

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

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

7 <sup>1</sup> Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

8 <sup>2</sup> Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

9 <sup>3</sup> Graduate field of Computational Biology, Cornell University, Ithaca, NY 14853.

10 <sup>4</sup> Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724.

11 <sup>5</sup> Department of Microbiology & Immunology, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

12 <sup>6</sup> Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853.

13 <sup>7</sup> Laboratory of Signaling and Gene Regulation, Cecil H. and Ida Green Center for Reproductive Biology Sciences,  
14 University of Texas Southwestern Medical Center, Dallas, TX 75390.

15 <sup>8</sup> Division of Basic Research, Department of Obstetrics and Gynecology, University of Texas Southwestern Medical  
16 Center, Dallas, TX 75390.  
17

## 18 \* **Address correspondence to:**

19 Charles G. Danko, Ph.D.  
20 Baker Institute for Animal Health  
21 Cornell University  
22 Hungerford Hill Rd.  
23 Ithaca, NY 14853  
24 Phone: 607-256-5620  
25 E-mail: [dankoc@gmail.com](mailto:dankoc@gmail.com)  
26  
27

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](mailto:asiepel@cshl.edu)

## 28 **Abstract:**

29 Transcriptional regulatory changes have been shown to contribute to phenotypic differences  
30 between species, but many questions remain about how gene expression evolves. Here we report  
31 the first comparative study of nascent transcription in primates. We used PRO-seq to map actively  
32 transcribing RNA polymerases in resting and activated CD4+ T-cells in multiple human,  
33 chimpanzee, and rhesus macaque individuals, with rodents as outgroups. This approach allowed us  
34 to measure transcription separately from post-transcriptional processes. We observed general  
35 conservation in coding and non-coding transcription, punctuated by numerous differences between  
36 species, particularly at distal enhancers and non-coding RNAs. We found evidence that  
37 transcription factor binding sites are a primary determinant of transcriptional differences between  
38 species, that stabilizing selection maintains gene expression levels despite frequent changes at  
39 distal enhancers, and that adaptive substitutions have driven lineage-specific transcription. Finally,  
40 we found strong correlations between evolutionary rates and long-range chromatin interactions.  
41 These observations clarify the role of primary transcription in regulatory evolution.

42           Following decades of speculation that changes in the regulation of genes could be a potent  
43 force in the evolution of form and function<sup>1-3</sup>, investigators have now empirically demonstrated the  
44 evolutionary importance of gene regulation across the tree of life<sup>4-12</sup>. Changes in gene expression  
45 are primarily driven by mutations to non-coding DNA sequences, particularly those that bind  
46 sequence-specific transcription factors<sup>13</sup>. Accordingly, adaptive nucleotide substitutions at  
47 transcription factor binding sites (TFBSs)<sup>9,10,14-16</sup> and gains and losses of TFBSs<sup>17-25</sup> both appear to  
48 make major contributions to the evolution of gene expression. These events are believed to modify  
49 a variety of rate-limiting steps early in transcriptional activation<sup>26</sup>. In addition, transcriptional  
50 activity is generally correlated with various epigenomic and structural features, including post-  
51 translational modifications to core histones, the locations of architectural proteins such as CTCF,  
52 and the organization of topological associated domains. Like TFBSs, these features display general  
53 conservation across species, yet do exhibit some variation, with differences between species  
54 roughly proportional to evolutionary distance<sup>27</sup>. Moreover, differences between species in these  
55 features correlate with differences in gene expression<sup>8,24,28-30</sup>.

56           Nevertheless, many open questions remain about the roles of TFBSs, chromatin  
57 organization, and posttranscriptional regulation in the evolution of gene expression. For example,  
58 there is a surprisingly limited correlation between differences in binding events and differences in  
59 mRNA expression levels<sup>31-33</sup>. Possible reasons for this discordance include non-functional TF  
60 binding<sup>31,32,34</sup>, compensatory gains and losses of TFBSs<sup>20,35-38</sup>, difficulties associating distal  
61 enhancers with target genes<sup>39</sup>, and a dependency of TF function on chromatin or chromosomal  
62 organization<sup>40</sup>. In addition, some changes in mRNA expression appear to be "buffered" at the post-  
63 transcriptional level<sup>41-43</sup>. Finally, it remains unclear to what degree epigenomic differences  
64 between species are causes and to what degree they are effects of differences in gene expression.

65           One reason why it has been difficult to disentangle these contributions to gene expression is  
66 that expression is typically measured in terms of the abundance of mRNA, which is subject to post-  
67 transcriptional processing<sup>44</sup> and therefore is an indirect measure of the transcription of genes by  
68 RNA polymerase II. An alternative and complementary approach is to measure the production of  
69 nascent RNAs using Precision Run-On and sequencing (PRO-seq) and related technologies<sup>45-50</sup>.  
70 These nascent RNA sequencing methods directly measure active transcription and are highly  
71 sensitive to immediate and transient transcriptional responses to stimuli<sup>51</sup>. In addition, they can  
72 detect active regulatory elements as well as target genes, because these elements themselves  
73 display distinctive patterns of transcription, which are highly underrepresented in RNA-seq data  
74 owing to their rapid degradation<sup>34,52,53</sup>. Indeed, the latest nascent RNA sequencing methods, such as  
75 PRO-seq<sup>46</sup>, in combination with new computational tools for regulatory element prediction<sup>54</sup>, serve  
76 as powerful single-assay systems for both identifying regulatory elements and measuring  
77 transcription levels.

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

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

84 We developed nucleotide-resolution maps of RNA polymerase in CD4+ T-cells isolated from  
85 five mammalian species. Samples were collected under resting and activated conditions from three  
86 unrelated individuals representing each of three primate species—humans, chimpanzees, and  
87 rhesus macaques—spanning ~25-30 million years of evolution (MYR) (**Fig. 1a**). To compare with  
88 studies that focus on longer evolutionary branch lengths, we also collected resting samples from a  
89 single individual in each of two rodent species—mouse and rat—which together serve as an  
90 outgroup to the primates (~80 MYR divergence). We used flow cytometry to validate the purity of  
91 isolated CD4+ cells (**Supplementary Fig. 1**). In addition, we used measurements of transcriptional  
92 activity of T-cell subset markers for T-helper type 1 (Th1), Th2, Th17, T-regulatory, and T-follicular  
93 helper cells to demonstrate that the population of CD4+ T-cell subsets within the total CD4+  
94 population was largely similar across species (**Supplementary Fig. 2**).

95 PRO-seq<sup>46,49</sup> libraries were sequenced to a combined depth of ~1 billion uniquely mapped  
96 reads (~78-274 million per species) (**Supplementary Table 1**). According to a principal  
97 component analysis, the first, second, and third sources of variation in the complete dataset were,  
98 respectively, the rodent vs. primate species, variation across the primate species, and the treatment  
99 condition (**Supplementary Fig. 3**). Similarly, hierarchical clustering of these data grouped the  
100 primate samples first by cell type or treatment condition and subsequently by species, with the  
101 rodent samples as outgroups that were more similar to untreated primate CD4+ T-cells than to  
102 other samples included in the comparison (**Fig. 1b**). The correlation between untreated samples  
103 decreased linearly with evolutionary time (**Supplementary Fig. 4**), consistent with reports that  
104 differences between species arise, on average, at a roughly constant evolutionary rate<sup>27,55</sup>.

105 To gain further insight into the evolution of the response to CD4+ T-cell stimulation, we  
106 compared transcriptional activity under resting and activated conditions within and between  
107 species. Here we focused on 42,556 GENCODE-annotated transcription units (TUs) best supported  
108 by PRO-seq data for human CD4+ T-cells<sup>56</sup>. In humans, we found that PMA and ionomycin ( $\pi$ )  
109 significantly altered the transcription levels of 6,940 (13%) of these TUs ( $p < 0.01$ , deSeq2<sup>57</sup>; **Fig.**  
110 **1c**). Parallel analyses in chimpanzee and rhesus macaque revealed many similarities in  
111 transcriptional changes following  $\pi$  treatment (**Supplementary Fig. 5a-b**). We identified a core set  
112 of 3,157 TUs that undergo evolutionarily conserved transcriptional changes in all three species  
113 following 30 min. of  $\pi$ -treatment, including many of the classical response genes (e.g., IFNG, TNF $\alpha$ ,  
114 IL2, and IL2RA), as well as numerous novel genes and lincRNAs (**Supplementary Fig. 5a-c**). Active  
115 transcriptional regulatory elements (TREs) undergoing changes following  $\pi$ -treatment were  
116 enriched for a similar set of transcription factor binding motifs across species, including those for  
117 NF-kB and the AP-1 heterodimers FOS and JUN, which are known to be activated by canonical T-cell  
118 receptor signaling (**Fig. 1d**; **Supplementary Note 1**). Thus, the core regulatory principles  
119 responsible for T-cell signaling and activation appear to remain broadly conserved across primate  
120 evolution.

121

## 122 **Rapid evolutionary changes in transcribed enhancers**

123 We used dREG<sup>54</sup> to identify 30,357 active TREs in human CD4+ T-cells, based on patterns of  
124 enhancer-templated RNA (eRNA) or upstream antisense (uarNA) transcription evident from PRO-  
125 seq data (**Online Methods**). We classified these predicted TREs as either protein-coding promoters

126 or candidate enhancers based on their proximity to gene annotations. The predicted TREs in each  
127 group were highly concordant with other marks of regulatory function in human CD4+ T-cells used  
128 previously to define groups of candidate enhancers, including acetylation of histone 3 lysine 27  
129 (H3K27ac), mono- and trimethylation of histone 3 lysine 4 (H3K4me1 and H3K4me3), and DNase-I-  
130 seq signal<sup>58</sup> (**Fig. 1e**). Notably, dREG identified >83% of DNase-I hypersensitive sites (DHS) marked  
131 by H3K27 acetylation in human CD4+ T-cells, consistent with prior work suggesting that  
132 transcription patterns alone can recover the majority of active enhancers defined by independent  
133 criteria<sup>34,52,54</sup>. Furthermore, we identified 88% and 91% of DHSs marked, respectively, by H3K9ac  
134 and H4K16ac, two other markers of regulatory function. Taken together, these data suggest that  
135 PRO-seq patterns reveal the locations of TREs with high sensitivity. In the analysis that follows, we  
136 will refer to these dREG-identified distal TREs as “enhancers” for simplicity.

137 Extending our dREG analysis to untreated CD4+ T-cells from additional species revealed  
138 71,748 TREs that were active in untreated T-cells in at least one species (ranging between 27,581  
139 and 39,387 TREs in each species). We defined two types of changes between species: (1) changes  
140 in the abundance of Pol II at TREs that were present across all species, and (2) complete gains or  
141 losses in at least one species (see **Supplementary Note 2, Supplementary Fig. 6, and Online**  
142 **Methods**). We found that 52% of enhancers showed evidence of changes in Pol II transcriptional  
143 activity in at least one of the three primate species and 81% showed changes at the longer  
144 evolutionary distance between primates and rodents (**Fig. 2a**), similar to recent observations in  
145 other systems<sup>10,24</sup> (**Supplementary Note 3**). Enhancers were predicted to be completely gained or  
146 lost at nearly eight times the rate of promoters (35% of enhancers; 12% of promoters;  $p < 2.2e-16$   
147 by Fisher’s exact test), consistent with recent observations based on H3K27ac and H3K4me3<sup>24</sup>. By  
148 contrast, TREs induced by  $\pi$  treatment were much more likely to be conserved, and showed similar  
149 conservation at promoters and enhancers (**Fig. 2a**).

150 Next we tested whether evolutionary changes in transcriptional activity correlate with the  
151 enrichment of other marks of active enhancers. Predicted lineage-specific human enhancers were  
152 enriched for both active and repressive enhancer marks (**Fig. 2b; Supplementary Fig. 7**). Whereas  
153 apparent human gains were enriched for high levels of the active enhancer marker H3K27ac, sites  
154 with reduced transcriptional activity in humans showed much lower enrichments of H3K27ac.  
155 Furthermore, locations at which the dREG signal was completely lost in a human-specific fashion  
156 displayed levels of H3K27ac approaching those of randomly selected background sites (**Fig. 2b**).  
157 Intriguingly, many of the losses on the human branch retained H3K4me1, which marks both active  
158 and inactive enhancers<sup>59,60</sup>, and these losses displayed higher levels of chromatin marks associated  
159 with transcriptional repression than a random background (**Fig. 2b**), indicating that, at least in  
160 some cases, an active ancestral primate enhancer retains a ‘poised’ chromatin state in human,  
161 despite losing both transcriptional activity and H3K27ac. Thus, evolutionary changes in poised and  
162 active marks may commonly occur as distinct events.

163

### 164 **Transcriptional changes correlate with DNA sequence differences**

165 To investigate whether changes in TRE activity are accompanied by changes in DNA  
166 sequence, we compared phyloP sequence conservation scores<sup>61</sup> at transcriptionally conserved TREs  
167 with phyloP scores at TREs that display evolutionary changes in transcription. Because signatures  
168 of sequence conservation in TREs are likely to be most pronounced in transcription factors binding

169 sites (TFBS), we restricted our sequence conservation analyses to matches to 567 clusters of TF  
170 binding motifs selected based on their distinct DNA binding specificities (see **Online Methods**)<sup>62,63</sup>.  
171 We required that motif matches were present in dREG sites, and adopted a threshold at which  
172 nearly half of the motifs discovered were bona fide TFBSs, as measured by ChIP-seq (positive  
173 predictive value [PPV]= 0.47).

174 TFBSs found in transcriptionally conserved dREG sites showed a marked enrichment for  
175 higher phyloP scores relative to surrounding regions, indicating local sequence conservation (**Fig.**  
176 **3a**). By contrast, TFBSs in lineage-specific dREG sites had substantially reduced enrichments in  
177 phyloP scores (**Fig. 3a, cyan/blue**). Notably, TFBSs in dREG sites lost on the human lineage  
178 showed enhanced conservation compared with those in human-specific gains. This observation is  
179 consistent with losses evolving under conservation in other mammalian species (which contribute  
180 to the phyloP scores) and gains emerging relatively recently. Each of these patterns was robust to  
181 corrections for potentially confounding differences in the distribution of sites, as well to choices of  
182 motif score thresholds (**Supplementary Fig. 8a**). Relaxing the motif score threshold to provide  
183 sensitivity for larger numbers of TFBSs at the expense of specificity, revealed patterns of  
184 conservation that correlate with the information content of positions within the DNA sequence  
185 motif (**Supplementary Fig. 8b**), further supporting TF binding as the functional property driving  
186 sequence conservation at these sites. Together, these analyses support the hypothesis that the  
187 sequences in TFBSs are a primary driver of transcriptional differences between species.

188 We searched for examples of DNA sequence differences that might be responsible for  
189 transcriptional changes following  $\pi$  treatment, hypothesizing that they might be characterized  
190 more easily than sequences responsible for transcription in the untreated condition, because these  
191 transcriptional changes were likely driven by a limited number of master TFs<sup>64</sup> (**Fig. 1e**). In one  
192 example, we found nucleotide substitutions in three apparent NF- $\kappa$ B binding sites in the proximal  
193 promoter and an internal enhancer of *SGPP2* that correlate with differences in *SGPP2* expression  
194 (**Fig. 3b; Supplementary Fig. 9**). Two of these putative binding sites were bound by NF- $\kappa$ B in  
195 human cell lines according to ChIP-seq data from ENCODE. Moreover, substitutions observed in  
196 human were found to disrupt the same position in the motif as NF- $\kappa$ B binding QTLs<sup>65</sup> (see **Online**  
197 **Methods**), and showed a general trend toward higher NF- $\kappa$ B binding in the human alleles  
198 (**Supplementary Fig. 9**). To test the hypothesis that observed DNA sequence changes produced  
199 differential transcriptional activity, we cloned DNA from each primate species into a reporter  
200 vector driving luciferase activity in an MCF-7 cell model, which recapitulates the primary  
201 transcriptional features of the *SGPP2* locus<sup>66</sup> (**Supplementary Fig. 9**). Differences in basal  
202 luciferase activity were generally concordant with those observed between species  
203 (**Supplementary Fig. 10**). Moreover, both the proximal promoter of *SGPP2* and the internal  
204 enhancer both activated luciferase expression more strongly following NF- $\kappa$ B activation when  
205 human DNA was cloned, but not with orthologous DNA from the other primates (**Fig. 3c**).

206 To determine whether these TREs affect the expression of *SGPP2* in its native genomic  
207 context, we silenced each TRE by using CRISPRi, which targets a catalytically dead CAS9 fused to  
208 the Krüppel-associated box repressor (dCAS9-KRAB), to specifically tri-methylate lysine 9 of  
209 histone 3 (H3K9me3)<sup>67</sup>. Three independent single-guide RNAs (sgRNAs) targeting the internal  
210 enhancer and two designed for the proximal promoter reduced *SGPP2* transcription to 50-60% of  
211 its resting level ( $p = 1.5e-3$  and  $2.6e-2$ , respectively, by a two-tailed t-test; **Fig. 3d**), consistent with

212 these TREs directly contributing to *SGPP2* expression in MCF-7 cells. Three sgRNAs targeting the  
213 upstream enhancer also had a significant effect on *SGPP2* expression ( $p = 1.8e-4$ ). Notably, the  
214 genome assemblies for chimpanzee and rhesus macaque harbor deletions of this upstream TRE that  
215 appear likely to affect its activity (**Supplementary Fig. 9**). However, although silencing individual  
216 enhancers reduced the transcription level of *SGPP2* following NF- $\kappa$ B activation, silencing individual  
217 enhancers was insufficient to completely abolish induction of *SGPP2* (**Fig. 3d**). Taken together, our  
218 findings show that at least two of the three TREs regulating *SGPP2* drove expression patterns  
219 matching PRO-seq data in a reporter assay, but none completely explained *SGPP2* activation *in situ*.  
220 These observations suggest that that multiple causal substitutions in NF- $\kappa$ B binding sites may work  
221 in concert to achieve human-specific activation of *SGPP2*.

222 In several cases, as with *SGPP2*, we observed numerous nucleotide substitutions within  
223 individual or clustered TFBSs. These clusters of substitutions are highly unlikely to occur by chance  
224 and suggest that positive selection may have driven adaptation of these binding sites. Indeed,  
225 *SGPP2* falls in a region recently identified as having an excess of derived alleles in modern humans  
226 compared with the sequenced Neanderthal (**Fig. 3d**)<sup>68</sup>, potentially consistent with recent positive  
227 selection driving evolutionary changes in *SGPP2* transcription. To more directly measure the  
228 impact of positive selection, we used INSIGHT<sup>69</sup> to compare patterns of within-species  
229 polymorphism and between-species sequence divergence in TREs that had undergone human  
230 lineage-specific transcriptional changes. This analysis indicated that dREG sites are most strongly  
231 influenced by weak negative selection (**Fig. 3f**), based on an excess of low-frequency derived alleles  
232 in human populations, as has been reported previously for regulatory sequences<sup>9</sup>. Nevertheless,  
233 TREs with lineage-specific transcriptional changes in human CD4+ T-cells showed reduced weak  
234 negative selection and were strikingly enriched for adaptive nucleotide substitutions ( $p < 0.01$   
235 INSIGHT likelihood ratio test; **Fig. 3f**), consistent with positive selection at these sites. We estimate  
236 a total of at least 121 adaptive substitutions since the human/chimpanzee divergence within TFBSs  
237 that undergo transcriptional changes in human CD4+ T-cells. Despite limited power to detect the  
238 specific contributions of many individual TFs at our stringent motif match score threshold, we did  
239 note significant excesses of putatively adaptive substitutions in the predicted binding sites of  
240 several TFs, including motifs recognized by forkhead box family, POU-domain containing, and ELF/  
241 ETS family (**Supplementary Fig. 11**;  $p < 0.01$ , INSIGHT likelihood ratio test). These estimates  
242 highlight the substantial contribution of adaptive evolutionary changes in TFBSs that may influence  
243 the transcriptional activity of TREs.

244

## 245 **Correlation between protein-coding and non-coding transcription**

246 We noticed that evolutionary changes in protein-coding gene transcription frequently  
247 correlate with changes in non-coding transcription units (TU) located nearby. To examine this  
248 pattern more generally, we adapted our recently reported hidden Markov model (HMM)<sup>70</sup> to  
249 estimate the location of TUs genome-wide, based on patterns of aligned PRO-seq reads and the  
250 location of TREs. Using this method, we annotated 54,793 TUs active in CD4+ T-cells of at least one  
251 of the primate species, approximately half of which overlap existing GENCODE annotations or their  
252 associated upstream antisense RNAs (**Supplementary Fig. 12a**). A cross-species comparison of  
253 the transcription levels for various TU classes (**Fig. 4a**) revealed that non-coding RNAs evolve in  
254 expression most rapidly and protein-coding genes evolve most slowly. GENCODE-annotated

255 lincRNAs undergo evolutionary changes in expression about as frequently as the unannotated non-  
256 coding RNAs predicted by our HMM, which are likely enriched for bi-directionally unstable eRNA  
257 species. The broad similarity in evolutionary conservation between these two non-coding RNA  
258 classes may be consistent with observations that some lincRNAs function as enhancers for nearby  
259 protein-coding genes, and that stable accumulation of the transcript is dispensable for this  
260 biological function<sup>71</sup>.

261 We next measured the extent to which non-coding and protein-coding transcriptional  
262 activities are correlated through evolutionary time. We found that evolutionary changes in protein-  
263 coding gene expression among any of the primate species were highly correlated with those at both  
264 upstream (Pearson's  $R = 0.85$ ,  $p < 2.2e-16$ ) and internal ( $R = 0.66$ ,  $p < 2.2e-16$ ) antisense transcripts  
265 of the same genes. Moreover, changes in the transcriptional activity of gene promoters correlate  
266 with changes in the activity of matched distal enhancers defined by various criteria: namely,  
267 enhancers to which the promoters loop according to cell-type matched ChIA-PET data ( $R = 0.45$ -  
268  $0.61$ ,  $p < 2.2e-16$ ; depending on analysis assumptions)<sup>72</sup>, enhancers located nearby the promoters  
269 ( $R = 0.69$ ,  $p < 2.2e-16$ ), or enhancers that share the same topological associated domain as the  
270 promoter ( $R = 0.62$ ,  $p < 2.2e-16$ )<sup>73</sup>. Using a generalized linear model to integrate expression  
271 changes in multiple types of TUs, we can explain 74% of the variance in gene transcription levels  
272 when we observe differences between species ( $R^2 = 0.74$  in a held-out set of sites,  $p < 2.2e-16$ ; **Fig.**  
273 **4b**) based on the activities of looped TREs, nearby TREs in the same topological associated domain,  
274 internal antisense TUs, and the upstream antisense TU. Thus evolutionary changes that result in  
275 differences in Pol II recruitment to protein-coding genes are well correlated across all interacting  
276 TREs, indicating a shared evolutionary pressure at proximal and distal TREs.

277

### 278 **Rates of Enhancer Evolution Vary with Evidence for Gene Interactions**

279 Despite this overall positive correlation, transcription at enhancers evolves rapidly and is  
280 frequently unaccompanied by transcriptional changes at nearby protein-coding genes. For example,  
281 *CCR7* transcription is highly conserved among both primate and rodent species (**Fig. 5a**;  
282 **Supplementary Fig. 2**) in spite of several apparent changes in enhancer activity within the same  
283 locus (gray vertical bars). These findings are consistent with recent observations that changes in  
284 enhancers within densely populated loci often do not have appreciable effects on the transcription  
285 of genes within the locus<sup>36,37</sup>.

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

296 These simple distance-based observations, however, ignore the critical issue of chromatin  
297 interactions between enhancers and promoters. To account for such loop interactions, we extracted

298 6,520 putative TRE interactions from Chromatin Interaction Analysis with Paired End Tag  
299 sequencing (ChIA-PET) data recognizing loops marked with H3K4me2 in human CD4+ T-cells<sup>72</sup>.  
300 We found that 55% of enhancers that participate in these loops were conserved between primate  
301 species compared to only 47% of non-looped enhancers (Fig. 5c;  $p = 5.6e-4$ , Fisher's exact test).  
302 Moreover, higher transcriptional conservation at looped enhancers does not depend on the  
303 distance to the transcription start site. Parallel analysis of promoter-capture Hi-C data<sup>74</sup> revealed  
304 that the strength of chromatin interaction was correlated with evolutionary conservation of distal  
305 TREs, corroborating the result obtained using ChIA-PET ( $p < 1e-3$ , bootstrap test). We observed  
306 similar levels of conservation at recently defined super-enhancers<sup>75</sup>, although this conservation  
307 may simply reflect an enrichment for loop interactions (48% of TREs in super-enhancers loop  
308 according to ChIA-PET, compared to 15% of all TREs). Looped enhancers were also enriched for  
309 elevated phyloP scores relative to either non-looped enhancers or randomly selected DNA  
310 sequences (Supplementary Fig. 14; phyloP > 0.75;  $p < 2.2e-16$ , Wilcoxon Rank Sum Test). That the  
311 enhancers participating in loop interactions are more highly conserved at both the transcription  
312 and DNA-sequence levels indicates that these enhancers have a disproportionately large effect on  
313 fitness, presumably owing to a more direct role in transcriptional regulation.

314

### 315 **Enhancer-Promoter Interactions Contribute to Constraint on Gene Transcription** 316 **Levels**

317 Distal loop interactions do not fully account for the disparity between enhancers and  
318 promoters in evolutionary rates—even looped enhancers still evolve significantly faster than  
319 promoters ( $p = 3e-3$ , Fisher's exact test). We hypothesized that redundancy in enhancers may help  
320 to explain rapid enhancer evolution. More specifically, we asked whether redundancy makes  
321 protein-coding genes regulated by multiple distal TREs, such as *CCR7* (Fig. 5a), more robust to  
322 enhancer turnover than those influenced by fewer distal TREs. Indeed, we found that evolutionary  
323 conservation of promoter TRE transcription is remarkably strongly correlated with the number of  
324 loop interactions a promoter has with distal sites (Fig. 6a, weighted Pearson's correlation = 0.87;  $p$   
325 <  $1e-3$  by a bootstrap test). A similar trend was observed between the number of loop interactions  
326 made by a target promoter and DNA sequence conservation in transcription factor binding motifs at  
327 the promoter, although the effect was weaker and did not meet our criteria for statistical  
328 significance (Fig. 6b, weighted Pearson's correlation = 0.65;  $p = 0.07$  by a bootstrap test).

329 But how does redundancy in distal TREs relate to the evolutionary conservation of the  
330 distal TREs themselves? If redundant distal TREs compensate for one another in some way,  
331 perhaps each one will be less, rather than more, conserved when their associated promoters have  
332 larger numbers of loop interactions. To address this question, we examined the rate of  
333 conservation of looped distal TREs as a function of the number of loops in which their gene-  
334 proximal partners participated. We found that DNA sequence conservation in putative TFBSs  
335 negatively correlates with the number of loops at the proximal end (Fig. 6d; weighted Pearson's  
336 correlation =  $-0.80$ ;  $p = 2e-3$  by a bootstrap test). We noted a similar trend toward a negative  
337 correlation between the conservation of distal TRE transcription and the number of loop  
338 interactions (weighted Pearson's correlation =  $-0.67$ ;  $p = 0.059$ , two-sided bootstrap test, Fig. 6c).  
339 These results suggest that each associated distal TFBS is individually less essential at genes having  
340 larger numbers of loop interactions with distal sites, and they are therefore consistent with a model



341 in which such TFBSs are more freely gained and lost during evolution. Taken together, our results  
342 imply that distance, looping, and redundancy of enhancers all contribute to constraints on the  
343 evolutionary rates of changes in gene transcription.

344

#### 345 **Discussion:**

346 We have carried out the first comparative analysis of primary transcription in any  
347 phylogenetic group, focusing on CD4+ T-cells in primates. Using PRO-seq and various  
348 computational tools, we estimated the locations and abundance of transcription units with high  
349 resolution and accuracy. In comparison to previous studies in primates<sup>29,33,76-78</sup>, this approach  
350 separated primary transcription from post-transcriptional processing, allowing us to study eRNAs,  
351 lincRNAs, and other rapidly degraded non-coding RNAs, as well as protein-coding genes. We found  
352 clear relationships between the DNA sequences of TFBSs and differential transcription across  
353 species and treatment conditions. We also found evidence that some transcriptional changes in  
354 humans were driven by adaptive evolution in nearby binding sites. Overall, our study provides new  
355 insights into the mode and tempo of recent evolutionary changes in transcription in primates.

356 Perhaps our most striking observation is that many non-coding transcription units,  
357 particularly eRNAs and lincRNAs, have undergone rapid evolutionary changes in comparison to  
358 protein-coding genes. Similar observations have been reported previously for lincRNAs<sup>79</sup>, but, to  
359 our knowledge, the observation for eRNAs is new, and it raises a number of questions. First, why  
360 are some enhancers more conserved than others? In particular, we find that enhancers proximal to,  
361 or that loop to, annotated promoters tend to be most constrained (**Fig 5b-c**). These enhancers may  
362 simply be most crucial for activating their target genes, but other factors may also contribute to  
363 their constraint. For example, perhaps these enhancers are enriched for tissue-specific functions,  
364 and are less constrained due to reduced pleiotropy<sup>80</sup>. Or perhaps many of them simply are not  
365 functional at all, and are transcribed as a by-product of other processes.

366 Second, how do protein-coding genes maintain stable transcription levels across species  
367 despite the rapid turnover of associated enhancers? One possibility is that many rapidly evolving  
368 enhancers are either not functional or act on targets other than the ones we have identified.  
369 However, several of our findings argue against this possibility; for example, we find that even  
370 looped enhancers, for which we have direct evidence of a promoter interaction, evolve significantly  
371 faster than promoters, and that eRNA conservation is strongly correlated with the number of loop  
372 interactions at associated promoters (**Fig. 6**). An alternative explanation, which appears more  
373 plausible to us, is that stabilizing selection on transcription levels drives enhancers to compensate  
374 for one another as they undergo evolutionary flux. This explanation would be compatible with  
375 reports from model systems<sup>36,38</sup>. Our finding that sequence conservation at distal enhancers is  
376 negatively correlated with the number of loop interactions at associated promoters  
377 (**Supplementary Fig. 14**) is also consistent with this explanation. The possibility of pervasive  
378 stabilizing selection on transcription levels in primates has been noted previously based on RNA-  
379 seq data<sup>81</sup>, but our data allow for more direct observations of both active transcription and  
380 associated regulatory elements.

381 A third question is, if most transcribed enhancers do indeed influence gene expression, then  
382 why are so many of them weakly maintained by natural selection? At present, we can only  
383 speculate on the answer to this question. One possibility is that some of the apparent turnover

384 events we have observed actually represent enhancers that have simply switched cell types in  
385 activity, as has been reported in some cases<sup>19</sup>. But it is also possible that selection tends to act  
386 diffusely on enhancers across an entire locus, rather than strongly on individual enhancers, as has  
387 been proposed in cancer evolution<sup>82</sup>. Our observation that multiple DNA sequence changes at the  
388 *SGPP2* locus appear functional provides some initial support this hypothesis. It will be possible  
389 to evaluate these hypotheses more rigorously as better data describing enhancers and enhancer-  
390 promoter interactions across many cell types become available for these and other groups of  
391 species.

392 **Methods:**

393

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

411

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

422

423 *T-cell treatment and PRO-seq library generation.* CD4+ T-cells were allowed to equilibrate in RPMI-  
424 1640 supplemented with 10% FBS for 2-4 hours before starting experiments. Primate CD4+ T-cells  
425 were stimulated with 25ng/mL PMA and 1mM Ionomycin (P/I or  $\pi$ ) or vehicle control (2.5uL EtOH  
426 and 1.66uL DMSO in 10mL of culture media). We selected the minimum concentrations which  
427 saturate the production of IL2 and IFNG mRNA after 3 hours of treatment (data not shown). A 30  
428 min. treatment duration was selected after observing a sharp increase in ChIP-qPCR signal for RNA  
429 Pol II phosphorylated at serine 5 on the C-terminal domain on the IFNG promoter at 30 min. (data  
430 not shown). To isolate nuclei, we resuspended cells in 1 mL lysis buffer (10 mM Tris-Cl, pH  
431 8, 300 mM sucrose, 10 mM NaCl, 2 mM MgAc<sub>2</sub>, 3 mM CaCl<sub>2</sub> and 0.1% NP-40). Nuclei were  
432 washed in 10 mL of wash buffer (10 mM Tris-Cl, pH 8, 300 mM sucrose, 10 mM NaCl and 2 mM  
433 MgAc<sub>2</sub>) to dilute free NTPs. Nuclei were washed in 1 mL, and subsequently resuspended in 50  
434  $\mu$ L, of storage buffer (50 mL Tris-Cl, pH 8.3, 40% glycerol, 5 mM MgCl<sub>2</sub> and 0.1 mM EDTA), snap

435 frozen in liquid nitrogen and kept for up to 6 months before making PRO-seq libraries. PRO-  
436 seq libraries were created exactly as described previously<sup>46</sup>. In most cases, we completed library  
437 preps with one member of each species (usually one human, chimpanzee, and rhesus macaque) to  
438 prevent batch effects from confounding differences between species. Samples were sequenced on  
439 an Illumina Hi-Seq 2000 or NextSeq500 at the Cornell University Biotechnology Resource Center.

440  
441 *Mapping PRO-seq reads.* We mapped PRO-seq reads using standard informatics tools. Our PRO-seq  
442 mapping pipeline begins by removing reads that fail Illumina quality filters and trimming adapters  
443 using cutadapt with a 10% error rate<sup>83</sup>. Reads were mapped with BWA<sup>84</sup> to the appropriate  
444 reference genome (either hg19, panTro4, rheMac3, mm10, or rn6) and a single copy of the Pol I  
445 ribosomal RNA transcription unit (GenBank ID# U13369.1). Mapped reads were converted to  
446 bigWig format for analysis using BedTools<sup>85</sup> and the bedGraphToBigWig program in the Kent  
447 Source software package<sup>86</sup>. The location of the RNA polymerase active site was represented by the  
448 single base, the 3' end of the nascent RNA, which is the position on the 5' end of each sequenced  
449 read. After mapping reads to the reference genome, three samples (one human, U and PI, one  
450 chimpanzee, U and PI, and one rhesus macaque, U and PI) were identified as having poor data  
451 quality on the basis of the number of uniquely mapped reads, and were excluded from downstream  
452 analysis.

453  
454 **Mapping 1:1 orthologs between different species.** During all comparative analyses, the genomic  
455 coordinates of mapped reads, dREG scores, and other parameters of interest were converted to the  
456 human assembly (hg19) using CrossMap<sup>87</sup>. We converted genomic coordinates between genome  
457 assemblies using reciprocal-best (rbest) nets<sup>88</sup>. Reciprocal-best nets have the advantage that  
458 comparisons between species are constrained to 1:1 orthologues. This constraint on mapping is  
459 enforced by requiring each position to map uniquely in a reciprocal alignment between the human  
460 reference and the other species in the comparison. We downloaded rbest nets for hg19-mm10,  
461 hg19-panTro4, hg19-rn6 from the UCSC Genome Browser. We created rbest nets for hg19-rheMac3  
462 using the doRecipBets.pl script provided as part of the Kent Source software package.

463  
464 **Analysis of transcriptional regulatory elements.** *Defining a consensus set of transcriptional*  
465 *regulatory elements.* We predicted TREs using dREG<sup>54</sup> separately in each species' reference  
466 genome. dREG uses a support vector regression model to score each site covered in a PRO-seq  
467 dataset based on its resemblance to features associated with transcription start sites in a reference  
468 training dataset. The dREG model was trained to recognize DNase-I-hypersensitive sites that also  
469 show substantial evidence of GRO-cap data in six PRO-seq or GRO-seq datasets measuring  
470 transcription in resting K562 cells. dREG scores were computed in the reference genome of each  
471 species in order to provide as much information as possible on the native context of each locus. In  
472 all cases, we combined the reads from all individuals for each species in order to maximize power  
473 for the discovery of TREs. In the primate species, treated and untreated CD4+ T-cells were  
474 analyzed separately.

475 We then defined a consensus set of TRE annotations, each of which bore the signature of an  
476 active TRE in at least one species and treatment condition. To define such a set, dREG scores were  
477 first converted to human reference genome (hg19) coordinates using CrossMap and the reciprocal-

478 best nets. The advantage of converting dREG scores between the reference genome is that  
479 individual bases transfer more completely than genomic intervals using CrossMap and related  
480 tools. We then identified TREs in each species separately by thresholding the dREG scores. In all  
481 analyses, we selected a threshold of 0.3, which corresponds to a predicted false discovery rate of  
482 <7% compared with other sources of genomic data in human CD4+ T-cells. In addition, parallel  
483 analyses at separate thresholds (0.25 and 0.35) provided results that were in all cases consistent  
484 with those reported in the main manuscript ([Supplementary Table 2](#)). The set of overlapping  
485 TREs from each species were reduced to a single element containing the union of all positions  
486 covered by the set using bedops, and sites within 500 bp of each other were further merged. We  
487 assigned each putative TRE the maximum dREG score for each species and for each treatment  
488 condition.

489  
490 *Identifying differences in TREs between species.* Differences in TRE transcription in 3-way (human-  
491 chimp-rhesus macaque) or 5-way (human-chimp-rhesus macaque-mouse-rat) species comparisons  
492 were identified using a combination of heuristics and statistical tests. Starting with the consensus  
493 set of TREs in hg19 coordinates, we first excluded potential one-to-many orthologs, by eliminating  
494 TREs that overlapped gaps in the reciprocal-best nets that were not classified as gaps in the  
495 standard nets. The remaining TREs were classified as unmappable when no orthologous position  
496 was defined in the rbest nets. Complete gains and losses were defined as TREs that were mappable  
497 in all species and for which the dREG score was less than 0.05 in at least one species and greater  
498 than 0.30 in at least one other species (see [Supplementary Note 1](#)). Gains and losses were  
499 assigned to a lineage based on an assumption of maximum parsimony under the known species  
500 phylogeny. We defined a set of TREs that displayed high-confidence changes in activity by  
501 comparing differences in PRO-seq read counts between species using deSeq2 and thresholding at a  
502 1% false discovery rate (as described below). Changes in TRE activities were compared to histone  
503 modification ChIP-seq, DNase-I-seq, and DNA methyl immunoprecipitation data from the  
504 Epigenome Roadmap project<sup>58</sup>.

505  
506 *TRE classification.* For some analyses, TREs were classified as likely promoters or enhancers on the  
507 basis of their distance from known protein-coding gene annotations (GENCODE v.19). TRE classes  
508 of primary interest include (see also [Supplementary Fig. 7](#)): (1) Promoters: near an annotated  
509 transcription start site (<100 bp); (2) Enhancers: distal to an annotated transcription start site  
510 (>5,000 bp)

511  
512 *Covariates that correlate with TRE changes.* We compared the frequency at which evolutionary  
513 changes in transcription occur at TREs in a variety of different genomic contexts. We examined the  
514 rate of change as a function of distance from the nearest annotated transcription start site in  
515 GenCode v.19. TREs were binned by distance in increments of 0.02 on a log<sub>10</sub> scale and we  
516 evaluated the mean rate at which evolutionary changes in TRE transcription arise in each bin. We  
517 also compared the rate of changes in TRE transcription across a variety of functional associations,  
518 including loop interactions, within the same topological associated domain, and in super-enhancers.  
519 H3K4me2 ChIA-PET data describing loop interactions were downloaded from the Gene Expression  
520 Omnibus (GEO) database (GSE32677) and the genomic locations of loops were converted from

521 hg18 to hg19 coordinates using the liftOver tool. We also analyzed a separate dataset profiling loop  
522 interactions based on promoter capture Hi-C data in human CD4+ T-cells taken from the  
523 supplementary materials of ref. <sup>74</sup>. Topological associated domains (TADs) based on Hi-C data for  
524 GM12878 cells were also downloaded from GEO (GSE63525). Super-enhancers in CD4+ T-cells  
525 were taken from the supplementary data for ref. <sup>75</sup>. In all cases we excluded sites with potential  
526 one-to-many orthology in any of the species included in the comparison (typically just the three  
527 primates). Potential one-to-many orthologs were defined based on differences in the standard and  
528 reciprocal-best nets for each species pair.

529  
530 *Refining the location of active TREs using dREG-HD.* During analyses of transcription factor binding  
531 motifs we further refined the location of TREs to the region between divergent paused RNA  
532 polymerase using a strategy that we call dREG-HD (manuscript in preparation, preliminary version  
533 available at <https://github.com/Danko-Lab/dREG.HD>). Briefly, we used an epsilon-support vector  
534 regression (SVR) with a Gaussian kernel to map the distribution of PRO-seq reads to smoothed  
535 DNase-I signal intensities. Training was conducted on randomly chosen positions within dREG  
536 peaks extended by 200bp on either side. Selection of feature vectors was optimized based on  
537 Pearson correlation coefficients between the imputed and experimental DNase-I score over the  
538 validation set. PRO-seq data was normalized by sequencing depth and further scaled such that the  
539 maximum value of any prediction dataset is within 90 percentile of the training examples. We  
540 chose a step size to be 60bp and extending 30 steps on each direction. The final model was trained  
541 using matched DNase-I and PRO-seq data in K562 cells.

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

548  
549 **DNA sequence analysis.** *Finding candidate transcription factor binding motifs.* All motif analyses  
550 focused on 1,964 human TF binding motifs from the CisBP database<sup>62</sup> clustered using an affinity  
551 propagation algorithm into 567 maximally distinct DNA binding specificities (see ref <sup>63</sup>). Scores,  
552 which reflect a log<sub>e</sub>-odds ratio comparing each candidate motif model to a third-order Markov  
553 background model, were calculated using the RTFBSDB package<sup>63</sup>. We selected two separate motif  
554 thresholds for different analyses. Scores >10 were used in analyses which mix multiple TF binding  
555 motifs, and strike a tradeoff that focuses on minimizing false positives at the expense of sensitivity.  
556 We dropped the cutoff score to motifs >8 in analyses that use individual motifs in order to increase  
557 statistical power. For each of these thresholds, we estimated the mean genome-wide positive  
558 predictive values to be 0.47 and 0.38, respectively, for motif cutoffs of 10 and 8, by comparing  
559 motifs to ChIP-seq peak calls in K562 cells. During comparative analyses we scanned each primate  
560 reference genome separately with each motif to allow the detection of a putative binding site in any  
561 of the species included in the analysis, and then moved scores to a human (hg19) reference genome  
562 using the CrossMap tool. We chose this strategy because changes in TRE activity may reflect  
563 changes in binding in any of the primate species. For example, human gains may be explained by

564 either a new binding site for a transcriptional activator in the human genome, or a loss in binding of  
565 a transcriptional repressor that was bound in both primate species.

566  
567 *Motif enrichment in TREs that change during CD4+ T-cell activation.* Motifs enriched in up- or down-  
568 regulated dREG-HD TREs during CD4+ T-cell activation ( $p < 0.01$ ) were selected using Fisher's exact  
569 test with a Bonferroni correction for multiple hypothesis testing. Up- or down-regulated TREs  
570 were compared to a background set of >2,500 GC-content matched TREs that do not change  
571 transcription levels following  $\pi$  treatment (log-2 fold change <0.5-fold in magnitude and  $p > 0.25$ )  
572 using the *enrichmentTest* function in RTFBSDB<sup>63</sup>. To test for motif robustness, the background  
573 resampling was repeated 100 times and motifs were selected that achieve a significant result in  
574 >90%.

575  
576 *DNA sequence conservation analysis.* For our evolutionary conservation analysis, we used phyloP  
577 scores<sup>61</sup> based on the 100-way genome alignments available in the UCSC Genome Browser (hg19).  
578 In all cases, bigWig files were obtained from the UCSC Genome Browser and processed using the  
579 bigWig package in R. We represented evolutionary conservation as the mean phyloP score in each  
580 identified TFBS in the indicated set of dREG-HD sites.

581  
582 *Enrichment of DNA sequence changes in motifs.* We identified single-nucleotide DNA sequence  
583 differences at sites at which two of three primate species share one base and the third species  
584 diverges. We intersected these species-specific divergences with matches to transcription factor  
585 binding motifs found within dREG-HD sites that undergo transcriptional changes between primate  
586 species. Because many motifs in Cis-BP are similar to one another, we first partitioned the motifs  
587 using clustering (as described above), and examined enrichments at the level of these clusters.  
588 Motifs were ranked by the Fisher's exact test p-value of the enrichment of species divergences in  
589 dREG-HD sites that change transcription status (where changes in DNA sequence and transcription  
590 occur on the same branch) to dREG-HD sites that do not change. We also compute the enrichment  
591 ratio, which we define as the number of species divergences in each TF binding motif in dREG-HD  
592 sites that change on the same branch normalized to the same statistic in sites that do not change.

593  
594 *INSIGHT analysis.* We examined the modes by which DNA sequences evolve in human lineage-  
595 specific dREG-HD sites or DHSs using INSIGHT<sup>69</sup>. We passed INSIGHT either complete DHSs, dREG-  
596 HD sites, or TFBS within dREG-HD sites that undergo the changes (see *Identifying differences in*  
597 *TREs between species*) indicated in the comparison. Human gains and losses, for example, were  
598 comprised of 4,384 dREG-HD sites with 9,924 separate regions (median length of 16 bp) after  
599 merging overlapping TFBSs with a log-odds score greater than 10. We also analyzed 24  
600 transcription factors each of which has more than 900 occurrences in dREG-HD sites that change on  
601 the human branch (log-odds score >8). All analyses were conducted using the INSIGHT web server  
602 (<http://compgen.cshl.edu/INSIGHT/>) with the default settings enabled.

603  
604 *bQTL analysis.* Frequency shift estimates for all variants in Teranchi et al. (2016) were provided by  
605 the Frasier lab and converted to a queryable database filtered to include only variants with  
606 coverage by 25 reads (75<sup>th</sup> percentile) or more to avoid noise at low read counts. For each

607 sequence/variant query, a set of four equivalent sequences/alternate allele pairs was constructed  
608 by swapping which allele was the reference and getting the reverse complement for both alleles.  
609 For example, given a sequence:variant:position combination of AATCGAA:C:3, the other queries  
610 produced were AACCGAA:T:3 (allele swap), TTCGATT:G:5 (reverse complement), and TTCGGTT:A:5  
611 (reverse complement allele swap). Frequency shifts were computed by taking the post-ChIP  
612 frequency minus the pre-ChIP frequency for the human reference allele. Since k-mers longer than 7  
613 had few hits, we allowed for wildcards (N) in longer sequences that would match any base.  
614 Wildcards were introduced into a k-mer by matching the k-mer sequence to the NF-kB motif and  
615 replacing the 3 lowest information content positions with N(s). Systematic shifts from 0 were  
616 tested using a one-tailed t-test. P-values for systematic differences at multiple sites were combined  
617 using Fisher's method.

618  
619 **De novo discovery of transcription units.** *Identification of transcription units (TU) using a three-*  
620 *state hidden Markov model.* We inferred transcription units (TU) using a three-state hidden Markov  
621 model (HMM) similar to those we have recently published<sup>51,70</sup>. Each TU begins at a TRE identified  
622 using dREG and continues through the entire region inferred to be transcribed, which can covers  
623 tens- to hundreds- of kilobases. Three states were used to represent background (i.e., outside of a  
624 transcription unit), the TU body, and a post-polyA decay region. The HMM transition structure is  
625 shown in [Supplementary Fig. 13a](#). We allow skipping over the post-polyA state, as unstable  
626 transcripts do not have these two-phase profiles. We took advantage of dREG as a potential signal  
627 for transcription initiation by incorporating the dREG score (maximum value in the interval from a  
628 given positive read-count position until the next, clamped to the zero-one interval) as a transition  
629 probability from the background to the transcription body state. PRO-seq data is generally sparse,  
630 so we applied a transformation that encoded only non-zero positions and the distance between  
631 such consecutive positions ([Supplementary Fig. 13a](#)). Our model described this transformed data  
632 using emissions distributions based on two types of variables. The first type of emission variable  
633 defines the PRO-seq read counts in non-zero positions. These counts were modeled using Poisson  
634 distributions in the background and post-polyA states, and using a Negative Binomial distribution  
635 in the transcription body state. The negative binomial distribution can be seen as a mixture of  
636 Poisson distributions with gamma-distributed rates and therefore allows for variation in TU  
637 expression levels across the genome. The second type of emission variable describes the  
638 distribution of distances in base pairs between positions having non-zero read counts. This  
639 distribution was modeled using a separate geometric distribution for each of the three states.  
640 Maximum likelihood estimates of all free parameters were obtained via Expectation Maximization,  
641 on a per-chromosome basis. TU predictions were then obtained using the Viterbi algorithm with  
642 parameters fixed at their maximum-likelihood values. Finally these predictions were mapped from  
643 the transformed coordinates back to genomic coordinates. Source code for our implementation is  
644 publicly available on GitHub: <https://github.com/andrelmartins/tunits.nhp>.

645  
646 *Inferring TU boundaries in the common great ape ancestor.* We identified the most likely TU  
647 locations in the great ape ancestor by maximum parsimony. TUs were identified and compared in  
648 human reference coordinates (hg19) for all species. We used the bedops package to find the  
649 intersection between the predicted TU intervals in each pair of species (i.e., human-chimp, human-



650 rhesus macaque, and chimp-rhesus macaque). Intersections ( $\geq 1$ bp) between pairs of species  
651 were merged, resulting in a collection of TUs shared by any two pairs of species, and therefore  
652 likely to be a TU in the human-chimp ancestor. All steps were applied independently on the plus  
653 and minus strands. These steps identified 37,626 putative TUs active in CD4+ T-cells of the primate  
654 ancestor. We added 17,167 TUs that did not overlap ancestral TUs but were found in any one of the  
655 three primate species.

656  
657 *Transcription unit classification.* TUs were classified by annotation type using a pipeline similar to  
658 ones that we have described recently<sup>51,70,89</sup>. Before classifying TUs we applied a heuristic to refine  
659 TUs on the basis of known annotations. TUs that completely overlap multiple gene annotations  
660 were broken at the transcription start site provided that a dREG site overlapped that transcription  
661 start site. Classification was completed using a set of rules to iteratively refine existing annotations,  
662 as shown in [Supplementary Fig. 13A](#). Unless otherwise stated, overlap between a TU and a  
663 transcript annotation was defined such that  $>50\%$  of a TU matched a gene annotation and covers at  
664 least  $50\%$  of the same annotation. TUs overlapping GENCODE annotations ( $>50\%$  overlap, defined  
665 as above) were classified using the biotype in the GENCODE database into protein coding, lincRNA  
666 (lincRNA or processed transcript), or pseudogene. The remaining transcripts were classified as  
667 annotated RNA genes using GENCODE annotations, the rnaGenes UCSC Genome Browser track  
668 (converted from hg18 to hg19 coordinates), and miRBase v20<sup>90</sup>. As many RNA genes are processed  
669 from much longer TUs, we required no specific degree of overlap for RNA genes. Upstream  
670 antisense (i.e., divergent) TUs were classified as those within 500bp of the transcription start site of  
671 any GENCODE or higher level TU annotation (including lincRNAs). Antisense transcripts were  
672 defined as those with a high degree of overlap ( $>50\%$ ) with annotated protein coding genes in the  
673 opposite orientation. The remaining transcripts with a high degree of overlap ( $>50\%$ ) to annotated  
674 repeats in the repeatmasker database (rmsk) were classified as repeat transcription. Finally, any  
675 TUs still remaining were classified as unannotated, and were further divided into those which are  
676 intergenic or that partially overlapping existing annotations.

677  
678 **Comparing transcription between conditions and species.** *Comparing transcription before and*  
679 *after CD4+ T-cell activation.* We compared  $\pi$  treated and untreated CD4+ T-cells within each of the  
680 primate species using gene annotations (GENCODE v19). We counted reads in the interval between  
681 500 bp downstream of the annotated transcription start site and either the end of the gene or  
682 60,000 bp into the gene body (whichever was shorter). This window was selected to avoid (1)  
683 counting reads in the pause peak near the transcription start site, and (2) to focus on the 5' end of  
684 the gene body affected by changes in transcription during 30 minutes of  $\pi$  treatment assuming a  
685 median elongation rate of 2 kb/ minute<sup>51,91</sup>. We limited analyses to gene annotations longer than  
686 500 bp in length. To quantify transcription at enhancers, we counted reads in the window covered  
687 by each dREG-HD site plus an additional 250 bp on each end. Differential expression analysis was  
688 conducted using deSeq2<sup>57</sup>.

689  
690 *Comparing transcription between species.* Read counts were compared between different species in  
691 hg19 coordinates. In all analyses, reads were transferred to the hg19 reference genome using  
692 CrossMap with rbest nets. Our analysis focused on transcription units or on the union of dREG sites

693 across species. We focused our analysis of transcription units on the interval between 250 bp  
694 downstream of the annotated transcription start site and either the end of the gene or 60,000 bp  
695 into the gene body (whichever was shorter). We limited our analyses to TUs longer than 500 bp in  
696 length. Reads counts were obtained within each transcription unit, gene annotation, or enhancer,  
697 abbreviated here as a 'region of interest' (ROI), that has confident one-to-one orthology in all  
698 species examined in the analysis. This strategy of focusing on blocks of one-to-one orthology avoids  
699 errors caused by systematic differences in mappability or repeat content of species-specific  
700 genomic segments. We broke each ROI into segments that have conserved orthology between  
701 hg19 and all species examined in the analysis, which included either a three-way (human-chimp-  
702 rhesus macaque) or five-way (human-chimp-rhesus macaque-mouse-rat) species comparison. We  
703 defined intervals of one-to-one orthology as those represented in levels 1, 3, and 5 of the reciprocal  
704 best nets (with gaps defined in levels 2, 4, and 6)<sup>88</sup>. Reads that map to regions that have orthology  
705 defined in all species were counted using the bigWig package in R using reads mapped to hg19  
706 coordinates. Final counts for each ROI were defined as the sum of read counts within the regions of  
707 orthology that intersect that ROI. ROIs without confident one-to-one orthologs in all species  
708 analyzed were discarded. Our pipeline makes extensive use of the bigWig R package, Kent source  
709 tools, as well as the bedops and bedtools software packages<sup>85,92</sup>. Differential expression was  
710 conducted between species using the deSeq2 package for R, as described above.

711  
712 **MCF-7 G11 cell culture.** MCF7 G11 tamoxifen resistant cells, were a gift from Dr. Joshua LaBaer.  
713 Cells were maintained in DMEM with 5% FBS, antibiotics, and 1uM tamoxifen. MCF-7 G11 dCas9-  
714 KRAB stable cell lines were made (as described below) and were maintained in DMEM with 5%  
715 FBS, antibiotics, and 1uM tamoxifen. MCF-7 G11 dCas9-KRAB sgRNA stable cell lines were  
716 maintained in DMEM with 5% FBS, antibiotics, 1uM tamoxifen, and 150ug/ul Hygromycin B.

717  
718 **Luciferase assays.** Genomic DNA was isolated from human, chimp, and rhesus macaque PBMCs  
719 depleted for CD4+ cells using a Quick-DNA Miniprep Plus Kit (#D4068S; Zymo research) following  
720 the manufacturer's instructions. Putative enhancer regions were amplified from the genomic DNA,  
721 restriction digested with KpnI and MluI, and cloned into the pGL3-promoter vector (Promega). The  
722 same orthologous regions were amplified from all three species with identical primers where  
723 possible or species-specific primers covering orthologous DNA in diverged regions. Vectors were  
724 co-transfected with pRL-SV40 Renilla (Promega) in a 10:1 ratio (500ng pGL3 to 50ng pRL-SV40) in  
725 MCF7 G11 cells cultured in 1uM tamoxifen. Transfected cells were treated with either 25ng/ml  
726 TNFa or water 21 hours after transfection. 24 hours post-transfection, luminescence was measured  
727 in triplicate using the Dual-Luciferase® Reporter Assay System (Promega).

728  
729 **Silencing endogenous TREs using dCAS9-KRAB.** *Cloning single-guide RNAs (sgRNAs).* Single-  
730 guide RNAs (sgRNAs) were designed using the CRISPR design tool (<http://crispr.mit.edu>) and  
731 sequences are shown in **Supplementary Table 3**. Forward and reverse sgRNAs were synthesized  
732 separately by IDT and annealed. T4 Polynucleotide Kinase (NEB) was used to phosphorylate the  
733 forward and reverse sgRNA during the annealing. 10x T4 DNA Ligase Buffer, which contains 1mM  
734 ATP, was incubated for 30 minutes at 37°C and then at 95°C for 5 minutes, decreasing by 5°C every 1  
735 minute until 25°C. Oligos were diluted 1:200 using Molecular grade water. sgRNAs were inserted

736 into the pLenti SpBsmBI sgRNA Hygro plasmid from addgene (#62205) by following the authors  
737 protocol<sup>93</sup>. The plasmid was linearized using BsmBI digestion (NEB) and purified using gel  
738 extraction (QIAquick Gel Extraction Kit). The purified linear plasmid was then dephosphorylated  
739 using Alkaline Phosphatase Calf Intestinal (CIP) (NEB) to ensure the linear plasmid did not ligate  
740 with itself. A second gel extraction was used as before to purify the linearized plasmid. The purified  
741 dephosphorylated linear plasmid and phosphorylated annealed oligos were ligated together using  
742 the Quick Ligation Kit (NEB). The ligated product was transformed into One Shot Stbl3 Chemically  
743 Competent E. coli (ThermoFisher Scientific). 100ul of the transformed bacteria were plated on  
744 Ampicillin (200ug/ml) plates. Single colonies were picked, sequenced, and the plasmid was isolated  
745 using endo free midi-preps from Omega.

746  
747 *Transfection of MCF-7 G11 cell lines.* We used lentivirus to transfect MCF-7 cells. Lentivirus was  
748 made using lipofectamine 3000 from Invitrogen. Phoenix Hek cells (grown in DMEM with 10% FBS  
749 and antibiotics) were seeded in a 6-well plate at 400,000 cells/plate. Cells were grown until ~90%  
750 confluent. 1ug of pHAGE\_EF1a\_dCas9-KRAB plasmid from addgene (#50919) plasmid or the pLenti  
751 SpBsmBI sgRNA Hygro (addgene #62205) containing each sgRNA, 0.5ug of psPAX (addgene  
752 #12260), and 0.25ug pMD2.G (addgene #12259) were mixed.

753 MCF-7 G11 cells were plated at ~200,000 cells/well in a 6-well plate. 24 hours later  
754 3ml/well of virus was mixed with 10ug/ml polybrene and incubated for 5 minutes at room  
755 temperature. This mix was added to the cells and centrifuged for 40 minutes at 800g at 32C (Viju  
756 Vijayan Pillai, personal communication). 12-24 hours later the virus was removed and fresh media  
757 was added. 24-48 hours later the cells were selected with 2ug/ml puromycin for 2 weeks. The MCF-  
758 7 G11 dCas9-KRAB stable cell lines was grown and maintained in puromycin. A second lentiviral  
759 infection was done using the stable MCF-7 G11 dCas9-KRAB cells. The same protocol was used. 24-  
760 48 hours later the cells were selected with 150ug/ml Hygromycin B. New stable cell lines were  
761 grown and maintained in hygromycin B.

762  
763 *TNF $\alpha$  treatment.* Prior to TNF $\alpha$  treatment, cells were grown for 3 days in DMEM with 5% FBS,  
764 antibiotics, tamoxifen and hygromycin. Cells were then left untreated or treated for 40 min with  
765 25ng/ml TNF $\alpha$ . RNA was extracted using TRIzol Reagent (Invitrogen). We reverse transcribed 1ug  
766 of RNA and used this as input for real-time quantitative PCR (RT-PCR) to analyze *SGPP2* expression.  
767 Primers for *SGPP2* were designed targeting a sequence in intron 1, upstream of the intronic  
768 enhancer. Raw *Cp* values were transferred to units of expression using a standard dilution curve  
769 comprised of a mixture of cDNA from each sample within the biological replicate. We included four  
770 serial dilutions, each of which covered a two-fold difference in expression. Each sample was further  
771 normalized for differences in RNA content by primers recognizing the 18S rRNA control. The ratio  
772 between normalized *SGPP2* expression in each sgRNA-transfected MCF-7 cell line and the empty  
773 vector control was log-2 transformed and tested for differences from 0 using a two-sided t-test.

774  
775 **Data availability.** PRO-seq data was deposited into the Gene Expression Omnibus database under  
776 accession number GSE85337.

777

778 **Code availability.** All data analysis scripts and software are publicly available on GitHub:  
779 <https://github.com/Danko-Lab/CD4-Cell-Evolution>.  
780

781 **References:**

- 782 1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3**,  
783 318–356 (1961).
- 784 2. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357  
785 (1969).
- 786 3. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**,  
787 107–116 (1975).
- 788 4. Rockman, M. V. *et al.* Ancient and recent positive selection transformed opioid cis-regulation in  
789 humans. *PLoS Biol.* **3**, e387 (2005).
- 790 5. Prabhakar, S. *et al.* Human-specific gain of function in a developmental enhancer. *Science* **321**,  
791 1346–1350 (2008).
- 792 6. Capra, J. A., Erwin, G. D., McKinsey, G., Rubenstein, J. L. R. & Pollard, K. S. Many human  
793 accelerated regions are developmental enhancers. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**,  
794 20130025 (2013).
- 795 7. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific  
796 traits. *Nature* **471**, 216–219 (2011).
- 797 8. Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic  
798 limb. *Cell* **154**, 185–196 (2013).
- 799 9. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor  
800 binding sites. *Nat. Genet.* **45**, 723–729 (2013).
- 801 10. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and  
802 Chimpanzee Neural Crest. *Cell* **163**, 68–83 (2015).
- 803 11. Siepel, A. & Arbiza, L. Cis-regulatory elements and human evolution. *Curr. Opin. Genet. Dev.* **29**,  
804 81–89 (2014).
- 805 12. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–  
806 216 (2007).
- 807 13. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21.  
808 *Science* **322**, 434–438 (2008).
- 809 14. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer  
810 genomics. *Science* **342**, 1235587 (2013).
- 811 15. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.-D. & Wray, G. A. Promoter regions of many  
812 neural- and nutrition-related genes have experienced positive selection during human  
813 evolution. *Nat. Genet.* **39**, 1140–1144 (2007).
- 814 16. Torgerson, D. G. *et al.* Evolutionary processes acting on candidate cis-regulatory regions in  
815 humans inferred from patterns of polymorphism and divergence. *PLoS Genet.* **5**, e1000592  
816 (2009).
- 817 17. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription  
818 factor binding. *Science* **328**, 1036–1040 (2010).
- 819 18. Ballester, B. *et al.* Multi-species, multi-transcription factor binding highlights conserved control  
820 of tissue-specific biological pathways. *Elife* **3**, e02626 (2014).
- 821 19. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory  
822 evolution. *Science* **346**, 1007–1012 (2014).
- 823 20. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five *Drosophila*  
824 species show functional enhancer conservation and turnover during cis-regulatory evolution.  
825 *Nat. Genet.* **46**, 685–692 (2014).
- 826 21. Doniger, S. W. & Fay, J. C. Frequent gain and loss of functional transcription factor binding sites.  
827 *PLoS Comput. Biol.* **3**, e99 (2007).
- 828 22. Zheng, W., Zhao, H., Mancera, E., Steinmetz, L. M. & Snyder, M. Genetic analysis of variation in  
829 transcription factor binding in yeast. *Nature* **464**, 1187–1191 (2010).

- 830 23. Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in  
831 transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* **8**,  
832 e1000343 (2010).
- 833 24. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
- 834 25. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-  
835 option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).
- 836 26. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II  
837 transcription in vivo. *Nature* **461**, 186–192 (2009).
- 838 27. Carvunis, A.-R. *et al.* Evidence for a common evolutionary rate in metazoan transcriptional  
839 networks. *Elife* **4**, (2015).
- 840 28. Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene expression  
841 variation in primates. *Genome Biol.* **15**, 547 (2014).
- 842 29. Cain, C. E., Blekhman, R., Marioni, J. C. & Gilad, Y. Gene expression differences among primates  
843 are associated with changes in a histone epigenetic modification. *Genetics* **187**, 1225–1234  
844 (2011).
- 845 30. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392  
846 (2012).
- 847 31. Paris, M. *et al.* Extensive divergence of transcription factor binding in *Drosophila* embryos with  
848 highly conserved gene expression. *PLoS Genet.* **9**, e1003748 (2013).
- 849 32. Cusanovich, D. A., Pavlovic, B., Pritchard, J. K. & Gilad, Y. The functional consequences of  
850 variation in transcription factor binding. *PLoS Genet.* **10**, e1004226 (2014).
- 851 33. Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene  
852 expression in mammals. *Genome Res.* **25**, 167–178 (2015).
- 853 34. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active  
854 estrogen receptor binding sites. *Genome Res.* **23**, 1210–1223 (2013).
- 855 35. Domené, S. *et al.* Enhancer turnover and conserved regulatory function in vertebrate evolution.  
856 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20130027 (2013).
- 857 36. Wunderlich, Z. *et al.* Krüppel Expression Levels Are Maintained through Compensatory  
858 Evolution of Shadow Enhancers. *Cell Rep.* **12**, 1740–1747 (2015).
- 859 37. Cannavò, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory  
860 Networks. *Curr. Biol.* **26**, 38–51 (2016).
- 861 38. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a  
862 eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
- 863 39. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene  
864 promoters. *Nature* **489**, 109–113 (2012).
- 865 40. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal  
866 domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
- 867 41. Khan, Z. *et al.* Primate transcript and protein expression levels evolve under compensatory  
868 selection pressures. *Science* **342**, 1100–1104 (2013).
- 869 42. Bauernfeind, A. L. *et al.* Evolutionary Divergence of Gene and Protein Expression in the Brains  
870 of Humans and Chimpanzees. *Genome Biol. Evol.* **7**, 2276–2288 (2015).
- 871 43. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science*  
872 **347**, 664–667 (2015).
- 873 44. Pai, A. A. *et al.* The contribution of RNA decay quantitative trait loci to inter-individual  
874 variation in steady-state gene expression levels. *PLoS Genet.* **8**, e1003000 (2012).
- 875 45. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and  
876 divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
- 877 46. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how  
878 promoters direct initiation and pausing. *Science* **339**, 950–953 (2013).
- 879 47. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at

- 880 nucleotide resolution. *Nature* **469**, 368–373 (2011).
- 881 48. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to  
882 RNA Processing. *Cell* **161**, 526–540 (2015).
- 883 49. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using  
884 precision nuclear run-on (PRO-seq). *Nat. Protoc.* **11**, 1455–1476 (2016).
- 885 50. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228  
886 (2016).
- 887 51. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in  
888 breast cancer cells. *Cell* **145**, 622–634 (2011).
- 889 52. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at  
890 mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
- 891 53. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally  
892 distinct RNA species. *Nat. Commun.* **5**, 5336 (2014).
- 893 54. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq  
894 data. *Nat. Methods* **12**, 433–438 (2015).
- 895 55. Khaitovich, P. *et al.* A Neutral Model of Transcriptome Evolution. *PLoS Biol.* **2**, e132 (2004).
- 896 56. Dukler, N. *et al.* Nascent RNA sequencing reveals a dynamic global transcriptional response at  
897 genes and enhancers to the natural medicinal compound celastrol. *bioRxiv* 117689 (2017).  
898 doi:10.1101/117689
- 899 57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for  
900 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- 901 58. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human  
902 epigenomes. *Nature* **518**, 317–330 (2015).
- 903 59. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of  
904 enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).
- 905 60. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of  
906 enhancers with distinct cellular functions. *Genome Res.* **21**, 1273–1283 (2011).
- 907 61. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution  
908 rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
- 909 62. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence  
910 specificity. *Cell* **158**, 1431–1443 (2014).
- 911 63. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription  
912 factor binding site analysis. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw338
- 913 64. Macian, F. NFAT proteins: key regulators of T-cell development and function. *Nat. Rev.*  
914 *Immunol.* **5**, 472–484 (2005).
- 915 65. Tehranchi, A. K. *et al.* Pooled ChIP-Seq Links Variation in Transcription Factor Binding to  
916 Complex Disease Risk. *Cell* **165**, 730–741 (2016).
- 917 66. Franco, H. L., Nagari, A. & Kraus, W. L. TNF $\alpha$  signaling exposes latent estrogen receptor binding  
918 sites to alter the breast cancer cell transcriptome. *Mol. Cell* **58**, 21–34 (2015).
- 919 67. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing  
920 of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
- 921 68. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- 922 69. Gronau, I., Arbiza, L., Mohammed, J. & Siepel, A. Inference of natural selection from  
923 interspersed genomic elements based on polymorphism and divergence. *Mol. Biol. Evol.* **30**,  
924 1159–1171 (2013).
- 925 70. Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated  
926 and cell type-specific transcription units from global run-on sequencing data. *BMC*  
927 *Bioinformatics* **16**, 222 (2015).
- 928 71. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription  
929 and splicing. *Nature* **539**, 452–455 (2016).

- 930 72. Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide  
931 enhancer-promoter interactions reveals co-expression of interacting genes and modes of  
932 higher order chromatin organization. *Cell Res.* **22**, 490–503 (2012).
- 933 73. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of  
934 chromatin looping. *Cell* **159**, 1665–1680 (2014).
- 935 74. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding  
936 Disease Variants to Target Gene Promoters. *Cell* **167**, 1369–1384.e19 (2016).
- 937 75. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947  
938 (2013).
- 939 76. Barreiro, L. B., Marioni, J. C., Blekhman, R., Stephens, M. & Gilad, Y. Functional comparison of  
940 innate immune signaling pathways in primates. *PLoS Genet.* **6**, e1001249 (2010).
- 941 77. Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P. & White, K. P. Expression profiling in primates  
942 reveals a rapid evolution of human transcription factors. *Nature* **440**, 242–245 (2006).
- 943 78. Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K. & Gilad, Y. Gene regulation in primates  
944 evolves under tissue-specific selection pressures. *PLoS Genet.* **4**, e1000271 (2008).
- 945 79. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression.  
946 *PLoS Genet.* **8**, e1002841 (2012).
- 947 80. Lewis, J. J., van der Burg, K. R. L., Mazo-Vargas, A. & Reed, R. D. ChIP-Seq-Annotated Heliconius  
948 erato Genome Highlights Patterns of cis-Regulatory Evolution in Lepidoptera. *Cell Rep.* **16**,  
949 2855–2863 (2016).
- 950 81. Gilad, Y., Oshlack, A. & Rifkin, S. A. Natural selection on gene expression. *Trends Genet.* **22**, 456–  
951 461 (2006).
- 952 82. Bailey, S. D. *et al.* Noncoding somatic and inherited single-nucleotide variants converge to  
953 promote ESR1 expression in breast cancer. *Nat. Genet.* **48**, 1260–1266 (2016).
- 954 83. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.  
955 *EMBnet.journal* **17**, 10–12 (2011).
- 956 84. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
957 *Bioinformatics* **25**, 1754–1760 (2009).
- 958 85. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.  
959 *Bioinformatics* **26**, 841–842 (2010).
- 960 86. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief.*  
961 *Bioinform.* **14**, 144–161 (2013).
- 962 87. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome  
963 assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
- 964 88. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution’s cauldron: duplication,  
965 deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.*  
966 **100**, 11484–11489 (2003).
- 967 89. Luo, X., Chae, M., Krishnakumar, R., Danko, C. G. & Kraus, W. L. Dynamic reorganization of the  
968 AC16 cardiomyocyte transcriptome in response to TNF $\alpha$  signaling revealed by integrated  
969 genomic analyses. *BMC Genomics* **15**, 155 (2014).
- 970 90. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep  
971 sequencing data. *Nucleic Acids Res.* **42**, D68–73 (2014).
- 972 91. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing,  
973 and elongation rate in cells. *Mol. Cell* **50**, 212–222 (2013).
- 974 92. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**,  
975 1919–1920 (2012).
- 976 93. Pham, H., Kearns, N. A. & Maehr, R. Transcriptional Regulation with CRISPR/Cas9 Effectors in  
977 Mammalian Cells. *Methods Mol. Biol.* **1358**, 43–57 (2016).



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

990  
991 **Author contributions:** LAC, BAM, CGD, EJR, and ETW performed CD4+ T-cell extraction, validation,  
992 and PRO-seq experiments. CGD, ZW, TC, ALM, LAC, and ND analyzed the data. CGD, AS, JTL, WLK,  
993 and SAC supervised data collection and analysis. CGD and AS wrote the paper with input from the  
994 other authors.

995  
996 **Competing financial interests:** The authors declare no competing financial interests.

997  
998 **Author information:** PRO-seq data was deposited into the Gene Expression Omnibus database  
999 under accession number GSE85337. All data analysis scripts and software are publicly available on  
1000 GitHub: <https://github.com/Danko-Lab/CD4-Cell-Evolution>.

1 **Figure Legends:**

2

3 **Fig. 1 | Maps of primary transcription in CD4+ T-cells. (a)** CD4+ T-cells were isolated from the

4 blood or spleen of individuals from five vertebrate species, including human, chimpanzee, rhesus

5 macaque, mouse, and rat. **(b)** Hierarchical clustering of PRO-seq signal intensities in gene bodies

6 groups CD4+ T-cell samples first by treatment condition and second by species. The color scale

7 represents Spearman's rank correlation between normalized transcription levels in active gene

8 bodies. Colored boxes (top) represents the species and treatment condition of each sample. **(c)** MA

9 plot shows the log<sub>2</sub> fold-change following  $\pi$  treatment in human CD4+ T-cells (y-axis) as a function

10 of the mean transcription level in GENCODE annotated genes (x-axis). Red points indicate

11 statistically significant changes ( $p < 0.01$ ). Several classical response genes that undergo well-

12 documented changes in transcript abundance following CD4+ T-cell activation (e.g., *IL2*, *IFNG*,

13 *TNF $\alpha$* , and *EGR3*) are marked. **(d)** Enrichment of TF binding motifs in TREs that increase

14 transcription levels following  $\pi$  treatment in the indicated species compared to TREs whose

15 transcription abundance does not change. Table shows the Bonferroni corrected p-value, based on

16 a Fisher's exact test (circle size), and the fold-enrichment over a group of unchanged background

17 sequences (color scale). Motif logos and the candidate transcription factor or Cis-BP motif ID are

18 shown. **(e)** Heatmaps show the distribution of PRO-seq (red and blue indicate transcription on the

19 plus and minus strand, respectively), ChIP-seq for H3K27ac, H3K4me1, and H3K4me3, and DNase-

20 I-seq signal intensity. Plots are centered on transcriptional regulatory elements (TREs) predicted

21 in untreated human CD4+ T-cells using dREG-HD (see Online Methods). All plots are ordered based

22 on the maximum dREG score in the window.

23

24 **Fig. 2 | Frequency of changes in TRE transcription. (a)** The fractions of TREs active in untreated

25 CD4+ T-cells that are present in the human reference genome and are conserved across all species

26 (blue), are not detectable and are therefore inferred as gains or losses (teal-white) or undergo

27 significant changes (green) in at least one species, or fall in regions for which no ortholog occurs in

28 at least one of the indicated genomes (pink). Inferred gains or losses are colored according to the

29 FDR corrected p-value associated with changes in RNA polymerase abundance (deSeq2). Plots

30 labeled "Primate" illustrate frequency of changes in a three-way comparison of human, chimpanzee,

31 and rhesus macaque focusing on the untreated condition, whereas those labeled "Mammal"

32 summarize a five-way comparison also including rat and mouse.  $\pi$  treatment denotes a comparison

33 between human untreated and PMA+Ionomycin treated CD4+ T-cell samples. **(b)** Boxplots show

34 the ChIP-seq signal near dREG sites classified as conserved, gains, losses, or complete losses for the

35 indicated chromatin or DNA modification in units of reads per kilobase. The box represents the

36 25th and 75th percentile. Whiskers represent 1.5 times the interquartile range, and points outside

37 of this range are not shown.

38

39 **Fig. 3 | Evolutionary changes in TRE transcription correlate with DNA sequence**

40 **conservation. (a)** Mean phyloP scores near TFBSs that are conserved (red), gained (blue), or lost

41 (cyan) on the human branch. Motifs (score > 10) are at least 100 bp from the nearest annotated

42 exon. **(b)** UCSC Genome Browser track shows transcription near *SGPP2* and *FARSB* in untreated (U)

43 and PMA+ionomycin ( $\pi$ ) treated CD4+ T-cells isolated from the indicated primate species. PRO-

44 seq tracks show transcription on the plus (red) and minus (blue) strands. Axes for the PRO-seq  
45 data are in units of reads per kilobase per million mapped (RPKM). Transcription units inferred  
46 from the PRO-seq data are shown above the plot. dREG tracks show the distribution of dREG signal.  
47 The Green et. al. (ref<sup>64</sup>) selective sweep scan track (top) represents the enrichment of derived  
48 alleles in modern human where Neanderthal has the ancestral allele. Points below the line  
49 represent a statistically significant number of derived alleles in modern human (line indicates a Z-  
50 score of -2). Net synteny tracks show the position of regions that have one-to-one orthologs in the  
51 chimpanzee and rhesus macaque genomes. **(c)** Luciferase signal driven by the *SGPP2* promoter or  
52 the internal enhancer in MCF-7 cells using DNA from each primate species. Bars show the mean  
53 fold-induction following 3 hours of stimulation with TNF $\alpha$ . Error bars represent the standard error  
54 of the mean. Red \*\* denotes  $p < 1e-3$  by a two-tailed t-test. **(d)** Transcription of *SGPP2* using  
55 primers targeting intron 1 following 0 or 40 min. of TNF $\alpha$  treatment after silencing the indicated  
56 TRE using dCAS9-KRAB. Bars represent the median of three independent biological replicates of  
57 two gRNAs targeting the promoter, three targeting the internal enhancer, and four targeting the  
58 upstream enhancer. Error bars represent the standard error. Red \* denotes  $p < 5e-2$  and \*\*  $p < 5e-3$   
59 by a two-tailed t-test. **(e)** INSIGHT estimates of the fraction of nucleotides under selection ( $\rho$ ), the  
60 expected density of segregating polymorphisms under weak negative selection ( $E[Pw]/kbp$ ), or the  
61 expected density of human nucleotide substitutions driven by positive selection ( $E[Dp]/kbp$ ) in  
62 human populations in the indicated class of sites. Red \* denotes conditions significantly enriched  
63 over random background sequences ( $p < 0.01$ ; two-tailed  $X^2$ -test).

64  
65 **Fig. 4 | Changes in non-coding RNA transcription predict changes in gene transcription. (a)**  
66 The fraction of each indicated class of RNAs that undergo changes in transcription in human CD4+  
67 T-cells (see Online Methods). The relationships among the indicated classes of transcription units  
68 are depicted at top. **(b)** Scatterplot shows the magnitude of changes in transcription predicted for  
69 protein-coding genes using changes in the transcription of nearby non-coding RNAs (y-axis) as a  
70 function of changes observed (x-axis). The line has a slope of 1 and an intercept of 0.

71  
72 **Fig. 5 | TRE conservation correlates with loop interactions and distance to gene promoters.**  
73 **(a)** UCSC Genome Browser tracks show transcription, dREG signal, and ChIA-PET loop interactions  
74 near the *CCR7* superenhancer in the human genome. PRO-seq tracks show transcription on the  
75 plus (red) and minus (blue) strands in units of RPKM. Net synteny tracks show regions of one-to-  
76 one orthology with the chimpanzee and rhesus macaque genomes. **(b)** Scatterplot shows the  
77 percentage of TREs conserved among all three primate species (y-axis) as a function of distance,  
78 either upstream or downstream, from the nearest annotated protein-coding transcription start site  
79 (x-axis). The size of each point represents the amount of data in the corresponding distance bin.  
80 **(c)** The percentage of all dREG sites that are conserved in each indicated class of TRE. TREs are  
81 separated into three bins based on the distance relative to the nearest transcription start site.  
82 Error bars reflect a 1,000-sample bootstrap.

83  
84 **Fig. 6 | Stabilizing selection on protein coding gene transcription. (a-b)** Scatterplot shows  
85 promoter conservation **(a)** or DNA sequence conservation **(b)** as a function of the number of loop  
86 interactions made by that site to distal sites across the genome (x-axis). **(c-d)** TRE conservation **(c)**

87 or DNA sequence conservation **(d)** as a function of the number of loop interactions made by the  
88 sequence at the distal end of the loop interaction (x-axis). In all panels the size of each point is  
89 proportional to the number of examples in the corresponding bin, following the scale shown in the  
90 center.

91

92

93 **Supplementary Figure Legends:**

94

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

98

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

105

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

112

113 **Supplementary Fig. 4 | Sample correlation plotted as a function of estimated evolutionary  
114 divergence time between species.** Scatterplot shows the evolutionary divergence time (X-axis) as  
115 a function of Spearman's correlation in gene body transcription between each sample collected in  
116 the untreated condition and the mean gene expression in untreated human CD4+ T-cells (Y-axis).  
117 The red line shows the best linear fit and dotted lines indicate the 99% confidence interval. We  
118 assume the following evolutionary divergence estimates for each species pair with respect to  
119 human, 12 MYR for chimp-human [Moorjani et. al. (2016)], 25 MYR for human-rhesus [Rogers  
120 (2013)], and 75 MYR for human-rodent [Chinwalla et. al. (2002)].

121

122 **Supplementary Fig. 5 | Changes in gene transcription following PMA+Ionomycin treatment in  
123 chimpanzee and rhesus macaque CD4+ T-cells. (a-b)** MA plot shows the log<sub>2</sub> fold change  
124 following  $\pi$  treatment (y-axis) as a function of the mean transcription level in GENCODE annotated  
125 genes (x-axis) in data from chimpanzee (left) and rhesus macaque (right) CD4+ T-cells. Red points  
126 indicate statistically significant changes ( $p < 0.01$ ). Several classical response genes that undergo  
127 well-documented changes in transcript abundance following CD4+ T-cell activation (e.g., *IL2*, *IFNG*,  
128 *TNF*, and *EGR3*) are marked. **(c)** UCSC genome browser track shows transcription in the *IFNG* locus  
129 in untreated (U) and PMA+ionomycin ( $\pi$ ) treated CD4+ T-cells isolated from the primate species

130 indicated at left. PRO-seq tracks show transcription on the plus (red) and minus (blue) strands.  
131 dREG tracks show the distribution of dREG signal. The net-synteny tracks show the fraction of the  
132 genomic area that is mappable in the indicated species. The location of transcription units inferred  
133 in the common ancestor of human and chimpanzee, and the location of RefSeq gene annotations,  
134 are shown at the bottom. **(d-f)** Scatterplots show the correlation between changes in gene  
135 expression (log-2 scale) following  $\pi$  treatment in the species indicated on the axes. Color scale  
136 indicates the density of points in the region.

137

138 **Supplementary Fig. 6 | Evolutionary changes in TREs.** **(a)** Venn diagram illustrating raw  
139 changes in TREs among primate species. In all cases, TREs were discovered in untreated CD4+ T-  
140 cells using dREG (threshold > 0.3). **(b)** Q-Q plot showing observed p-values (deSeq2 in human  
141 compared to the other two primate species) among TREs that were not identified by dREG in at  
142 least one species (red), all TREs identified (black), and a set of conserved TREs (gray).

143

144 **Supplementary Fig. 7 | Evolutionary changes in TREs correlate with chromatin and DNA**  
145 **modifications.** ChIP-seq signal for H3K27ac and H3K4me1 near dREG sites classified as gains,  
146 losses, or complete losses of TRE signal (dREG score < 0.05) on the human branch.

147

148 **Supplementary Fig. 8 | PhyloP scores in transcription factor (TF) binding motifs.** **(a)**  
149 Evolutionary conservation centered on matches to a TF binding motif at the indicated cut off score  
150 (left), or adjusted for distance to the nearest annotated transcription start site by subsampling  
151 (right) **(b)** PhyloP scores that fall within the binding motifs recognized by STAT2 (M6494\_1.02),  
152 YY1 (M4490\_1.02), CREB1 (M6180\_1.02), and ELF1 (M6203\_1.02). In all cases motifs fall in dREG-  
153 HD that are gained (blue) or lost (cyan) on the human branch, or are conserved among all primate  
154 species (red). **(c)** The distribution of human derived alleles near dREG sites that are gained (blue)  
155 or lost (cyan) on the human branch, or are conserved among all primate species (red).

156

157 **Supplementary Fig. 9 | Candidate causal DNA sequence differences underlying changes in**  
158 **SGPP2 transcription.** UCSC genome browser track shows transcription near *SGPP2* and *FARSB* in  
159 untreated (U) and PMA+ionomycin ( $\pi$ ) treated human CD4+ T-cells or in human MCF-7 cells. PRO-  
160 seq tracks show transcription on the plus (red) and minus (blue) strands. Axes for the PRO-seq  
161 data are in units of reads per kilobase per million mapped (RPKM) or in raw reads (MCF-7). dREG  
162 tracks show the distribution of dREG signal. Heatmap (top) shows Hi-C signal in GM12878  
163 lymphoblastoid cell lines. Insert (bottom) shows lack of orthology in chimpanzee and rhesus  
164 macaque in an active TRE (human) that binds a number of TFs in ENCODE cell lines (left) and  
165 substitutions in NF-kB binding motifs near *SGPP2*. Two motif occurrences in the proximal  
166 promoter were bound by RELA, a subunit of NF-kB, based on human ChIP-seq data in ENCODE cell  
167 lines (green boxes). Positions where human carries a derived allele are indicated by yellow  
168 highlights. PRO-seq reads matched the human reference allele in all positions (15/ 15 reads match  
169 C and 26/ 26 match the reference T allele in the NF-kB binding site in the promoter; 11/ 11 reads  
170 match the G and 11/ 11 match the T reference allele in the NF-kB binding site in the promoter; and  
171 24/ 24 reads support the TG human reference sequence in the internal enhancer). Scatterplots  
172 show the relative frequencies of the human allele in RELA (NF-kB) ChIP-seq data matching NF-kB

173 binding QTLs that mimic the human and ancestral alleles, while controlling for the flanking  
174 sequence indicated below the plot. The red dot denotes the mean. All four human-specific DNA  
175 sequence changes in NF-kB binding motifs in the proximal promoter together show trend toward  
176 higher NF-kB binding in human ( $p = 0.017$ , using Fisher's method to combine p-values).

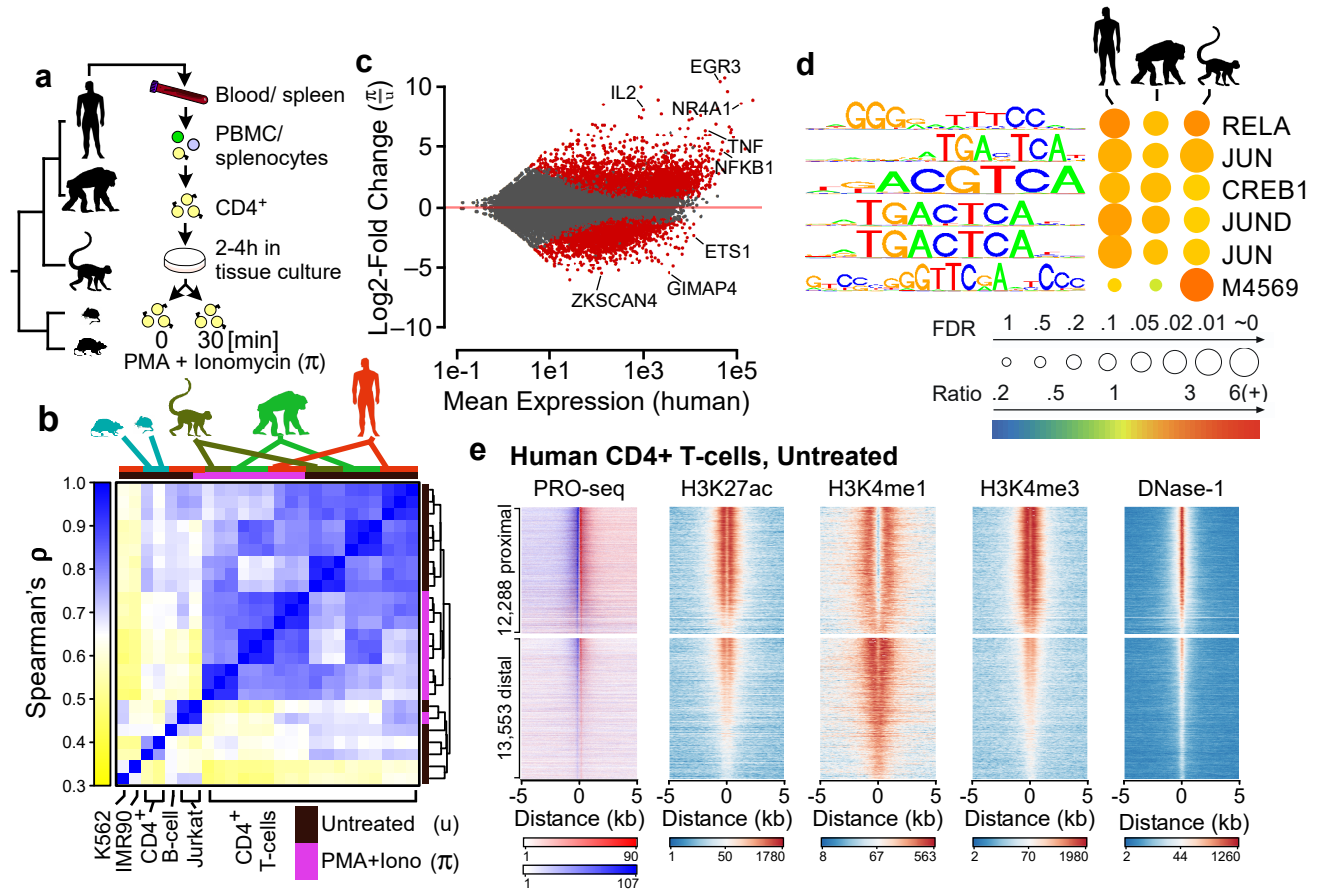
177  
178 **Supplementary Fig. 10 | Luciferase assays for TREs identified near *SGPP2*.** The Y-axis shows  
179 the luciferase signal driven by the *SGPP2* promoter or the internal enhancer in MCF-7 cells using  
180 DNA from each primate species following 3 hours of stimulation with TNF $\alpha$  or vehicle control. Bars  
181 show the mean luciferase activity in each species, over the empty vector and renilla controls. Error  
182 bars represent the standard error of the mean.

183  
184 **Supplementary Fig. 11 | Adaptive substitutions in specific TF binding motifs.** Adaptive  
185 substitutions in TF binding motifs (TFBM) occurring commonly (>900 times) in human lineage-  
186 specific dREG-HD sites. Columns denote the TF name annotated in CisBP (TF), number of sites  
187 (Sites), the number of bases (Bases), the expected number of adaptive substitutes per kilobase  
188 ( $E[A]$ ), the standard error in the expected substitutions per kilobase ( $E[A]_{stderr}$ ), and the  
189 estimated number of adaptive substitutions (# Adaptive Substitutions). TFBSs may be bound by  
190 any TF that recognizes a similar motif. TFBM in which  $E[A]$  is significantly larger than 0 are  
191 highlighted in bold fold. The estimated number of adaptive substitutions for each of these sites is  
192 shown.

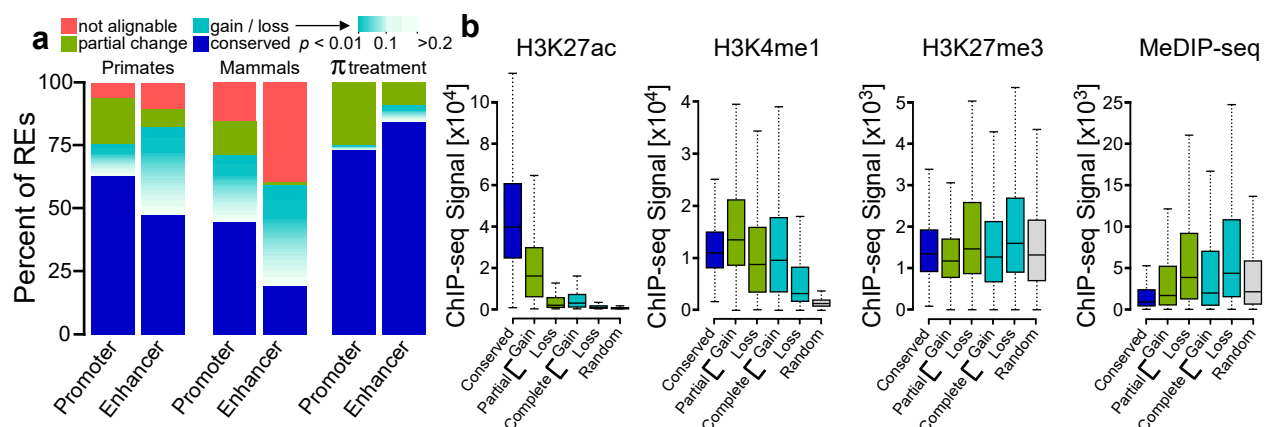
193  
194 **Supplementary Fig. 12 | Discovery of transcription units (TU) in primate T-cells. (a)** A novel  
195 three-state hidden Markov model (HMM) was used to discover transcription units. States  
196 correspond to non-transcribed background sequence, transcribed sequence, and post polyA  
197 transcription. TUs were classified into one of seven classes as indicated in the cartoon. **(b)** The  
198 number and fraction of transcription units that fall into each TU classification. **(c)** Example of the  
199 hidden Markov model (HMM) in a typical region. TUs largely agree with RefSeq gene annotations  
200 when available.

201  
202 **Supplementary Fig. 13 | DNA sequence conservation as a function of genomic distance to the**  
203 **nearest start site.** Scatterplot shows the percentage of TREs undergoing complete gains and  
204 losses (left), undergoing a partial change in the abundance of Pol II (center), or that are not  
205 alignable between species (right) as a function of distance from the nearest annotated transcription  
206 start site (x-axis). The size of each point represents the amount of data in the corresponding  
207 distance bin.

208  
209 **Supplementary Fig. 14 | Evolutionary conservation of DNA sequence mirrors functional**  
210 **conservation at looped- and un-looped enhancers.** Cumulative distribution function of phyloP  
211 scores from the 100-way alignments in the indicated class of dREG site. The insert shows the  
212 fraction of sites in each class exceeding a phyloP score cutoff of 0.75.

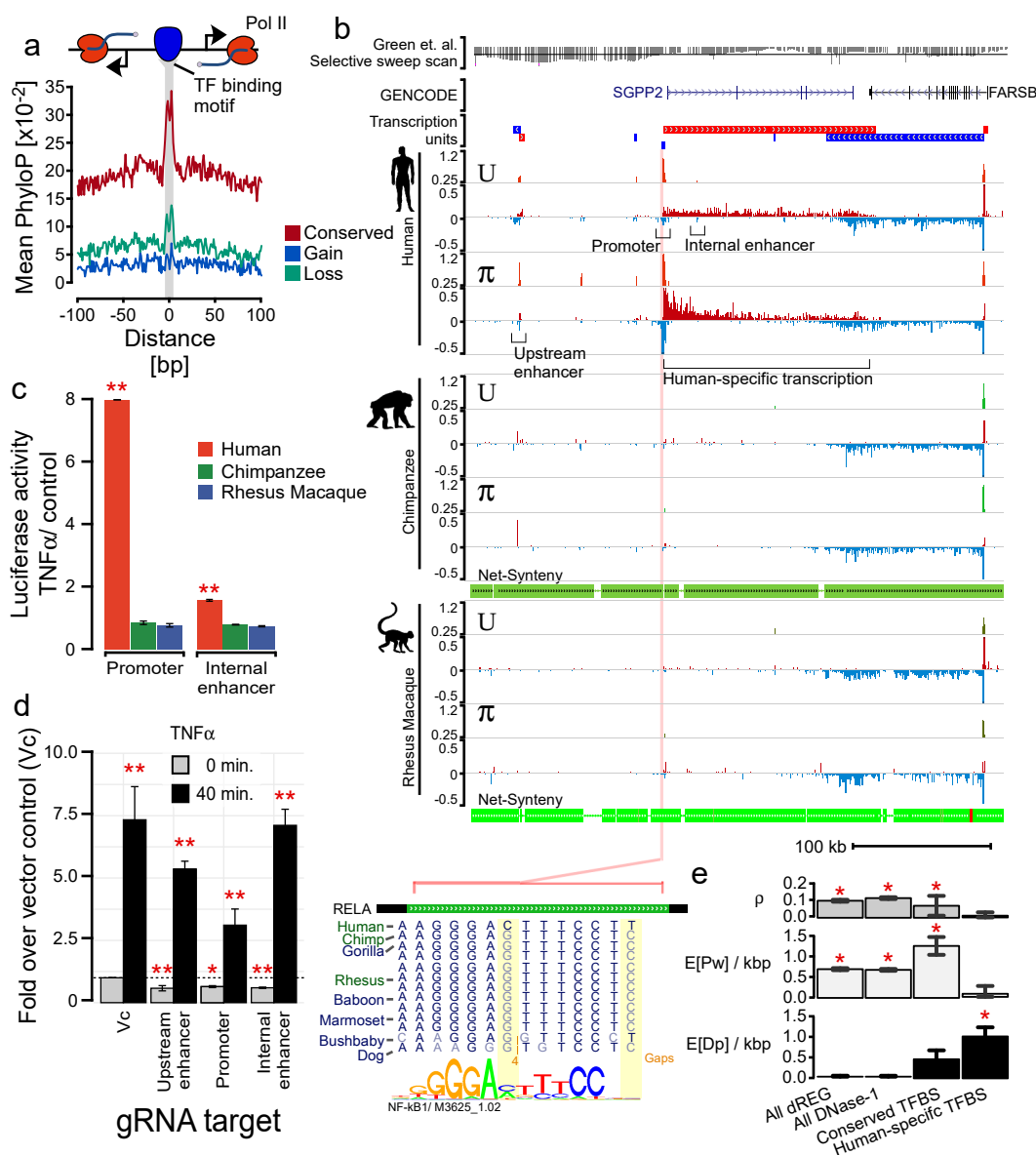


**Fig. 1 | Maps of primary transcription in CD4+ T-cells.** (a) CD4+ T-cells were isolated from the blood or spleen of individuals from five vertebrate species, including human, chimpanzee, rhesus macaque, mouse, and rat. (b) Hierarchical clustering of PRO-seq signal intensities in gene bodies groups 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 log2 fold-change following  $\pi$  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, TNF $\alpha$ , and EGR3) are marked. (d) Enrichment of TF binding motifs in TREs that increase transcription levels following  $\pi$  treatment in the indicated species 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 logos and the candidate transcription factor or Cis-BP 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.

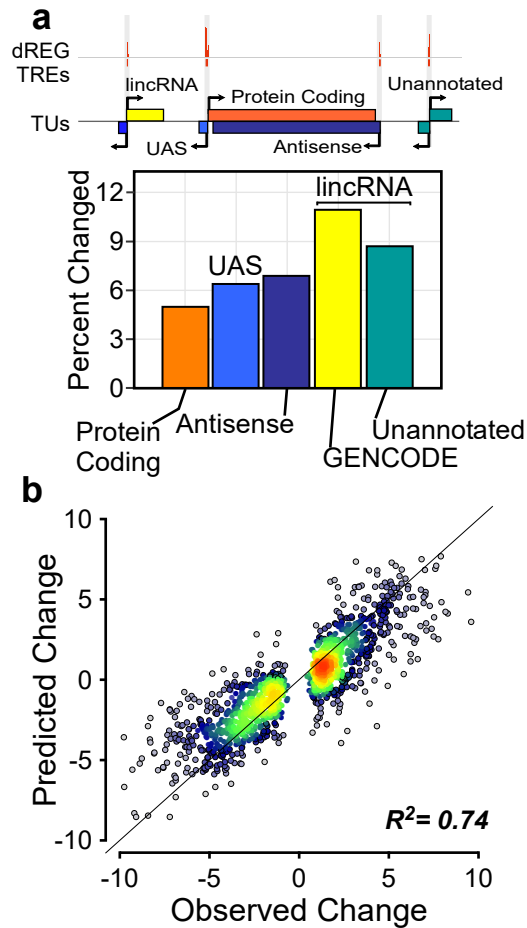


**Fig. 2 | Frequency of changes in TRE transcription. (a)** The fractions of TREs active in untreated CD4<sup>+</sup> T-cells that are present in the human reference genome and are conserved across all species (blue), are not detectable and are therefore inferred as gains or losses (teal-white) or undergo 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.  $\pi$  treatment denotes a comparison between human untreated and PMA+Ionomycin treated CD4<sup>+</sup> T-cell samples. **(b)** Boxplots show the ChIP-seq signal near dREG sites classified as conserved, gains, losses, or complete losses 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.

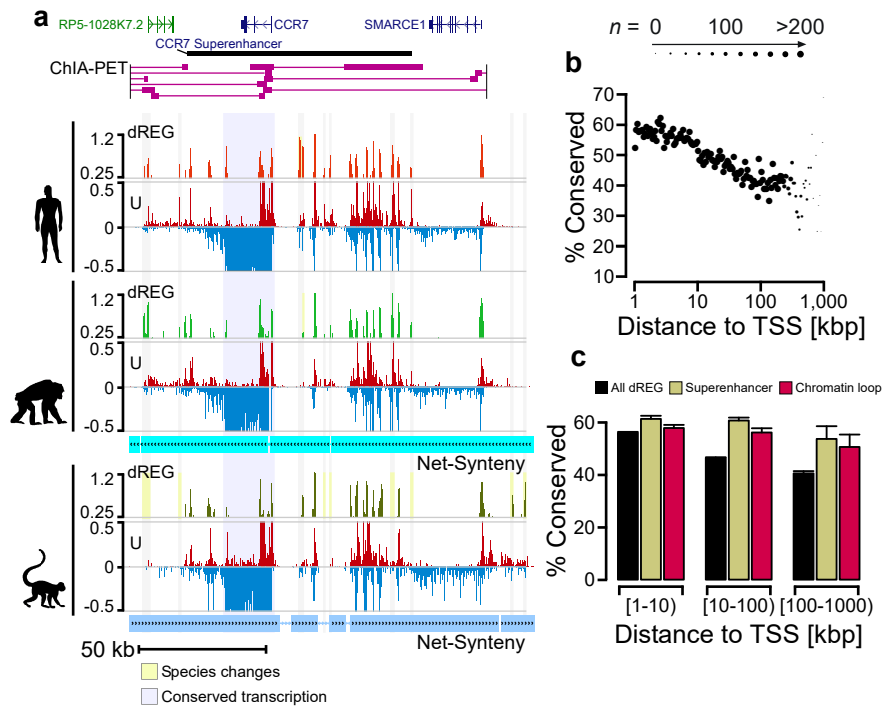




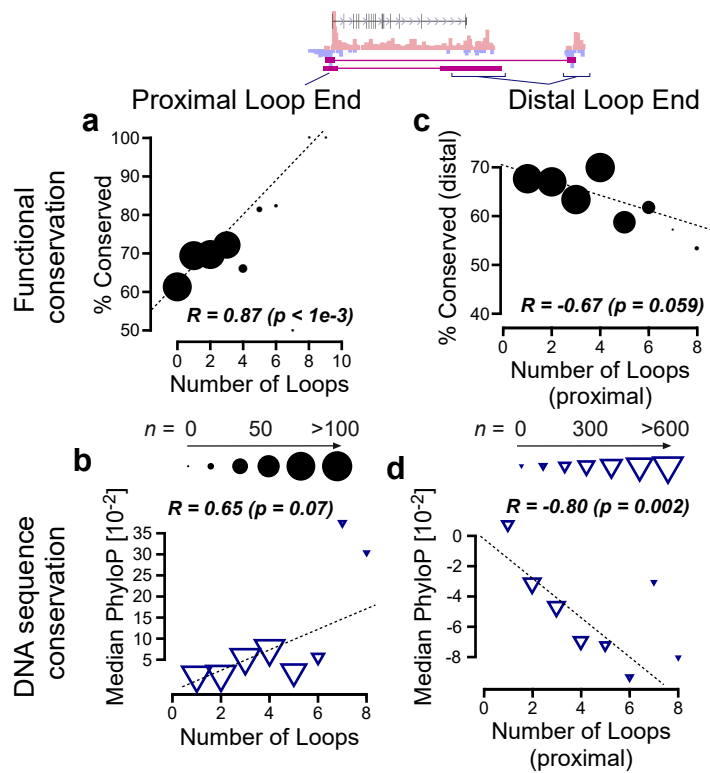
**Fig. 3 | Evolutionary changes in TRE transcription correlate with DNA sequence conservation.** (a) Mean phyloP scores near TFBSs that are conserved (red), gained (blue), or lost (cyan) on the human branch. Motifs (score > 10) are at least 100 bp from the nearest annotated exon. (b) UCSC Genome Browser track shows transcription near SGPP2 and FARSB in untreated (U) and PMA-ionomycin ( $\pi$ ) treated CD4<sup>+</sup> 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). Transcription units inferred from the PRO-seq data are shown above the plot. dREG tracks show the distribution of dREG signal. The Green et. al. (ref64) 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 TNF $\alpha$ . Error bars represent the standard error of the mean. Red \*\* denotes p < 1e-3 by a two-tailed t-test. (d) Transcription of SGPP2 using primers targeting intron 1 following 0 or 40 min. of TNF $\alpha$  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-2 and \*\* p < 5e-3 by a two-tailed t-test. (e) INSIGHT estimates of the fraction of nucleotides under selection ( $\rho$ ), the expected density of segregating polymorphisms under weak negative selection (E[Pw]/kbp), or the expected density of human nucleotide substitutions driven by positive selection (E[Dp]/kbp) in human populations in the indicated class of sites. Red \* denotes conditions significantly enriched over random background sequences (p < 0.01; two-tailed Chi-squared-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.