enrichment of other active enhancer marks. We used the known phylogeny for these species to identify likely lineage-specific gains and losses of enhancer activity in human CD4+ T-cells, and examined independent sources of genomic data[55] at the corresponding genomic locations. Overall, predicted lineage-specific human enhancers were enriched for both active and repressive enhancer marks (**Fig. 2b; Supplementary Fig. 8**). Whereas human gains were enriched for high levels of H3K27ac, sites with reduced transcriptional activity in humans showed much lower enrichments of this mark of active enhancers. Furthermore, locations of complete human lineage-specific loss of dREG signal displayed levels of H3K27ac approaching the background, consistent with a complete loss of enhancer activity (**Fig. 2b**). Intriguingly, H3K4me1, which marks active and some inactive enhancers[58,59], was enriched at predicted human lineage-specific losses in enhancer activity (**Fig. 2b**), indicating that, at least in some cases, an active ancestral primate enhancer retains a 'poised' chromatin state in human, despite losing transcriptional activity and H3K27ac. This result suggests a possible model of TRE assembly and activation in which changes in poised and active marks are evolutionarily distinct events.

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[60] at conserved TREs with scores at

171 TREs that display evolutionary changes in transcription. Because signatures of sequence
172 conservation in TREs are likely to be most pronounced in transcription factors binding sites (TFBS),
173 we restricted our sequence conservation analyses to strong matches to 1,964 human TF binding
174 motifs. We grouped these binding motifs by their DNA sequence similarity into 567 clusters and we
175 excluded motifs for TFs that are transcriptionally inactive in CD4+ T-cells[61,62].
176 TFBSs found in transcriptionally conserved dREG sites showed a marked enrichment for
177 higher phyloP scores relative to surrounding regions, indicating that these TFBSs are enriched for
178 evolutionary conservation in their DNA sequences (**Fig. 3a**). By contrast, TFBSs in lineage-specific
179 dREG sites had lower enrichments in phyloP scores (**Fig. 3a, red/blue**). Notably, TFBSs in dREG
180 sites lost on the human lineage show enhanced conservation compared with those in human-
181 specific gains, consistent with losses evolving under conservation in other mammalian species
182 (which contribute to the phyloP scores) and gains emerging relatively recently. Restricting these
183 analyses to binding sites for specific TFs revealed patterns of conservation that correlate with the
184 information content of the DNA sequence motif (**Fig 3b; Supplementary Fig. 9**), further supporting
185 TF binding as the functional property driving evolutionary conservation at these sites. This
186 analysis suggests that the sequences in TFBSs are a primary driver of transcriptional differences
187 between species.
188 To estimate the specific contribution of each TF to transcriptional changes, we compared
189 lineage-specific single-nucleotide substitutions within TFBSs in lineage-specific TREs to the same
190 statistic for TFBSs in conserved TREs. This analysis identified 363 TFs with statistically significant
191 associations between lineage-specific transcription and nucleotide substitutions in their TFBSs in
192 resting CD4+ T-cells (**Fig. 3c**). The most strongly enriched of these TFs include several with known
193 or predicted roles in CD4+ T-cell biology, such as ELF1, YY1, and CREB1. ELF1 bindings sites, for
194 example, are enriched in non-coding GWAS SNPs putatively affecting human autoimmune
195 phenotypes in CD4+ T-cell TREs[63]. We applied the same analysis to TREs displaying lineage-
196 specific changes in π-dependent activation, and identified enrichments for the motifs binding NFAT,
197 AP-1 heterodimers (FOS and JUN), and NF-kB (p = 0.03), consistent with our motif discovery results
198 within each individual primate species. In one example at the *SGPP2* locus (**Fig. 3d**), two NF-kB
199 binding motifs found in the proximal promoter and an internal enhancer of *SGPP2* (**Fig. 3e;**
200 **Supplementary Fig. 10**) are candidates for changes that cause humans to activate *SGPP2*
201 transcription following π treatment.
202 In some cases, we observed numerous nucleotide substitutions within short TFBSs, either
203 individually or in clusters of nearby TFBSs, which is unlikely to occur by chance and suggests a
204 possible role for positive selection in the evolution of these binding sites. For example, upstream of
205 *SGPP2* we noted an excess of derived alleles in modern humans in cases where the sequenced
206 Neanderthal has the ancestral allele (**Fig. 3d**)[64], potentially consistent with positive selection
207 driving evolutionary changes in *SGPP2* transcription since the divergence of modern humans and
208 Neanderthals. To more directly gauge the impact of positive selection, we used INSIGHT[65] to
209 compare patterns of within-species polymorphism and between-species sequence divergence in
210 TREs that had undergone human lineage-specific transcriptional changes. As has been reported
211 previously for regulatory sequences[9], dREG sites appear to be most strongly influenced by weak
212 negative selection, which is reflected in an excess of low frequency derived alleles in human
213 populations (**Fig. 3f**). Nevertheless, TREs with lineage-specific transcriptional changes in human

CD4+ T-cells showed reduced weak negative selection and were strikingly enriched for adaptive nucleotide substitutions ($p < 0.01$ INSIGHT likelihood ratio test; **Fig. 3f**), consistent with positive selection at these sites. We estimate that at least 160 adaptive substitutions have occurred since the human/chimpanzee divergence within TFBSs that undergo transcriptional changes in human CD4+ T-cells. Although we are underpowered to detect the specific contribution of many TFs in this analysis, we did note statistically significant excesses of predicted adaptive substitutions in the bindings sites of several TFs, including FOXO1, GATA3, IRF4, RARG, and ZEB1 (**Supplementary Fig. 11**; $p < 0.01$, INSIGHT likelihood ratio test). These estimates highlight the substantial contribution of adaptive evolutionary changes in TFBSs that influence the transcriptional activity of TREs.

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. *SGPP2*, for instance, undergoes a ~60-fold increase in basal transcription in humans, which broadly correlates with changes in nearby non-coding TUs (**Fig. 3d; Supplementary Fig. 10**). To examine this pattern more generally, we designed a hidden Markov model (HMM) to estimate the boundaries of TUs genome-wide, based on patterns of aligned PRO-seq reads. Using this method, we annotated 32,602 TUs active in CD4+ T-cells of at least two of the three primates, likely indicating a TU in the human-chimpanzee ancestor, as well as an additional 13,085 TUs that are specific to one of the primate species (**Supplementary Fig. 12a**). Approximately half of the TUs identified using this approach overlap either annotated protein-coding genes or their associated upstream antisense RNAs, a small fraction overlap annotated lincRNAs, and approximately half are completely un-annotated (**Supplementary Fig. 12b**).

A cross-species comparison of the transcription levels for various TU classes (**Fig. 4a**) revealed that enhancer RNAs evolve in expression most rapidly and protein-coding genes evolve most slowly. Upstream antisense RNAs are relatively conserved compared to more distal TUs, perhaps because they share regulatory sequences with protein-coding genes. Intriguingly, lincRNAs—including both those annotated in GENCODE and novel lincRNAs predicted by our HMM—undergo evolutionary changes in expression about as frequently as enhancer RNAs. Since patterns of nascent transcription at active lincRNA promoters are highly similar to those at active enhancers[50,66], this finding suggests that, on average, lincRNA transcription is no more conserved than can be explained by the activity of the lincRNA promoter.

We measured the extent to which non-coding and protein-coding transcriptional activities are correlated through evolutionary time. Evolutionary changes in protein-coding gene expression were highly correlated with those at upstream (Pearson's $R = 0.81$, $p < 2.2e\text{-}16$) and internal ($R = 0.66$, $p < 2.2e\text{-}16$) antisense transcripts of the same genes. Moreover, changes in the transcriptional activity of gene promoters correlate with the activities of distal enhancers to which they loop according to cell-type matched ChIA-PET data ($R = 0.40\text{-}0.69$, $p < 2.4e\text{-}6$; depending on analysis assumptions)[67], which are located nearby ($R = 0.58$, $p < 2.2e\text{-}16$), or which share the same topological associated domain ($R = 0.53$, $p < 2.2e\text{-}16$)[68]. Using a generalized linear model to integrate expression changes in multiple types of TUs, we can explain 65% of the variance in gene transcription levels ($R^2 = 0.65$ in a heldout set of sites, $p < 2.2e\text{-}16$; **Fig. 4b**) based on the activities of looped TREs, nearby TREs, internal antisense TUs, and the upstream antisense TU. Thus

evolutionary changes that result in differences in Pol II recruitment to protein-coding genes are remarkably correlated across all interacting TREs, indicating a shared evolutionary pressure at proximal and distal TREs.

Rates of Enhancer Evolution Vary with Evidence for Gene Interactions

Transcription at enhancers evolves rapidly and is frequently unaccompanied by changes at nearby protein-coding genes. For example, *CCR7* transcription is highly conserved among both primate and rodent species (**Fig. 5a; Supplementary Fig. 2**) in spite of several apparent changes in enhancer activity within the same locus (gray vertical bars). These findings are consistent with recent observations that changes in enhancers within densely populated loci often do not have appreciable effects on the transcription of genes within the locus[35,36].

To explain this effect, we searched for genomic features correlated with conservation of transcription at enhancers. We found that one of the most strongly correlated features is the distance between enhancers and the nearest annotated RefSeq transcription start site (**Fig. 5b**). In particular, about 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. DNA sequences of proximal sites were also more highly conserved (**Supplementary Fig. 13**), correlating with the higher transcriptional conservation, though the effect was limited to the area highly proximal to the transcription start site.

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[67]. Strikingly, we found that, overall, 51% of enhancers that participate in these loops were conserved between primate species compared to 37% of non-looped enhancers (**Fig. 5c**;, p = 2.1e-13, Fisher's exact test). Moreover, this high degree of transcriptional conservation at looped enhancers does not depend on the distance to the transcription start site. We observed similar levels of conservation at recently defined superenhancers[69], which may simply reflect an enrichment for loop interactions (48% of TREs in superenhancers 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 (**Fig. 5d**; phyloP > 0.75; p < 2.2e-16, Wilcoxon Rank Sum Test). That the subset of enhancers which form loop interactions with distal sites is more highly conserved at both the transcription and DNA sequence levels suggests this subset has a disproportionately large effect on fitness, perhaps because it is more directly involved in transcriptional regulation.

Enhancer-Promoter Interactions Contribute to Constraint on Gene Transcription Levels

Distal loop interactions do not fully account for the disparity in evolutionary rates between enhancer and promoter transcription. Looped enhancers still evolve significantly faster than promoters (**Supplementary Fig. 14**; p = 3e-5, Fisher's exact test). We hypothesized that a higher redundancy in cis-regulatory signals makes protein-coding genes controlled by multiple TREs, such as *CCR7* (**Fig. 5a**), more robust to enhancer turnover. Indeed, we found that evolutionary

conservation of promoter TRE transcription is remarkably strongly correlated with the number of loop interactions with distal sites (**Fig. 5e**, weighted Pearson's correlation = 0.91; p < 1e-3 by a bootstrap test). Promoters involved in one or more chromatin loops change expression 15% more slowly than non-looped promoters (p = 4.7e-4; Fisher's exact test), and the probably of transcription conservation increases by ~4% with each additional loop interaction. We observed a similar correlation between the number of loop interactions made by a target promoter and DNA sequence conservation in transcription factor binding motifs at the promoter (**Fig. 5e; Supplementary Fig. 15a**).

We hypothesized that compensation among TREs may permit changes to distal TREs that loop to promoters which themselves have many other loop interactions. We therefore examined the conservation of looped TREs as a function of the number of loops in which their gene-proximal partners participate. We found that DNA sequence conservation in putative TFBSs is significantly reduced at the distal end of loop interactions and negatively correlates with the number of loops at the proximal end (**Fig. 5f; Supplementary Fig. 15b**). This result suggests that each associated distal TFBS is individually less essential at genes having multiple loop interactions with distal sites, and it is therefore consistent with a model in which such TFBSs are more freely gained and lost during evolution. This result may also explain why superenhancers have lower DNA sequence conservation than typical looped enhancers, in spite of their strong enrichment for loop interactions (**Fig. 5d**). 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:

We describe the first comparative analysis of primary transcription in any phylogenetic group, focusing on CD4+ T-cells in primates. Using PRO-seq and several new bioinformatic tools we estimated the locations and abundance of transcription units with high resolution and accuracy. In comparison to previous studies in primates[28,32,70–72], this approach separated primary transcription from post-transcriptional processing, allowing us to study eRNAs, lincRNAs, and other rapidly degraded non-coding RNAs, as well as protein-coding genes. We found clear relationships between the DNA sequences of TFBSs and differential transcription across species and treatment conditions. We also found evidence that some transcriptional changes in humans were driven by adaptive evolution in nearby binding sites. Overall, our study provides new insights into the mode and tempo of recent evolutionary changes in transcription in primates.

Perhaps our most striking observation is that many non-coding transcription units, particularly eRNAs and lincRNAs, have undergone rapid evolutionary changes in comparison to protein-coding genes. Similar observations have been reported previously for lincRNAs[73], but, to our knowledge, the observation for eRNAs is new, and it raises a number of questions. First, why are some enhancers more conserved than others? We find that enhancers proximal to, or that loop to, annotated promoters tend to be constrained (**Fig 5b-c**). These enhancers may simply be most crucial for activating their target genes, but other factors may also contribute to their constraint. For example, perhaps these enhancers are enriched for tissue-specific functions, and are less constrained due to reduced pleiotropy[74]. Or perhaps many of them are simply not functional at all, and are transcribed as a by-product of other processes.

Second, how do protein-coding genes maintain stable transcription levels across species despite the rapid turnover of associated enhancers? One possibility is that many rapidly evolving enhancers are either not functional or act on targets other than the ones we have identified. However, several of our findings argue against this possibility; for example, we find that even looped enhancers evolve significantly faster than promoters (**Supplementary Fig. 14**), that eRNA conservation is strongly correlated with the number of loop interactions at associated promoters (**Fig. 5e**), and that sequence conservation at distal enhancers is negatively correlated with the number of loop interactions at associated promoters (**Fig. 5f**). An alternative explanation, which appears more plausible to us, is that stabilizing selection on transcription levels drives enhancers to compensate for one another as they undergo evolutionary flux. This observation would be compatible with reports from model systems[35,37]. The possibility of pervasive stabilizing selection on transcription levels in primates has been noted previously based on RNA-seq data[75], but our data allow for more direct observations of both active transcription and associated regulatory elements.

Third, if most transcribed enhancers do indeed influence gene expression, then why are so many of them weakly maintained by natural selection? One possibility is that some of these enhancers have simply switched cell types, as has been reported in some cases[19]. Another possibility is that selection acts diffusely on enhancers across an entire locus, rather than strongly on individual enhancers, as has been proposed in cancer evolution[76]. A related idea is that, over long evolutionary time periods, it may be useful to maintain a collection of enhancers at each locus, even if all of them are not being used at any given time—they may, in a sense, serve as “spare parts" that can be mixed and matched, as opportunities allow, for cell-type and developmental-stage-specific functions. Similar suggestions have been made for noncoding transcription units such as eRNAs, upstream antisense RNAs, and lincRNAs[50,77], but it may hold as well for regulatory function. It will be possible to begin to evaluate some of these hypotheses as better data describing enhancers and enhancer-promoter interactions across many cell types become available for these and other groups of species.

369
370 **Methods:**
371
372 **Multiple species PRO-seq library generation.** *Isolation of primate CD4+ T-cells.* All human and
373 animal experiments were done in compliance with Cornell University IRB and IACUC guidelines.
374 We obtained peripheral blood samples (60-80 mL) from healthy adult male humans, chimpanzees,
375 and rhesus macaques. Informed consent was obtained from all human subjects. To account for
376 within-species variation in gene transcription we used three individuals to represent each primate
377 species. Blood was collected into purple top EDTA tubes. Human samples were maintained
378 overnight at 4C to mimic shipping non-human primate blood samples. Blood was mixed 50:50 with
379 phosphate buffered saline (PBS). Peripheral blood mononuclear cells (PBMCs) were isolated by
380 centrifugation (750x g) of 35 mL of blood:PBS over 15 mL Ficoll-Paque for 30 minutes at 20C. Cells
381 were washed three times in ice cold PBS. CD4+ T-cells were isolated using CD4 microbeads
382 (Miltenyi Biotech, 130-045-101 [human and chimp], 130-091-102 [rhesus macaque]). Up to 10^8
383 PBMCs were resuspended in binding buffer (PBS with 0.5% BSA and 2mM EDTA). Cells were
384 bound to CD4 microbeads (20uL of microbeads/ 10^7 cells) for 15 minutes at 4C in the dark. Cells
385 were washed with 1-2 mL of PBS/BSA solution, resuspended in 500uL of binding buffer, and passed
386 over a MACS LS column (Miltenyi Biotech, 130-042-401) on a neodymium magnet. The MACS LS
387 column was washed three times with 2mL PBS/BSA solution, before being eluted off the
388 neodymium magnet. Cells were counted in a hemocytometer.
389
390 *Isolation of CD4+ T-cells from mouse and rat.* Spleen samples were collected from one male mouse
391 (FVB) and one male rat (Albino Oxford) that had been sacrificed for IACUC-approved research not
392 related to the present study. Dissected spleen was mashed through a cell strainer using a sterile
393 glass pestle and suspended in 20 mL RPMI-1640. Cells were pelleted at 800xg for 3 minutes and
394 resuspended in 1-5mL of ACK lysis buffer for 10 minutes at room temperature to lyse red blood
395 cells. RPMI-1640 was added to a final volume 10 times that used for ACK lysis (10-40 mL). Cells
396 were pelleted at 800xg for 3 minutes, counted in a hemocytometer, and resuspended in RPMI-1640
397 to a final concentration of 250,000 cells per ml. CD4+ T-cells were isolated from splenocytes using
398 products specific for mouse and rat (Miltenyi Biotech, 130-104-453 [mouse], 130-090-319 [rat])
399 following instructions from Miltenyi Biotech, and as described above.
400
401 *T-cell treatment and PRO-seq library generation.* CD4+ T-cells were allowed to equilibrate in RPMI-
402 1640 supplemented with 10% FBS for 2-4 hours before starting experiments. Primate CD4+ T-cells
403 were stimulated with 25ng/mL PMA and 1mM Ionomycin (P/I or π) or vehicle control (2.5uL EtOH
404 and 1.66uL DMSO in 10mL of culture media). We selected the minimum concentrations which
405 saturate the production of IL2 and IFNG mRNA after 3 hours of treatment (data not shown). A 30
406 min. treatment duration was selected after observing a sharp increase in ChIP-qPCR signal for RNA
407 Pol II phosphorylated at serine 5 on the C-terminal domain on the IFNG promoter at 30 min. (data
408 not shown). To isolate nuclei, we resuspended cells in 1 mL lysis buffer (10 mM Tris-Cl, pH
409 8, 300 mM sucrose, 10 mM NaCl, 2 mM MgAc2, 3 mM CaCl2 and 0.1% NP-40). Nuclei were
410 washed in 10 mL of wash buffer (10 mM Tris-Cl, pH 8, 300 mM sucrose, 10 mM NaCl and 2 mM
411 MgAc2) to dilute free NTPs. Nuclei were washed in 1 mL, and subsequently resuspended in 50

412 μL, of storage buffer (50 mL Tris-Cl, pH 8.3, 40% glycerol, 5 mM MgCl2 and 0.1 mM EDTA), snap
413 frozen in liquid nitrogen and kept for up to 6 months before making PRO-seq libraries. PRO-
414 seq libraries were created exactly as described previously[45]. In most cases, we completed library
415 preps with one member of each species (usually one human, chimpanzee, and rhesus macaque) to
416 prevent batch effects from confounding differences between species. Samples were sequenced on
417 an Illumina Hi-Seq 2000 or NextSeq500 at the Cornell University Biotechnology Resource Center.
418
419 *Mapping PRO-seq reads.* We mapped PRO-seq reads using standard informatics tools. Our PRO-seq
420 mapping pipeline begins by removing reads that fail Illumina quality filters and trimming adapters
421 using cutadapt with a 10% error rate[78]. Reads were mapped with BWA[79] to the appropriate
422 reference genome (either hg19, panTro4, rheMac3, mm10, or rn6) and a single copy of the Pol I
423 ribosomal RNA transcription unit (GenBank ID# U13369.1). Mapped reads were converted to
424 bigWig format for analysis using BedTools[80] and the bedGraphToBigWig program in the Kent
425 Source software package[81]. The location of the RNA polymerase active site was represented by the
426 single base, the 3' end of the nascent RNA, which is the position on the 5' end of each sequenced
427 read. After mapping reads to the reference genome, three samples (one chimpanzee, U and PI, and
428 one rhesus macaque, PI) were identified as having poor data quality on the basis of the number of
429 uniquely mapped reads, and were excluded from downstream analysis. During comparative
430 analyses, the genomic coordinates of mapped reads were converted to the human assembly (hg19)
431 using CrossMap[82]. We converted genomic coordinates between genome assemblies using
432 reciprocal-best (rbest) nets[83]. We downloaded rbest nets for hg19-mm10, hg19-panTro4, hg19-rn6
433 from the UCSC genome browser. We created rbest nets for hg19-rheMac3 using the doRecipBets.pl
434 script provided as part of the Kent Source software package.
435
436 **Analysis of transcriptional regulatory elements.** *Defining a consensus set of transcriptional*
437 *regulatory elements.* We predicted TREs in each species' reference genome using dREG[52]. In all
438 cases, we combined the reads from all individuals for each species and T-cell treatment condition in
439 order to maximize power for the discovery of TREs. We then defined a consensus set of TRE
440 annotations, each of which bore the signature of an active TRE in at least one species and treatment
441 condition. To define such a set, dREG scores were first converted to human reference genome
442 (hg19) coordinates using CrossMap and the reciprocal-best nets. We then identified TREs in each
443 species separately by thresholding the dREG scores. Instead of simply merging overlapping TREs,
444 which tends to create large elements, we applied the thresholding procedure several times, taking
445 scores greater than 0.9, 0.8, and 0.7, and each time taking only TREs that did not intersect a
446 previously selected TRE at a higher threshold. Finally, the set of overlapping TREs from each
447 species were reduced to a single element containing the union of all positions covered by the set
448 using bedops, and sites within 500 bp of each other were further merged. We assigned each
449 putative TRE the maximum dREG score for each species and for each treatment condition.
450
451 *Identifying differences in TREs between species.* Differences in TRE transcription in 3-way (human-
452 chimp-rhesus macaque) or 5-way (human-chimp-rhesus macaque-mouse-rat) species comparisons
453 were identified using a combination of heuristics and statistical tests. Starting with the consensus
454 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 0.1 in at least one species and greater than 0.7 in at least one other species. 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 the Limma package[84] and thresholding at a 5% false discovery rate (as described below). These TREs were often active in all species. Changes in TRE activities were compared to histone modification ChIP-seq and DNase-I-seq data from the Epigenome Roadmap project [55].

TRE classification. For some analyses, TREs were classified into likely promoters and enhancers on the basis of their distance from known gene annotations (RefGene) and the predicted stability of the resulting transcription unit (TU). TRE classes of primary interest include (see also **Supplementary Fig. 7**): (1)promoters: near an annotated transcription start site (<500 bp) and producing a stable TU (instability score <0.1); (2)enhancers: distal to an annotated transcription start site (>10,000 bp) and producing an unstable TU (instability score >0.1). TU stability was defined using the posterior probability that a TRE yields an unstable transcription unit using the forward-backward tables in a hidden Markov model we described recently[50]. This approach produced a score for both the forward and the upstream divergent TU of each TRE. We took the minimum of these instability scores to represent the stability of that TRE.

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 context. We compared changes as a function of distance from the nearest annotated transcription start site in RefSeq. 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 in a variety of functional associations, including loop interactions, within the same topological associated domain, and in superenhancers. H3K4me2 ChIA-PET data describing loop interactions were downloaded from the Gene Expression Omnibus (GEO) database (GSE32677) and the genomic locations of loops were converted from hg18 to hg19 coordinates using the liftOver tool. Looped enhancers were defined as those within 5,000 bp of a loop center. Topological associated domains (TADs) based on Hi-C data for GM12878 cells were also downloaded from GEO (GSE63525). Superenhancers in CD4+ T-cells were taken from the supplementary data for ref. [69].

Refining dREG peak calls using dREG-HD. During analyses on 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. Testing the optimized dREG-HD (including both DNase-I imputation and peak calling) on GM12878, a GRO-seq dataset completely held out from model training and parameter optimization, revealed 82% sensitivity for DNase-I peaks within dREG sites at a 10% false discovery rate (FDR).

DNA sequence analysis. *DNA sequence conservation analysis.* For our evolutionary conservation analysis, we used phyloP scores[60] based on either (1) the primate species in the 46-way alignments, or (2) the 100-way genome alignments, both 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. We focused on 1,964 human TF binding motifs from the CisBP database[61] and clustered motifs using an affinity propagation algorithm into 567 maximally distinct DNA binding specificities (see ref[62]). We used TFBSs having a $\log_e$-odds score >10 in any of the primate reference genomes, with scores obtained by comparing each candidate motif model to a third-order Markov background model using the RTFBSDB package[62].

Motif enrichment in TREs that change during CD4+ T-cell activation. Motif enrichment analyses were completed using RTFBSDB[62] as described above, except that motifs were clustered into 621 maximally distinct DNA binding specificities (see ref[62]). We selected the motif whose canonical transcription factor is most highly transcribed in human CD4+ T-cells to represent each cluster. We defined a motif cutoff $\log_e$ odds ratio of 7.5 in a sequence compared with a third-order Markov model as background. Motifs enriched in up- or down-regulated dREG-HD TREs during CD4+ T-cell activation (>8-fold in magnitude and $p < 0.01$) 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 (<2-fold in magnitude and $p > 0.1$) using the *enrichmentTest* function in RTFBSDB[62].

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 into 567 clusters having highly similar binding preferences, 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[65]. We passed INSIGHT either complete DHSs, dREG-HD sites, or TFBS (log-odds score >7) 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 2,975 dREG-HD sites with 13,843 separate regions (median length of 29 bp) after merging overlapping TFBSs with a log-odds score greater than 7. We also analyzed 27 transcription factors each of which has more than 1,000 occurrences in dREG-HD sites that change on the human branch. All analyses were conducted using the INSIGHT web server (http://compgen.cshl.edu/INSIGHT/) with the default settings enabled.

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) to capture the distribution of PRO-seq read densities similar to those we have recently published[49,85]. Three states were used to represent background (i.e., outside of a transcription unit), 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 distribution 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

583 the transformed coordinates back to genomic coordinates. Source code for our implementation is
584 publicly available on GitHub: https://github.com/andrelmartins/tunits.nhp.
585
586 *Inferring TU boundaries in the common great ape ancestor.* We identified the most likely TU
587 boundaries in the great ape ancestor by maximum parsimony. TUs were identified and compared
588 in human reference coordinates (hg19) for all species. We used the bedops package to mark the
589 intersection between each pair of species (i.e., human-chimp, human-rhesus macaque, and chimp-
590 rhesus macaque). Intersections (>= 1bp) between pairs of species were merged, resulting in a
591 collection of TUs shared by any two pairs of species, and therefore likely to be a TU in the human-
592 chimp ancestor. All steps were applied independently on the plus and minus strands. These steps
593 identified 32,602 putative TUs active in CD4+ T-cells of the primate ancestor. We added 13,085
594 TUs that did not overlap ancestral TUs but were found in any one of the three primate species.
595
596 *Transcription unit classification.* TUs were classified into by annotation type using a pipeline similar
597 to ones that we have described recently[49,85,86]. Before classifying TUs we applied a heuristic to
598 refine TUs on the basis of known annotations. TUs that completely overlap multiple gene
599 annotations were broken at the transcription start site provided that a dREG site overlapped that
600 transcription start site. Classification was completed using a set of rules to iteratively refine
601 existing annotations, as shown in **Supplementary Fig. 13A**. Unless otherwise stated, overlap
602 between a TU and a transcript annotation was defined such that >50% of a TU matched a gene
603 annotation and covers at least 50% of the same annotation. TUs overlapping GENCODE annotations
604 (>50% overlap, defined as above) were classified using the biotype in the GENCODE database into
605 protein coding, lincRNA (lincRNA or processed transcript), or pseudogene. The remaining
606 transcripts were classified as annotated RNA genes using GENCODE annotations, the rnaGenes
607 UCSC genome browser track (converted from hg18 to hg19 coordinates), and miRBase v20 [87]. As
608 many RNA genes are processed from much longer TUs, we required no specific degree of overlap
609 for RNA genes. Upstream antisense (i.e., divergent) TUs were classified as those within 500bp of
610 the transcription start site of any GENCODE or higher level TU annotation (including lincRNAs).
611 Antisense transcripts were defined as those with a high degree of overlap (>50%) with annotated
612 protein coding genes in the opposite orientation. The remaining transcripts with a high degree of
613 overlap (>50%) to annotated repeats in the repeatmasker database (rmsk) were classified as
614 repeat transcription. Finally, any TUs still remaining were classified as unannotated, and were
615 further divided into those which are intergenic or that partially overlapping existing annotations.
616
617 **Comparing transcription between conditions and species.** *Comparing transcription before and*
618 *after CD4+ T-cell activation.* We compared π treated and untreated CD4+ T-cells within each of the
619 primate species using gene annotations (GENCODE v19). We counted reads in the interval between
620 500 bp downstream of the annotated transcription start site and either the end of the gene or
621 60,000 bp into the gene body (whichever was shorter). This window was selected to avoid (1)
622 counting reads in the pause peak near the transcription start site, and (2) to focus on the 5' end of
623 the gene body affected by changes in transcription during 30 minutes of π treatment assuming a
624 median elongation rate of 2 kb/ minute[49,88]. We limited analyses to gene annotations longer than
625 1,000 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[53].

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. 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)[83]. 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[80,89]. Differential expression was conducted between species using the LIMMA package for R[84].

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

Reviewer link:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=efivmcssnpezzop&acc=GSE85337

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

References (Paperpile):

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
2. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165,** 349–357 (1969).
3. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
4. Rockman, M. V. *et al.* Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol.* **3,** e387 (2005).
5. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321,** 1346–1350 (2008).
6. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20130025 (2013).
7. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471,** 216–219 (2011).
8. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154,** 185–196 (2013).
9. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45,** 723–729 (2013).
10. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163,** 68–83 (2015).
11. Siepel, A. & Arbiza, L. Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.* **29,** 81–89 (2014).
12. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206–216 (2007).
13. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322,** 434–438 (2008).
14. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
15. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39,** 1140–1144 (2007).
16. Torgerson, D. G. *et al.* Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* **5,** e1000592 (2009).
17. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
18. Ballester, B. *et al.* Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife* **3,** e02626 (2014).
19. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346,** 1007–1012 (2014).
20. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat. Genet.* **46,** 685–692 (2014).
21. Doniger, S. W. & Fay, J. C. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput. Biol.* **3,** e99 (2007).
22. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in

transcription factor binding in yeast. *Nature* **464,** 1187–1191 (2010).
23. Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol.* **8,** e1000343 (2010).
24. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
25. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351,** 1083–1087 (2016).
26. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461,** 186–192 (2009).
27. Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol.* **15,** 547 (2014).
28. Cain, C. E., Blekhman, R., Marioni, J. C. & Gilad, Y. Gene expression differences among primates are associated with changes in a histone epigenetic modification. *Genetics* **187,** 1225–1234 (2011).
29. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149,** 1381–1392 (2012).
30. Paris, M. *et al.* Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genet.* **9,** e1003748 (2013).
31. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* **10,** e1004226 (2014).
32. Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* **25,** 167–178 (2015).
33. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23,** 1210–1223 (2013).
34. Domené, S. *et al.* Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368,** 20130027 (2013).
35. Wunderlich, Z. *et al.* Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. *Cell Rep.* **12,** 1740–1747 (2015).
36. Cannavò, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol.* **26,** 38–51 (2016).
37. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403,** 564–567 (2000).
38. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
39. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10,** 1297–1309 (2015).
40. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* **342,** 1100–1104 (2013).
41. Bauernfeind, A. L. *et al.* Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. *Genome Biol. Evol.* **7,** 2276–2288 (2015).
42. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347,** 664–667 (2015).
43. Pai, A. A. *et al.* The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genet.* **8,** e1003000 (2012).
44. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–1848 (2008).
45. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 (2013).
46. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373 (2011).

47. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161,** 526–540 (2015).
48. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11,** 1455–1476 (2016).
49. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145,** 622–634 (2011).
50. 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).
51. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5,** 5336 (2014).
52. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12,** 433–438 (2015).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
54. Nguyen, D. H., Hurtado-Ziola, N., Gagneux, P. & Varki, A. Loss of Siglec expression on T lymphocytes during human evolution. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 7765–7770 (2006).
55. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
56. Macian, F. NFAT proteins: key regulators of T-cell development and function. *Nat. Rev. Immunol.* **5,** 472–484 (2005).
57. Ouyang, W., Beckett, O., Flavell, R. A. & Li, M. O. An essential role of the Forkhead-box transcription factor Foxo1 in control of T cell homeostasis and tolerance. *Immunity* **30,** 358–371 (2009).
58. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44,** 148–156 (2012).
59. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21,** 1273–1283 (2011).
60. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
61. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158,** 1431–1443 (2014).
62. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw338
63. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
64. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
65. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* **30,** 1159–1171 (2013).
66. Sun, M., Gadad, S. S., Kim, D.-S. & Kraus, W. L. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol. Cell* **59,** 698–711 (2015).
67. 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.* **22,** 490–503 (2012).
68. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
69. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
70. Barreiro, L. B., Marioni, J. C., Blekhman, R., Stephens, M. & Gilad, Y. Functional comparison of

innate immune signaling pathways in primates. *PLoS Genet.* **6,** e1001249 (2010).
71. Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. & White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440,** 242–245 (2006).
72. Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K. & Gilad, Y. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* **4,** e1000271 (2008).
73. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8,** e1002841 (2012).
74. Lewis, J. J., van der Burg, K. R. L., Mazo-Vargas, A. & Reed, R. D. ChIP-Seq-Annotated Heliconius erato Genome Highlights Patterns of cis-Regulatory Evolution in Lepidoptera. *Cell Rep.* **16,** 2855–2863 (2016).
75. Gilad, Y., Oshlack, A. & Rifkin, S. A. Natural selection on gene expression. *Trends Genet.* **22,** 456–461 (2006).
76. Bailey, S. D. *et al.* Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. *Nat. Genet.* **48,** 1260–1266 (2016).
77. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155,** 990–996 (2013).
78. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
81. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14,** 144–161 (2013).
82. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30,** 1006–1007 (2014).
83. Kent, W. J., 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.* **100,** 11484–11489 (2003).
84. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
85. 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).
86. Luo, X., Chae, M., Krishnakumar, R., Danko, C. G. & Kraus, W. L. Dynamic reorganization of the AC16 cardiomyocyte transcriptome in response to TNFα signaling revealed by integrated genomic analyses. *BMC Genomics* **15,** 155 (2014).
87. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42,** D68–73 (2014).
88. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* **50,** 212–222 (2013).
89. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28,** 1919–1920 (2012).

Acknowledgements: We thank M. Jin for assistance in establishing the magnetic separation of CD4+ T-cells, 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), NHLBI (National Heart, Lung, and Blood Institute) grant UHL129958A to CGD and JTL, NIGMS (National Institute of General Medical Sciences) grant GM102192 to AS, and NHGRI (National Human Genome Research Institute) grant HG0070707 to AS and JTL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

Author contributions: CGD, EJR, and ETW performed CD4+ T-cell extraction, validation, and PRO-seq experiments. CGD, ZW, TC, and ALM analyzed the data. CGD, AS, JTL, and WLK 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 availiable on GitHub: https://github.com/Danko-Lab/CD4-Cell-Evolution.

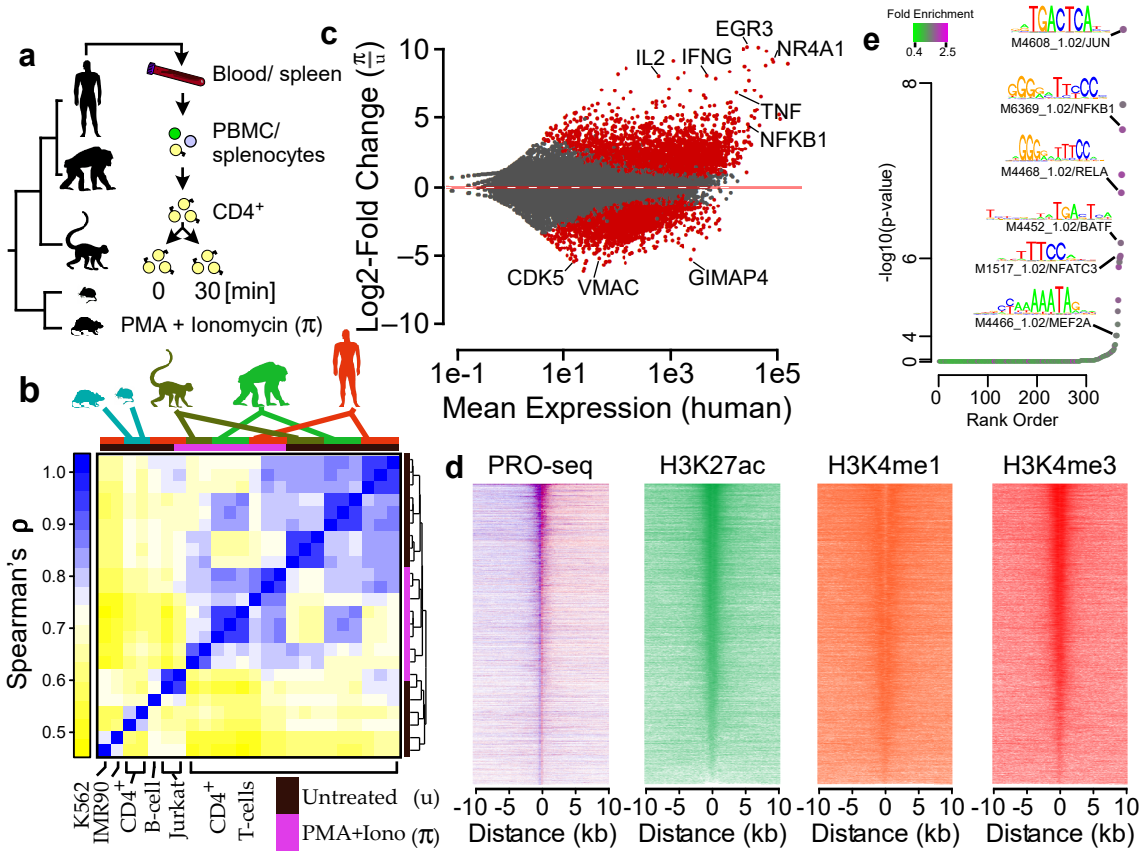

Fig. 1 | Maps of primary transcription in CD4+ T-cells. **(a)** CD4+ T-cells were isolated from the blood or spleen of individuals from five vertebrate species, including human, chimpanzee, rhesus macaque, mouse, and rat. **(b)** Hierarchical clustering of PRO-seq signal intensities in gene bodies groups 19 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 log2 fold-change following π treatment (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, TNF, and EGR3) are marked. **(d)** Heatmaps show the distribution of PRO-seq (red and blue indicate transcription on the plus and minus strand, respectively), H3K27ac, H3K4me1, and H3K4me3 ChIP-seq signal intensity. Plots are centered on transcriptional regulatory elements (TREs) predicted in untreated human CD4+ T-cells using dREG-HD (see Methods), including both promoters and enhancers. All plots are ordered based on H3K27ac ChIP-seq intensity. **(e)** Enrichment of TF binding motifs in TREs that increase transcription levels following π treatment compared to TREs transcribed in both conditions. Plot shows the -log 10 p-value, based on a Fisher's exact test and Bonferroni correction (y-axis), as a function of the rank order of enriched motifs (x-axis). The color scale denotes the degree of enrichment. Motifs, the Cis-BP ID number, and the target transcription factor for several outliers are shown in the plot.

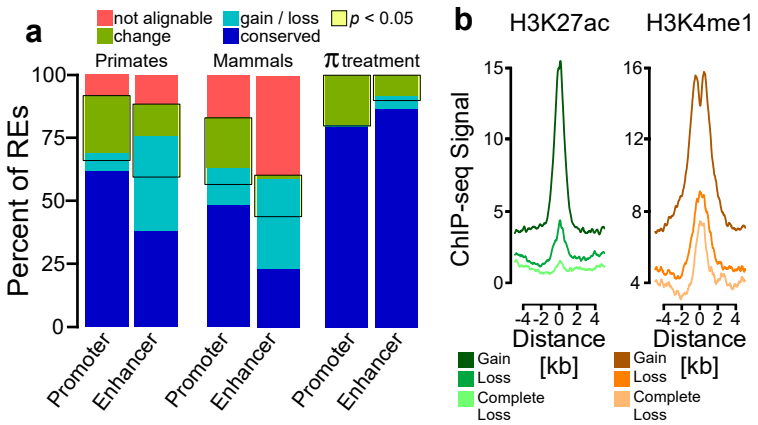

Fig. 2 | Frequency of changes in TRE transcription. (a) The fractions of TREs that are present in the human reference genome and are conserved across all species (blue), are not detectable and are therefore inferred as gains or losses (teal) or undergo significant changes (green) in at least one species, or fall in regions for which no ortholog occurs in at least one of the indicated genomes (pink). Plots labeled "Primate" illustrate frequency of changes in a three-way comparison of human, chimpanzee, and rhesus macaque, whereas those labeled "Mammal" summarize a five-way comparison also including rat and mouse. p treatment denotes a similar comparison between human untreated and PMA+Ionomycin treated CD4+ T-cell samples. A yellow box shows the complete fraction of TREs that undergo statistically significant changes in Pol II abundance (FDR-corrected p-value < 0.05). **(b)** ChIP-seq signal for H3K27ac (green) and H3K4me1 (orange) near dREG sites classified as gains, losses, or complete losses of TRE signal (dREG score < 0.1) on the human branch.

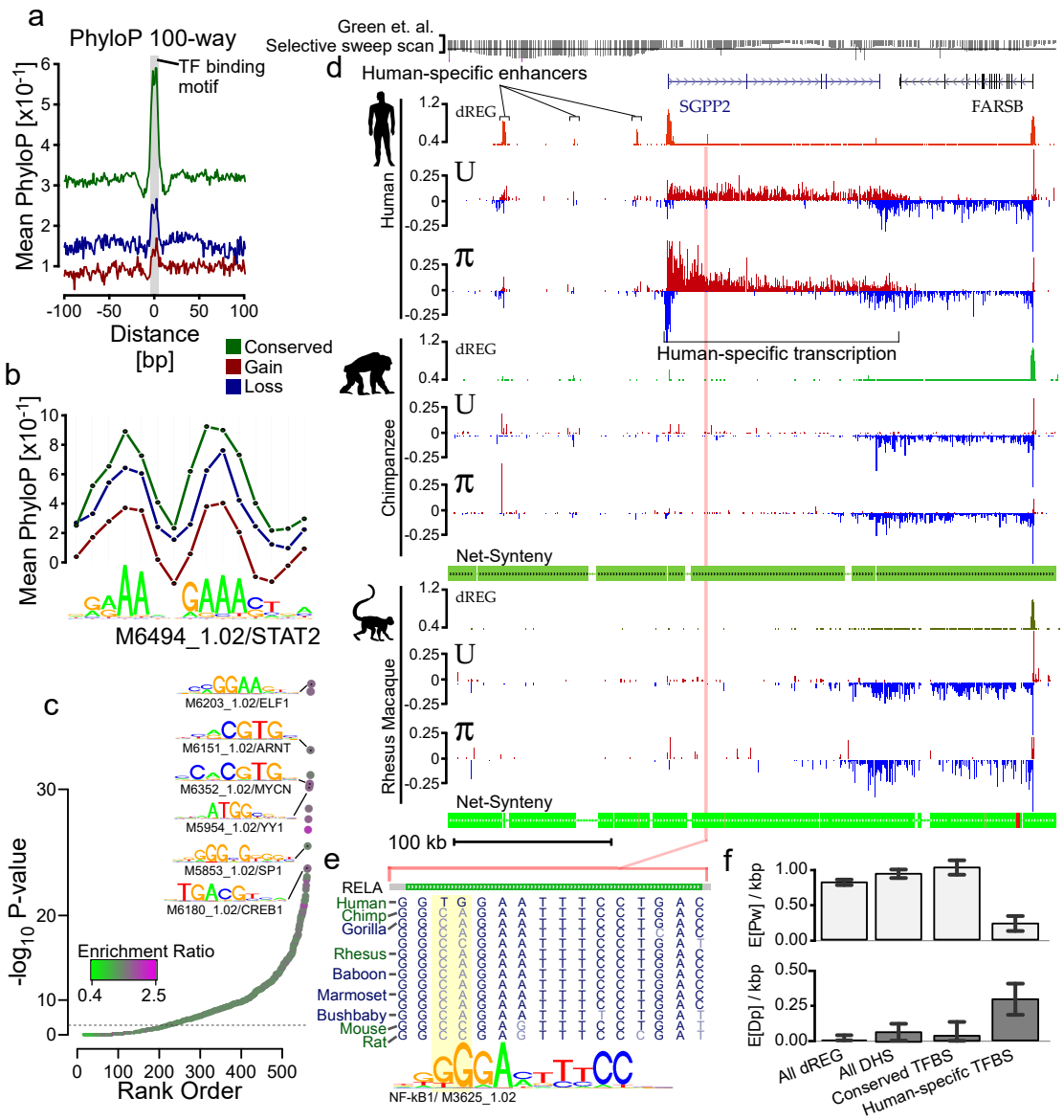

Fig. 3 | Evolutionary changes in TRE transcription correlate with DNA sequence conservation. **(a)** Mean phyloP scores near transcription factor binding motifs in dREG-HD sites that are conserved (green), gained (red), or lost (blue) on the human branch. PhyloP scores were computed based on the 100-way species alignments in the UCSC genome browser. **(b)** PhyloP scores that fall within the binding motif recognized by STAT2 (M6494_1.02). **(c)** Motifs were ranked based on the enrichment of DNA sequence differences found inside of dREG-HD sites that change in primate species compared to those which do not change. Plot shows the Bonferroni-corrected -log 10 p-value (y-axis) as a function of motif rank order (x-axis). **(d)** UCSC Genome Browser track shows transcription near SGPP2 and FARSB in untreated (U) and PMA+ionomycin (p) treated CD4+ T-cells isolated from the indicated primate 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. The Green et. al. (ref63) 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. **(e)** Multiple species alignment shows DNA sequence near one dREG site that contains a strong match to the NF-κB binding motif in human. This motif occurrence is bound by NF-κB subunit RELA based on ChIP-seq data in multiple ENCODE cell lines (green box). **(f)** INSIGHT estimates of the expected number of segregating polymorphisms under weak negative selection (E[Pw]/kbp) or the expected number of human nucleotide substitutions driven by positive selection (E[Dp]/kbp) in human populations in the indicated class of dREG-HD sites. All estimates are in units of nucleotides per kilobase.

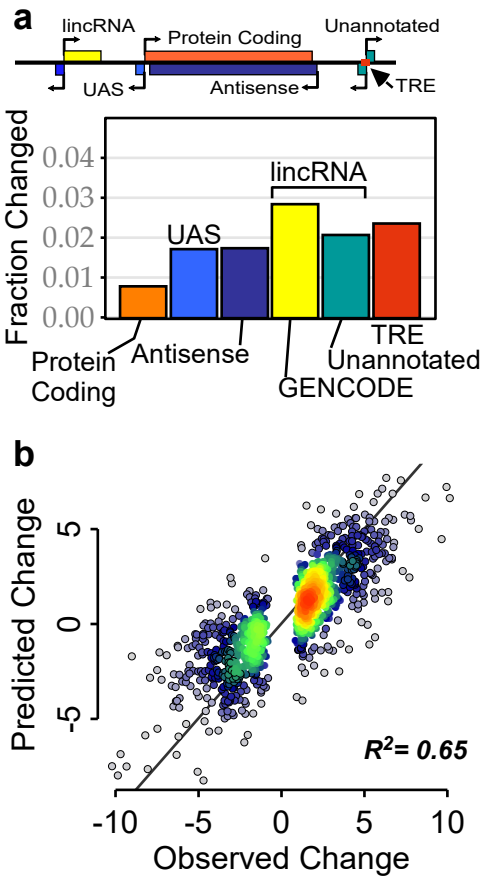

Fig. 4 | Changes in non-coding RNA transcription predict changes in gene transcription. (a) The fraction of each indicated class of RNAs that undergo changes in transcription in human CD4+ T-cells (see Methods). The relationships among the indicated classes of transcription units are depicted at top. **(b)** Scatterplot shows the magnitude of changes in transcription predicted for 928 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.

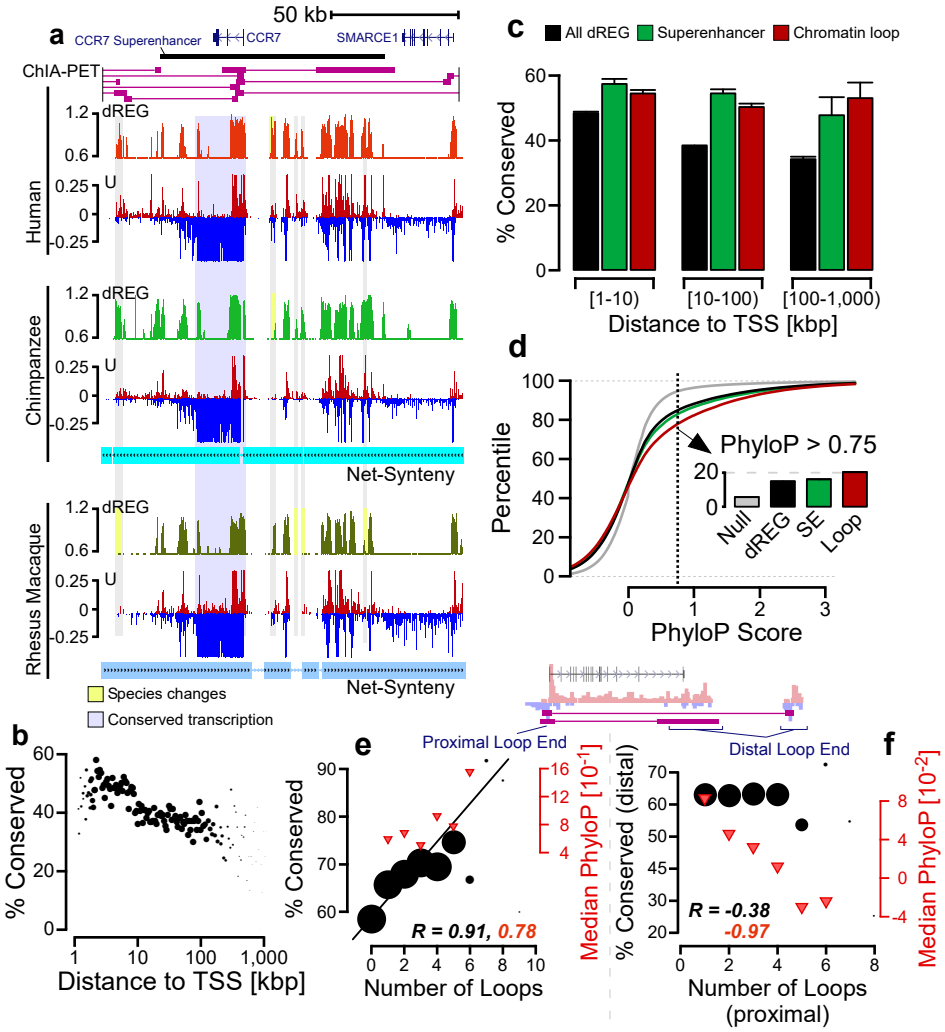

Fig. 5 | Stabilizing selection on protein-coding gene transcription. **(a)** UCSC Genome Browser tracks show transcription, dREG signal, and ChIA-PET loop interactions near the CCR7 superenhancer in the human genome. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands in units 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 from the nearest annotated 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. **(d)** 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. **(e)** Scatterplot shows promoter conservation (y-axis, black) or DNA sequence conservation (y-axis, red) as a function of the number of loop interactions made by that site to distal sites across the genome (x-axis). **(f)** TRE conservation (y-axis) or DNA sequence conservation (y-axis, red) as a function of the number of loop interactions made by the distal sequence element (x-axis). In panels (e-f) the size of each point is proportional to the number of examples in the corresponding bin.