Prevalent uses of exons as regulatory sequences are possible and inevitable

Jing Chen, Pengyu Ni, Meng Niu[#], Jun-tao Guo* and Zhengsheng Su*

Department of Bioinformatics and Genomics, the University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223

[#]Current address: Department of Genetics Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, 68198

* To whom correspondence should be addressed zcsu@uncc.edu or jguo4@uncc.edu

ABSTRACT

It has long been known that exons can encode transcriptional enhancers. However, the prevalence of such dual-use exons and related questions remain elusive. Our recently predicted highly accurate, large sets of *cis*-regulatory module candidates (CRMCs) and non-CRMCs in the human genome provide us an opportunity to address these questions. We find that exonic transcription factor binding sites(eTFBSs) occupy at least a third of the total exon lengths, suggesting exonic enhancers(eEHs) are more prevalent than originally thought. Moreover, active eTFBSs significantly overlap experimentally determined active eEHs, and enhance the transcription of nearby genes. Furthermore, both A/T and C/G in eTFBSs are more likely under evolutionary selection than those in non-CRMC exons, indicating the eTFBSs might be in dual-use. Interestingly, eTFBSs in codons tend to encode loops rather than more critical helices and strands in protein structures, while eTFBSs in untranslated regions (UTRs) tend to avoid positions where known UTR-related functions were located. Intriguingly, active eTFBSs are found to be in close physical proximity to distal promoters and involved in the activation of target genes. The close physical proximity between exons and promoters in topologically associating domains might render less critical exons to opt for parts of enhancers when non-exonic sequences are unavailable due to space constraints. It appears that nature avoids the dilemma of evolving a sequence for two unrelated functions by using less-critical, physically available exons for eTFBSs. Therefore, the prevalent dual-use of exons is not only possible but also inevitable.

INTRODUCTION

Eukaryotic genomes contain two major types of functional sequences: coding-sequences (CDSs) that encode proteins or RNAs, and cis-regulatory sequences or modules (CRMs) such as promoters and enhancers that control the transcription of their target CDSs(1-3). CRMs have long been known to reside in intergenic regions or in introns of genes(4-7). However, it has been found that exons including CDSs, 5'-untranslated regions(5'-UTRs) and 3'-UTRs could also function as CRMs such as enhancers and silencers for their own or different genes(8-24). For example, the second exon of the ADAMTS5 gene regulates its own transcription(10), and exon 15 of mouse gene Dync1i1 physically interacts with and activates the promoter of the *Dlx5/6* gene that is 900kb away(12). Mutations in such exonic enhancers (eEHs) can result in diseases by altering their enhancer activities(14,25-28). More recently, using a high throughput

enhancer characterization method, self-transcribing active regulatory regions sequencing (STARR-seq)(29), and its variants such as whole genome STARR-seq (WHG-STARR-seq) (30) and ATAC-STARR-seq(31), several groups found that 3~10% of tens of thousands of short active enhancers (~500bp) identified in cell lines were located in exons (CDSs or UTRs)(29,30). However, these exonic sequences were generally ignored, although they were likely parts of long eEHs that were at least partially functional when truncated in the episomal expression vectors(29-31). Moreover, it has been reported that at least 15% of codons in 86% of genes in the human genome were hypersensitive to DNase I treatment in 81 human cell/tissue types and thus were likely in dual-use for both encoding specific amino acid sequences for protein functions and encoding specific transcription factors (TFs) binding sites (TFBSs) for gene transcriptional regulation(32). These so-called duons were found to be more conserved than non-duon codons at the four-fold degenerate positions, and were thought to be under double selection constraints owing to the dual-use(32). Mutations in duons could lead to diseases due to altered activities of relevant eEHs(33).

A reanalysis of the DNase hypersensitive sites (DHSs), however, showed that duons actually evolved similarly to non-duon codons when the conservation levels or substitution rates of A/T and C/G were compared separately, using either phyloP(34) conservation scores or substitution rates derived from the most recent common ancestor of humans and gorilla compared with chimpanzee. The authors argued that the earlier conclusions about duons(32) were incorrectly drawn due to the higher C/G frequencies at the fourfold degenerate positions in duons than in non-duon CDSs, whereas C/G at the fourfold degenerate positions tend to be more conserved than A/T. This result casts doubts on the validity of the duon hypothesis(35). A similar conclusion was drawn based on an analysis of three known eEHs(36). On the other hand, it has been shown that DHSs are not a reliable predictor of CRMs, not to mention more subtle TFBSs, because a naked DNA segment may not necessarily be a CRM (37-41). It is highly likely that the predicted duons might have a high false discovery rate (FDRs)(37-41). Therefore, the evolutionary neutrality of the duons observed earlier(35) might be due to inaccuracy of their predictions. To clarify these contradictory results, sufficient experimentally identified eEHs are clearly needed. Unfortunately, such a dataset is still lacking because even ATAC-STARR-seq and WHG-STARR-seq peaks have up to 56%(31) and 87%(30) FDRs, respectively. Thus, a set of accurately predicted CRMs and constituent TFBSs might be an alternative solution to address these interesting and important questions about eEHs.

We recently predicted a set of 1.41M CRM candidate (CRMCs) and a set of 1.96M non-CRMCs in the human genome by integrating more than 6,000 TF ChIP-seq datasets from various cell/tissue types (41). Validation on experimentally determined CRM function-related sequence elements such as VISTA enhancers(42) and non-coding ClinVar variants(43) indicates that our predicted CRMCs are highly accurate, with an estimated FDR of 0.05% and false negative rate (FNR) of 1.27%(41). The FDR and FNR for predicting CRMs decrease further when a more stringent p-value cutoff is used (41). As expected, most (~94%) of the predicted CRMs and constituent TFBSs are located in non-exonic sequences (NESs), while the remaining ~6% overlap exons(41). These predicted TFBSs in exons promoted us to re-examine the duon hypothesis and address related fundamental questions for the dual-use of codons and UTRs.

MATERIALS AND METHODS

Datasets. We downloaded predicted CRMCs and non-CRMCs from the PCRMS database (https://cci-bioinfo.uncc.edu/)(41), and extracted putative CRMs with a p-value < $5x10^{-6}$ to minimize FDR, resulting in 428,628 CRMs containing 38,507,543 potentially overlapping putative TFBSs. The human genome assembly version GRCh38.p13 was used as the reference genome. Annotations (CDSs, 5-UTRs and 3'-UTRs) of genes and verified transcripts were downloaded from Ensembl Release 100 (https://useast.ensembl.org/index.html). For each gene, the longest annotated transcript (agreed by different databases, with support from mRNA data) was selected for CDS assignments. Protein structures were download from the Protein Data Bank (PDB)(44). The chromatin interaction analysis by paired-end tagging (ChIA-PET) dataset of the K562 cell line (45). was downloaded from the GEO database(46) with the access number GSE33664. TF ChIP-seq datasets were downloaded from the Cistrome database(47). WHG-STARR-seq peaks in four cell lines (A549, HepG2, MCH-7 and SH-HY5Y) and RNA-seq data in 19 human cell/tissue types (Table S1) were downloaded from the ENCODE data portal (https://www.encodeproject.org/about/data-access/). ATAC-STARR-seq peaks in the GM12878 cell line were downloaded from the work by Wang *et.al* (31).

Definition of degenerate positions and non-degenerate positions of the codons

We consider genome nucleotide positions at the third codon positions of the 21 synonymous codon sets to be degenerate, including 12 sets of two-fold degenerate positions, i.e.,

Arg(AG[AG]), Asn(AA[CT]), Asp(GA[CT]), Cys(TG[CT])), Gln(CA[AG]), Glu(GA[AG]), His(CA[CT]), Leu(TT[AG]), Lys(AA[AG]), Phe(TT[CT]), Ser(AG[CT]) and Tyr(TA[CT]); one set of three-fold degenerate positions, Ile(AT[ACT]), and eight sets of four-fold degenerate positions, i.e., Ala(GC*), Arg(CG*), Gly(GG*), Leu(CT*), Pro(CC*), Ser(TC*), Val(GT*), Thr(AC*). We consider genome nucleotide positions at the first codon positions of all the sense codons to be non-degenerate, except for Arg([AC]GA), Arg([AC]GG), Leu([CT]TA) and Leu([CT]TG). We also consider genome nucleotide positions at the second codon positions of all sense codons to be non-degenerate.

Assignment of secondary structure types. For proteins with known structures in PDB, we assigned secondary structure types to amino acid sequences using the DSSP program(48) following the widely used convention: H (α-helix), G (3_{10}-helix) and I (π-helix) states as helix type; E (extended strand) and B (residue in isolated β-bridge) states as strand type and all the other states as loop type. For proteins without known structures, we assigned secondary structure types as follows. We first generated a non-redundant protein sequence set (less than 30% sequence identity) for all annotated proteins in the human genome using CD-HIT(49). If a protein in the dataset has a highly homologous protein with known structure in PDB with at least 50% coverage and 80% sequence identity, the secondary structure types of the template protein were copied to the target protein. If no homologous structures were found in PDB, RaptorX, a highly accurate secondary structure prediction program, was applied to predict secondary structure types with default settings(50).

Assignment of conservation scores. The phyloP(34) and GERP(Genomic Evolutionary Rate Profiling)(51) scores of each nucleotide position in the human genome were downloaded from the UCSC Genome Browser database(52).

Prediction of active and non-active TFBSs in cell lines. If a TFBS entirely falls in a 1,000bp region centering on the summit of any TF ChIP-seq binding peak in a cell line, we predict the TFBS to be active in the cell line. Otherwise, we predict the TFBS to be inactive in the cell line. Clearly, the completeness of prediction of active TFBSs in a cell/tissue type depends on the number of available ChIP-seq datasets for active TFs in the cell/tissue type. We made predictions in 19 cell/tissue types (Table S1) with the highest numbers of available TF ChIP-seq datasets among all cell/tissue types that also had RNA-seq data available from the ENCODE data portal (https://www.encodeproject.org/about/data-access/).

Validation of eTFBSs using STARR-seq peaks. We counted the number of predicted active and inactive eTFBSs that overlapped a STARR-seq peak in a cell line.

Validation of eTFBSs using chromatin interactions. We calculated significant interactions between two loci using ChiaSig(53) based on the ChIA-PET reads from the K562 cells(45). We then counted the number of predicted active and inactive eTFBSs overlapping loci in close proximity to the promoter that was not the immediate upstream to host gene of the eTFBSs, to exclude the promoters of the host gene of eTFBSs.

Statistical analysis. Kolmogorov-Smirnov(K-S) test, Mann Whitney U test or χ^2 test were used to evaluate statistically significant levels of hypothesis tests as indicated in the text and figure legends.

RESULTS

eTFBSs are prevalent but unevenly distributed along CDSs and UTRs

To analyze potential eEHs in the human genome and address the related issues, we used the 428,628 CRMs containing 38,507,543 putative TFBSs predicted with a stringent p-value cutoff of $5x10^{-6}$ to minimize FDR. The CRMs and TFBSs have a total length of 982,470,181bp and 397,703,041bp, covering 31.81% and 12.88% of the mappable genome (3,088,269,832bp), respectively. As expected, the vast majority (94.42%) of the CRM positions are located in NESs, while the remaining 5.58% fall in exons. There are a total of 180,223 (34.08%) CRMs that at least partially overlap exons, and we refer them as exonic CRMs (eCRMs). Although a small number (0.16%) of eCRMs entirely fall in exons, most eCRMs have less than 10% of their lengths overlapping exons (Figure 1a). Figure S1 shows a few examples of the predicted eCRMs that overlap experimentally verified eEHs. An eCRM often overlaps one or few exons and spans adjacent introns and/or intergenic regions (Figure S1a~S1h), but it can also span the entire locus of a gene (Figure S1i~S1k). The vast majority (93.71%) of the TFBS positions are located in NESs, while the remaining 6.29% are located in exons (Figure 1b). A TFBS is considered as an nTFBS if it has at least one position located in a NES. Of the 38,507,543 putative TFBSs, 36,597,744 (95.04%) are nTFBSs, while the other 1,909,799 (4.96%) TFBSs are entirely located within an exon, and each is considered as an eTFBSs. We focus our

subsequent analyses mainly on eTFBSs in the eCRMs instead of the full-length eCRMs, since it is the constituent TFBSs in CRM that mainly determine the function of the CRM(54,55). If an eTFBS position is annotated to be in both a CDS and a UTR, we consider it to be in the CDS. If an eTFBS position is annotated to be in both a 5'-UTR and a 3'-UTR, we consider it as "others" and exclude it from analyses. Of all the TFBS positions (397,703,041bp), 2.80%, 0.90%, 2.58% and 0.01% are located in CDSs, 5'-UTRs, 3'-UTRs and "others" respectively (Figure 1b). If an eTFBS is entirely located in a CDS, a 5'-UTR or a 3'-UTR, we refer it as a cTFBS, a 5'-uTFBS or a 3'-uTFBS, respectively. Of the 38,507,543 putative TFBSs, 1,047,183 (2.72%) 129,518(0.34%) and 733,089(1.90%) are cTFBSs, 5'-uTFBSs and 3'-uTFBSs, respectively. The eTFBSs are located in exons of 19,060 (96.78%) of the 19,694 verified genes, and each gene harbored an average of 56 potentially overlapping eTFBSs (Figure 1c). More specifically, of the eTFBS positions, 54.71%, 7.46% and 37.82% are located in CDSs, 5'-UTRs and 3'-UTRs of 18,971 (96.32%), 12,903(65.52$) and 13,237(67.21%) genes, respectively (Figure 1d). Thus, the vast majority of genes have at least a coding exon, its 5'-UTR or its 3'-UTR at least partially overlapping an eTFBS. The cTFBSs, 5'-eTFBS and 3'-uTFBSs comprise 33.9%, 48.8% and 32.5% of the total length of annotated CDSs (32,797,669bp), 5'-UTRs (3,104,596bp) and 3'-UTRs (23,630,714bp), respectively (Figure 1e). Therefore, a third of CDS and 3'-UTR lengths are eTFBS positions, while almost a half of 5'-UTR lengths are eTFBS positions, suggesting that eTFBSs and eEHs/eCRMs might be more prevalent than originally thought (32,56). Interestingly, cTFBS and 3'-uTFBS positions tend to be clustered at the two ends of CDSs and 3'-UTRs, respectively, while 5'-uTFBS positions are evenly located along most part of 5'-UTRs but tend to avoid the two ends (Figure 1f).

Genes closest to active eTFBSs have higher transcription levels than those closest to inactive eTFBSs

To see whether or not eTFBSs are involved in transcriptional regulation, we predicted active and inactive eTFBSs in 19 human cell/tissue types with available RNA-seq data (Materials and Methods), and compared transcription levels in these cell/tissue types of genes closest to the predicted active eTFBSs with those of genes closest to the predicted inactive eTFBSs. To avoid confounding factors, we exclude the eTFBSs' host genes as the closest genes, although an eTFBS might regulate its host gene as demonstrated in some cases (8,10,15,16) and in the STARR-seq experiments(29). We predicted an average of 31.10%, 78.47% and 28.17% of the 1,047,183 cTFBSs, 129,518 5'-uTFBSs and 733,089 3'-uTFBSs to be active, and the remaining 68.90%, 21.53% and 71.83% to be inactive in the cell/tissue types, respectively. As shown in

Figure 2a, genes closest to the predicted active cTFBSs, 5'-uTFBSs or 3'-uTFBSs have significantly higher transcription levels than those closest to the predicted inactive counterpart eTFBSs in the cell/tissue types ($p<2.2x10^{-302}$ for all comparisons with Mann Whitney U test), indicating that eCRMs containing active eTFBSs might have enhancer activities.

Exonic STARR-seq peaks are significantly enriched in active eTFBSs

Exonic STARR-seq peaks might be the largest sets of eEH/eCRM candidates in the literature so far with demonstrated enhancer activities in episomal contexts. It is interesting to see if exonic STARR-seq peaks were enriched in our predicted active eTFBSs. To this end, we mapped our predicted active and inactive eTFBSs to ATAC-STARR-seq peaks in GM12878 cells and WHG-STARR-seq peaks in other four cell lines (A549, HepG2, MCH-7 and SH-HY5Y) (Materials and Methods). In the five cell lines, we predicted an average of 21.77%, 64.14% and 14.03% of the 1,047,183 cTFBSs, 129,518 5'-uTFBSs and 733,089 3'-uTFBSs to be active and the remaining 78.23%, 35.86% and 85.97% to be inactive, respectively. As shown in Figure 2b, active cTFBSs (3.33%), 5'-uTFBSs(7.62%) and 3'-uTFBSs(3.05%) have significantly higher proportion overlapping STARR-seq peaks than inactive cTFBSs (0.84%), 5'-uTFBSs (1.16%) and 3'-uTFBSs (0.60%), respectively ($p<2.18x10^{-6}$ for all comparison with χ^2 test). Notably, though the differences are highly significant, the enrichment levels are not very high, and this might be due to the facts that the STARR-seq methods can only work on sequences shorter than 500bp(30,31), while the vast majority of known mammal enhancers are longer than 1,000bp(41,42). It is highly likely that many active enhancers longer than 500bp were missed by the STARR-seq methods since the truncated forms of long CRMs might not be functional, even if they were included in STARR-seq episomal expression vectors, resulting in high FNRs. On the other hand, due to the episomal nature of the expression vectors used in STARR-seq methods, 56% and 87% FDRs have been reported for TATAC-STARR-seq (30) and WHG-STARR-seq (31,57), respectively.

To tease out confounding factors of using STARR-seq peaks to validate our predicted eTFBSs, we examined the expression levels in the five cell lines of genes closest to 1) active eTFBSs that overlap STARR-seq peaks (aeTFBSs+S^+), 2) inactive eTFBSs that overlap STARR-seq peaks (ieTFBSs+S^+), 3) active eTFBSs that do not overlap STARR-seq peaks (aeTFBSs+S^-), and 4) inactive eTFBSs that do not overlap STARR-seq peaks (ieTFBSs+S^-). As shown in Figure 2c, genes closest to aeTFBSs+S^+ (acTFBSs+S^+, a5'-uTFBSs+S^+ and a3'-uTFBSs+S^+) had significantly higher transcription levels than those of closest to corresponding

ieTFBSs+S^+(icTFBSs+S^+, i5'-uTFBSs+S^+ and i3'-uTFBSs+S^+, respectively, $p<2.23x10^{-302}$ for all comparisons with Mann Whitney U test), suggesting that STARR-seq peaks that overlap our predicted active eTFBSs might drive gene expression in native chromatin contexts, thus might be eEHs or components of long eEHs that are at least partially functional when truncated in the STARR-seq expression vectors. Interestingly, genes closest to aTFBS+ S^- (acTFBS+ S^-, a5'-uTFBSs+S^- and a3'-uTFBSs+S^-) also had significantly higher transcription levels than those of closest to corresponding icTFBS+ S^- (icTFBSs+S^-, i5'-uTFBSs+S^- and i3'-uTFBSs+S^-, respectively, $p<2.23x10^{-302}$ for all comparisons with Mann Whitney U test)(Figure 2c), suggesting that our predicted active eTFBSs that do not overlap STARR-seq peak might be involved in enhancer activities as well, and eCRMs containing these eTFBSs might be missed by STARR-seq methods due to their aforementioned limitations. On the other hand, genes closest to icTFBSs+S^+, i5'-uTFBSs+S^+ and i3'-uTFBSs+S^+ had similarly low expression levels to those closest to icTFBSs+S^-, 5'-iuTFBSs+S^- and 3'-iuTFBSs+S^-, respectively, Figure 2c), suggesting that STARR-seq peaks that overlap our predicted inactive eTFBSs might not drive gene expression in native chromatin contexts, either because they might be inactive in the cell lines or they might not be enhancers at all as suggested earlier(29-31,57). Taken together, these results unequivocally demonstrate that our predicted active eTFBSs might be involved in enhancer activities, and thus, are likely constituents of eEHs/eCRMs as demonstrated in a few experimentally validated examples (Figure S1).

C/G contents elevate at degenerate positions in cTFBSs as well as at 5'-uTFBS and nTFBS positions

To evaluate possible biased usages of A/T and C/G in our predicted TFBSs and non-CRMCs, we first compared nucleotide frequencies at the degenerate third codon positions of the 21 synonymous codon sets in cTFBSs and non-CRMC CDSs (Materials and Methods). As shown in Figure 3a, although the human genome is A/T-rich (59% A/T vs 41% C/G)(58), there are more C/G than A/T at the degenerate positions of 17 of the 21 synonymous codon sets in cTFBSs, except for Asn (AA[CT]), Ile (AT[ACT]), Ser (TC*) and Thr (AC*). In contrast, there are more A/T than C/G at the degenerate positions of 17 of the 21 synonymous codon sets in non-CRMC CDSs, except for Gln (CA[AG]), Leu (CT*), Leu (TT[AG]) and Val (GT*) (Figure 3a). We next compared nucleotide frequencies at the degenerate third codon positions of the 21 synonymous codon sets in cTFBSs relative to those in non-CRMC CDSs. As shown in Figure 3b, the log odd ratios of the frequencies indicate that C/G are more preferred at the degenerate positions in cTFBSs than in non-CRMC CDSs ($p< 2.2x10^{-16}$, χ^2 test) for all the synonymous

codon sets. This result is in agreement with the earlier finding that C/G was more preferred at the degenerate positions in the 21 synonymous codon sets in duons than those in non-duon CDSs(35). Finally, we compared A/T and C/G contents at the degenerate positions of the 21 synonymous codon sets in cTFBSs as well as at 5'-uTFBS, 3'-uTFBS and nTFBS positions with their non-CRMC counterpart positions. As expected, C/G contents at degenerate positions in cTFBSs are higher than those in non-CRMC CDSs (Figure 3c, $p< 2.2x10^{-16}$, χ^2 test). Interestingly, 5'-uTFBSs and nTFBSs also have higher C/G contents than non-CRMC 5'-UTRs (Figure 3c, $p< 2.2x10^{-16}$, χ^2 test) and non-CRMC NESs (Figure 3c, $p< 2.2x10^{-16}$, χ^2 test), respectively. However, 3'-uTFBSs have slightly but significantly lower C/G than non-CRMC 3'-UTRs (Figure 3c, $p< 2.2x10^{-16}$, χ^2 test). Furthermore, C/G contents in the degenerate positions, 5'-UTRs, 3'-UTRs are higher (Figure 3c, $p< 2.2x10^{-16}$, χ^2 test) than the neutral expectation (41%), regardless of whether they are in TFBSs or not (Figure 3c). It has been shown that C/G contents at four-fold degenerate positions are consistently elevated above the neutral expectation (59). Thus, this is particularly true for degenerate positions in cTFBSs, with C/G contents of 62% (Figure 3c). Interestingly, nTFBS positions also have higher C/G contents than the neutral expectation (41%), while non-CRMC NESs have the same C/G contents as the neutral expectation (41%) (Figure 3c). The elevation of C/G contents in cTFBSs and 5'-uTFBSs might be due to their prolonged nucleosome-free states during transcription and TF binding compared to their non-CRMC counterparts; while the elevation of C/G contents in nTFBS positions might be due to their prolonged nucleosome-free states during TF binding compared to non-CRMC NESs. The higher C/G contents in cTFBSs and 5'-uTFBSs than in nTFBS positions might be due to the longer time that cTFBSs and 5'-uTFBSs are in nucleosome-free state than are nTFBS positions. The slightly higher C/G contents in non-CRMC 3'-UTRs than in 3'-uTFBSs (Figure 3c) suggest that some 3'-UTR related functions mediated by the middle part of 3'-UTRs where 3'-uTFBSs are depleted (Figure 1f) might spend more time in nucleosome-free states than do 3'-uTFBSs.

Both non-degenerate and degenerate positions in cTFBSs have higher proportions under evolutionary selection than those in non-CRMC CDSs

As indicated earlier, the two earlier studies drew contradictory results about the evolution of duons (32,35). It is interesting to see how our predicted cTFBSs evolve when compared with non-CRMC CDSs. To this end, we compared the phyloP scores (34) of A/T and C/G at the non-degenerate positions at the first and the second codon positions as well as at the degenerate positions at the third codon positions of the 21 synonymous codon sets (Materials and Methods)

in cTFBSs with those in non-CRMC CDSs. The larger a positive phyloP score of a nucleotide position in the genome, the more likely it is under purifying selection; the smaller a negative phyloP score of a position, the more likely it is under positive selection; and a phyloP score around zero means that the position is selectively neutral or nearly so(34). For convenience of discussion, we consider a genomic nucleotide position with a phyloP score within an interval $[-\delta, \delta]$ ($\delta > 0$) to be selectively neutral or nearly so, a position with a score greater than δ to be under purifying selection, and a position with a score less than $-\delta$ to be under positive selection. We define the size of the area under the density curve of the distribution of the scores of a set of genomic nucleotide positions within the intervals $[-\delta, \delta]$, $(\delta, maximum)$ and $(minimum, -\delta)$, to be the proportion of the positions that are selectively neutral or nearly so, under purifying selection and under positive selection, respectively. The choices of δ =0.5, 1 and 1.5, yielded similar results (data not shown), so we used $\delta = 1$ in this study. Evolutionary constraints on the non-degenerate positions at the first and the second codon positions reflect selection pressures from the requirements of encoded amino acids for protein-related functions and other functions of CDSs, if any, such as direct TF binding; while evolutionary constraints on the degenerate positions at the third codon positions indicate selection pressures from the requirements of functions other than protein-coding, e.g. direct TF binding.

Both A/T and C/G at the nondegenerate positions of the first (Figure 4a) and the second (Figure 4b) codon positions as well as at the degenerate positions of the third codon positions (Figure 4c) in cTFBSs have significantly different ($p<2.23\times10^{-302}$, K-S test) phyloP score distributions than those in non-CRMC CDSs, strongly suggesting that cTFBSs evolve very differently from non-CRMC CDSs. Specifically, as expected, most (>60%) of the non-degenerate positions of either the first or the second codon positions are under purifying selection (Figures 4d and 4e). However, either A/T or C/G at the non-degenerate positions of both codon positions in cTFBSs tend to have higher proportions (total difference 30.63%%) under purifying selection than those in non-CRMC CDSs (Figures 4d and 4e, $p < 2.2\times10^{-16}$, χ^2 test). Only a small proportion (~5%) of non-degenerate positions of both codon positions are under positive selection (Figures 4d and 4e). Meanwhile, either A/T or C/G at the non-degenerate positions of both codon positions in the cTFBSs tend to have lower proportions (total difference 2.45%) under positive selection than those in non-CRMC CDSs (Figures 4d and 4e, $p < 2.2\times10^{-16}$, χ^2 test). Less than a third of the non-degenerate positions of both codon positions are selectively neutral or nearly so (Figures 4d~4e). Interestingly, either A/T or C/G at the two non-degenerate codon positions in cTFBSs tend to have lower proportions (total

difference 25.56%) to be selectively neutral or nearly so than those in non-CRMC CDSs (Figures 4d~4e, $p < 2.2x10^{-16}$, χ^2 test). These results indicate that the non-degenerate positions at the first and second codon positions in cTFBSs are more likely to be conserved than those in non-CRMC CDSs. The difference in the proportions of positions under evolutionary constraints suggests that about 25.56% of the non-degenerate positions of the first and the second codon positions in cTFBSs might be in direct contact with TFs, and thus, are in dual-use, i.e., for encoding specific peptides and direct TF binding interfaces, while non-CRMC CDSs might still have at least 25.56% selectively neutral non-degenerate positions that can potentially adopt other functions such as TF binding in the course of evolution without significant detrimental effects on protein functions.

In stark contrast to the non-degenerate positions with >60% under purifying selection (Figures 4d and 4e), less than a third of degenerate positions of the third codon positions are under purifying selection (Figure 4f), Meanwhile, A/T at the degenerate positions in cTFBSs have a lower proportion (20.82%) under purifying selection than those (23.31%) in non-CRMC CDSs (Figure 4f, difference 2.49%, $p < 2.2x10^{-16}$, χ^2 test), while the opposite (34.69% vs 31.74%) is true for C/G at the degenerate positions in cTFBSs (Figures 4f, difference 2.95%, $p < 2.2x10^{-16}$, χ^2 test). About 15% of the degenerate positions are under positive selection (Figure 4f). Interestingly, A/T at the degenerate positions in cTFBSs have higher proportions (20.97%) under positive selection than those (13.19%) in non-CRMC CDSs (Figure 4f, difference 7.78%, $p < 2.2x10^{-16}$, χ^2 test). More than half of the degenerate positions are selectively neutral or nearly so (Figure 4f). Meanwhile, both A/T (58.21%) and C/G (52.80%) at the degenerate positions in cTFBSs have a lower proportion of neutrality than those (63.51% and 56.39%) in non-CRMC CDSs (Figure 4f, total difference 8.89%, $p < 2.2x10^{-16}$, χ^2 test). These results indicate that the degenerate positions in cTFBSs are more likely under either purifying selection (C/G) (difference 2.95%) or positive selection (A/T) (difference 7.78%) than those in non-CRMC CDSs. The difference in the proportions of positions under evolutionary constraints suggests that about 8.89% of the degenerate positions in cTFBSs might be in direct contact with TFs. Therefore, this conclusion is in sharp contrast to either of the two earlier observations that degenerate positions in duons are more conserved than those in non-duons (32), or that both A/T and C/G at degenerate positions in duons are similarly or only slightly more conserved compared with those in non-duon CDSs(35). These results indicate that our predicted cTFBSs are distinct from the duons used in the earlier two studies, which were predicted based on DHSs alone, thus may not necessarily be cTFBSs. Taken together, these results suggest that about

34.45% (25.56%+8.89%) of nucleotide positions in cTFBSs might be in direct contact with TFs for transcriptional regulation, consistent with current knowledge about TF-TFBS interactions that about a third of positions in a TF binding motifs are highly conserved and thus might be involved in direct TF-DNA interactions (60).

Moreover, we reason that if the degenerate positions in cTFBSs are only for TF binding, then they should evolve similarly to nTFBSs that are entirely located in NESs. However, our results show that both A/T and C/G at the degenerate positions in cTFBSs have a higher proportion under purifying selection than those in nTFBSs (Figure 4f, total difference 25.19%, $p < 2.2\times10^{-16}$, χ^2 test), suggesting that about 25.19% of the degenerate positions in cTFBSs might have other functions in addition to TF binding, thus might be in dual-use. Furthermore, we reason that if the degenerate positions in non-CRMC CDSs do not have any function except for being the wobble bases in codons, and thus, are selectively neutral, then they should evolve similarly to nucleotide positions of non-CRMC NESs that are largely selective neutral (41). However, we found that both A/T and C/G at the degenerate positions in non-CRMC CDSs are more likely to be under either positive selection or purifying selection than non-CRMC NESs (Figure 4f, total difference 56.14%, $p < 2.2\times10^{-16}$, χ^2 test), suggesting that about 56.14% of the degenerate positions in non-CRMC CDSs might have other biological functions such as splicing regulation (61-63), than being wobble bases. Similar results were obtained when GERP evolutionary scores(64) were used for all the above analyses (Figures S2a~2f).

5'-uTFBSs and 3'-uTFBSs have higher proportions of positions under either purifying selection or positive selection than corresponding non-CRMC UTRs and nTFBSs

To assess how 5'-uTFBSs and 3'-uTFBSs evolve, we first compared the phyloP scores of their A/T and G/C positions with those in corresponding non-CRMC UTRs. We found that both A/T and C/G in either 5'-uTFBSs (Figure 5a) or 3'-uTFBSs (Figure 5b) had significantly different ($p<2.23\times10^{-302}$, K-S test) phyloP score distributions than those in corresponding non-CRMC UTRs, strongly suggesting that 5'-uTFBSs and 3'-uTFBSs evolve very differently from corresponding non-CRMC UTRs. Specifically, both A/T and C/G in either 5'-uTFBSs (Figure 5c, total difference 33.66%, $p < 2.2\times10^{-16}$, χ^2 test) or 3'-uTFBSs (Figure 5d, total difference 38.88%, $p < 2.2\times10^{-16}$, χ^2 test) have a higher proportion of positions under purifying selection than those in corresponding non-CRMC UTRs. Moreover, A/T in either 5'-uTFBSs (Figure 5c, total difference 3.75%, $p < 2.2\times10^{-16}$, χ^2 test) or 3'-uTFBSs (Figure 5d, total difference 0.43%, $p < 2.2\times10^{-16}$, χ^2 test) have a higher proportion of positions under positive selection than those in

corresponding non-CRMC UTRs. These results suggest that about 36.41% and 39.32% positions of 5'-uTFBSs and 3'-uTFBSs, respectively, are under strong evolutionary constraints, and thus, might be in direct contact with TFs. Furthermore, both A/T and C/G in either 5'-uTFBSs (Figure 5c, total difference 18.91%, $p < 2.2x10^{-16}$, χ^2 test) or 3'-uTFBSs (Figure 5d, total difference 21.19%, $p < 2.2x10^{-16}$, χ^2 test) have higher proportions of positions under purifying selection than those in nTFBSs that are entirely in NESs, suggesting that about 18.91% and 21.19% positions of 5'-uTFBSs and 3'-UTFBs, respectively, might have other UTR-related functions in addition to TF binding, and thus, might be in dual-use. As expected, both A/T and C/G in either non-CRMC 5'-UTRs (Figure 5c, total difference 7.13%, $p < 2.2x10^{-16}$, χ^2 test) or non-CRMC 3'-UTRs (Figure 5d, total difference 4.18%, $p < 2.2x10^{-16}$, χ^2 test) had higher proportions of positions under purifying selection than those in non-CRMC NESs, suggesting that about 7.13% and 4.18% positions of non-CRMC 5'-UTRs and non-CRMC 3'-UTRs, respectively, might be directly involved in UTR-related functions, such as ribosome binding in 5'-UTRs (65) and RNA-binding protein (RBPs) binding in 3'-UTRs (66). Similar results were seen using the GERP scores, with the exception that C/G in either 5'-uTFBSs or 3'-uTFBSs also have a significantly ($p < 2.2x10^{-16}$, χ^2 test) higher proportion of positions under positive selection than those in corresponding non-CRMC UTRs (Figures S3a~S3d).

cTFBSs preferentially code for loops rather than helices and strands

One of the puzzles for the dual-use of codons is how the two irrelevant functions of a DNA sequence can be possibly co-evolved. To address this, we mapped amino acids encoded by cTFBSs to known 3D structures of proteins in PDB (Materials and Methods). To reduce the biased distribution of structures to some protein families, we generated a non-redundant protein set with a 30% sequence identity cutoff (Materials and Methods), whose CDSs contained a total of 7,266,384bp cTFBS positions. Amino acids encoded by 7.49% of the 7,266,384bp cTFBS positions can be mapped to 1,761 known protein structures. These mapped amino acids are enriched in loops (48.73%, χ^2 test) compared with the proportion of length of loops in host proteins (46.65%, $p < 2.2x10^{-16}$) as well as in a non-redundant of all known protein structures(67) (47.10%, $p < 2.2x10^{-16}$, χ^2 test) (Figure 6a). On the other hand, these mapped amino acids are under-represented in helices (33.53%) and strands (17.74%) compared with the proportions of helices and strands in the host proteins (34.31% and 19.04%, respectively, $p < 2.2x10^{-16}$, χ^2 test) as well as in all proteins with known structures(67) (34.21% and 18.69%, respectively, $p < 2.2x10^{-16}$, χ^2 test) (Figure 6a). For amino acids encoded by the remaining

92.51% of the cTFBSs positions, which could not be mapped to any known protein structures, we predicted their secondary structure types using RaptorX (50). A similar pattern is found showing that amino acids encoded by these cTFBSs are enriched in the predicted loops but depleted in the predicted helices and strands (Figure 6b). Specifically, 60.79%, 30.01%, and 9.20% of the peptides encoded by cTFBSs are predicted to adopt loops, helices and strands, respectively, while these proportions are 57.30%, 31.92%, and 10.78% ($p < 2.2x10^{-16}$, χ^2 test), respectively, for the host proteins; and 58.20%, 31.44%, and 10.36% ($p < 2.2x10^{-16}$, χ^2 test), respectively, for all the proteins with predicted secondary structures (Figure 6b). It has been shown that loops are generally less conserved than helices and strands(67), and we see the same results for both known (Figures S4a and S4b) and predicted (Figures S4c and S4d) secondary structure types. Since the folding of a protein is mainly determined by its helix and strand structures, but not its loops, changes in amino acids in the loops are less likely to alter the overall folding and functions of the protein. In this regard, it is not surprising that the predicted cTFBSs tend to encode amino acids in loops where purifying selection are weaker (Figures S4a~S4d), and therefore they could adopt for specific TF binding without compromising the overall structures of proteins.

Active eTFBSs tend to be in close physical proximity to distal promotors in topologically associating domains

Using chromatin conformation capture techniques(68) such as HiC(69), it is now well established that the linear genomic DNA is folded into largely conserved topologically associating domains (TADs) in the nuclei of cells(70-72), where enhancers interact with promoters via looping over long distances for transcriptional regulation(73,74). Therefore, we hypothesize that active eTFBSs, but not inactive eTFBSs, must be in close physical proximity to the promoters of target genes in TADs and involved in transcriptional regulation, although the promoters may be linearly far away from the eTFBSs. To test this hypothesis, we used a "chromatin interaction analysis by paired-end tagging" (ChIA-PED) dataset generated in K562 cells using an antibody against hypomethylated Pol II at Ser2(45). ChIA-PED is a variant of HiC, which combines ChIP-seq with HiC to identify DNA loci that are physically close to sequences bound by the antibody-targeted protein(45). Pol II hypomethylated at Ser2 stalls at promoters in the preinitiation complex(75), thus the resulting ChIA-PED data are enriched for reads from loci in close physical proximity to promoters. Since an eTFBS tends to be close to the promoter of its host gene, we exclude the promoters of the host genes of eTFBSs (Materials and Methods). We predicted 31.38%, 72.90% and 24.58% of the 1,047,183 cTFBSs, 129,518 5'-uTFBSs and

733,089 3'-uTFBSs, respectively, to be active in K562 cells, and the remaining 68.62% cTFBSs, 27.10% 5'-uTFBSs and 75.42% 3'-uTFBSs to be inactive in the cells (Materials and Methods). Similar to the averaged results in 19 cell/tissue types (Figure 2a), in K562 cells, genes closest to predicted active cTFBSs, 5'-uTFBSs and 3'-uTFBSs also have significantly higher expression levels than those closest to inactive cTFBSs, 5'-uTFBSs and 3'-uTFBSs (Figure 6c, $p<2.2x10^{-302}$ for all comparisons with Mann Whitney U test), respectively, suggesting that our prediction of the functional states (active and inactive) of the eTFBSs are highly accurate in K562 cells. As shown in Figure 6d, the predicted active cTFBSs (18.44%), 5'-uTFBSs (23.78%) and 3'-uTFBSs (15.72%) have a significantly higher proportion in close physical proximity to distal promoters than the predicted inactive cTFBS (3.17%), 5'-UTFBSs (1.85%) and 3-uTFBSs (2.92%) ($p < 2.23x10^{-302}$ for all comparisons with χ^2 test), respectively.

To tease out confounding factors of using ChIP-PED data in validating our predicted eTFBSs, we examined the transcription levels in K562 cells of genes 1) closest to promotors that are in close proximity to active eTFBSs (aeTFBSs+P^+), 2) closest to promotors that are in close proximity to inactive eTFBSs (ieTFBSs+P^+), 3) closest to active eTFBSs that are not in close proximity to promoters (aeTFBSs+P^-), and 4) closest to inactive eTFBSs that are not in close proximity to promoters (ieTFBSs+P^-). As shown in Figure 6e, putative target genes of aeTFBSs+P^+ (acTFBSs+P^+, a5'-uTFBSs+P^+ and a3'-uTFBSs+P^+) have similarly high transcription levels as those of corresponding ieTFBSs+P^+(icTFBSs+P^+, i5'-uTFBSs+P^+ and i3'-uTFBSs+P^+, respectively), indicating that both predicted active and predicted inactive eTFBSs that are in close physical proximity to promoters might be involved enhanced transcription of the target genes, while we might falsely predicted the functional states of "inactive eTFBSs" that are in close physical proximity to promoters. Nonetheless, these false negatives consist of only a small portion (<3%) of the predicted inactive eTFBSs (Figure 6d). On the other hand, genes closest to our predicted aeTFBSs+P^- (acTFBSs+P^-, a5'-uTFBSs+P^- and a3'-uTFBSs+P^-) have significantly higher transcription levels than those closest to corresponding ieTFBSs+P^- (icTFBSs+P^-, i5'-uTFBSs+P^- and i3'-uTFBSs+P^-, respectively, $p<2.23x10^{-302}$ for all comparisons, Figure 6e), suggesting that our predicted active eTFBSs that do not overlap ChIA-PED peaks tend to activate nearby promoters. It is highly likely that these promoters might be missed by the ChIA-PED method. Taken together, these results unequivocally demonstrate that our predicted active eTFBSs indeed tend to be in close physical proximity to distal promoters, and involved in enhanced transcriptions of target genes, thus they might be constituent TFBSs of eEHs/eCRMs.

DISCUSSION

It has long been known that in addition to encoding amino acids and UTR function-related signals, exons can also encode other information, including splicing enhancers(61-63), overlapping ORFs(76), lincRNA(77) and transcriptional enhancers(8-24). It has been estimated that up to 25% of codons in the human genome may code for such overlapping functions based on evolutionary constraints on the degenerate positions in synonymous codons(56). Although there are numerous experimentally verified cases for each of these dual-use codons and UTRs, it is under hot debate as to how prevalent they are in the human genome(32,78), what evolutionary constraints these codons and UTRs have been subject to in the course of evolution(32,61,63,79), and how it is possible for a sequence to evolve for such two unrelated functions. For instance, Stergachis and colleagues (32) showed that up to 15% codons in 86.9% genes were duons predicted using DHSs in 81 cell/tissue types, and that these duons were under strong purifying selection. However, others (35,36) concluded that duons were largely selectively neutral and thus might not be functional. The discrepancies may result from different methods employed(32,35,61) and/or different datasets used (32,36,61,78). Clearly, the lack of a large experimentally verified positive and negative datasets for eEHs has largely hampered clarifying these contradictory results and addressing related issues of the dual-use of exons. Our recent prediction of large sets of CRMCs and non-CRMCs with high accuracy in the human genome provides us an opportunity to address these issues.

A CRM is made up of clusters of TFBSs and spacer sequences between adjacent TFBSs. TFBSs consist of 40% of the length of a CRM(41), but they play more critical roles in transcriptional regulation than spacers. Moreover, most eCRMs have less than 10% of their lengths overlapping exons. Therefore, in this study we used predicted eTFBSs located in experimentally verified exons to address the aforementioned questions regarding dual-use of exons. We showed that genes closest to active eTFBSs had elevated transcription levels compared to those closest to inactive eTFBSs using RNA-seq data in 19 cell/tissue types in which we were able to predict the functional states of the TFBSs. Although due to their high FDRs and FNRs and length limitation (30,31), STARR-seq peaks could not be used as gold standards to validate our predicted eCRMs and eTFBSs, they are significantly enriched in our predicted active eTFBSs compared to in inactive eTFBSs. More importantly, genes closest to active eTFBSs that overlap STARR-seq peaks have elevated transcription levels, suggesting that the exonic STARR-seq peaks are bona fide eEHs and the eTFBSs are parts of them.

Although most of our predicted active eTFBSs do not overlap STARR-seq peaks, gene closest to them have significantly higher transcription level than those closest to inactive eTFBSs, strongly suggesting that they are likely parts of long active eEHs that were missed by the STARR-seq methods. All these functional genomics data indicate that our predicted eTFBSs might be involved in transcriptional regulation of nearby genes. Surprisingly, we found that the cTFBSs, 5'-uTFBSs and 3'-uTFBSs comprised more than a third of the total length of annotated CDSs, 5'-UTRs, and 3'-UTRs, respectively (Figure 1e), and most genes had at least an exon partially overlapping a CRM. Therefore, eTFBSs and eEHs/eCRMs might be more prevalent than originally thought (32,56).

Evolutionary behaviors of the eTFBSs also support their TF binding functions. First, we found that both A/T and C/G at the non-degenerate positions of either the first (Figures 4a, 4d, S2a, S2d), or the second (Figures 4b, 4e, S2b, S2e) positions of codons in the cTFBSs were more likely to be under purifying selection than those in non-CRMC CDSs, strongly suggesting that the non-degenerate positions in the cTFBSs might be in dual-use for encoding specific peptides and TF binding interfaces. Second, degenerate positions at the third codon positions in the 21 synonymous codon sets in the cTFBSs are more likely under evolutionary constraints (either positive selection for A/T or purifying selection for CG) than those in non-CRMC CDSs(Figures 4c, 4f, S2c, S2f), suggesting that they might be involved in TF binding. Interestingly, both A/T and C/G at degenerate positions in cTFBSs also are more likely under purifying selection than those in nTFBSs (Figures 4c, 4f, S2c and S2f), suggesting that these degenerate positions might be in dual-use for TF binding and unknown non-coding functions. This conclusion therefore is different from the two earlier contradictory ones drawn by Stergachis et al. (32) and Xing/He (35) using the duons predicted based on DHSs. We (41) and others (37,39,40,80) have shown that enhancers predicted based solely on DHSs have high FDRs. It is highly likely that a high FDR of duons might lead to the earlier conclusions. Finally, 5'-uTFBSs and 3'-uTFBSs are more likely to be under purifying selection than their non-CRMC counterparts (Figures 5a~5d, S3a~S3d), suggesting that they might be involved in TF binding. Interestingly, both 5'-uTFBSs and 3'-uTFBSs also are more likely to be under purifying selection than nTFBSs, suggesting that the uTFBSs might be in dual-use for TF binding and other UTR-related functions.

The paradox concerning the dual-use of exons is how a DNA sequence could evolve two unrelated functions such as TF binding and encoding amino acids or UTR-related functions. For

instance, the amino acid coding function would require the DNA sequence to specify a specific peptide that plays a role in the protein's function, while the TF binding function would demand the sequence to adopt an specific interface to which the cognate TF can bind. Although an earlier study(11) suggests that nature solves this paradox by opting exonic remnants from whole-genome duplication for eEHs in some cases, this mechanism cannot explain the prevalent dual-use of exons as we show in this study. Our findings might provide a more general answer to the puzzle. First, at least a quarter of non-degenerate positions at the first and second codon positions in non-CRMC CDSs are selectively neutral or nearly so (Figures 4d, 4e, S2d, S3e), thus they are potentially allowed to evolve into cTFBSs without detrimental effects on protein functions. Interestingly, the proportion of selectively neutral non-degenerate positions in cTFBSs is smaller than in non-CRMC CDSs (Figures 4d, 4e, S2d, S3e), suggesting that once neutral positions become cTFBSs, they are under purifying selection. This might explain why the proportion of neutrality of non-degenerate positions in cTFBSs is lower than that of those in non-CRMC CDSs (Figures 4d, 4e, S2d and S2e). Second, amino acids encoded by cTFBSs tend to be located in structurally and functional less critical loops, and avoid structurally and functionally more important helices and strands (Figures 6a and 6b), thereby reducing detrimental effects of evolving codons into cTFBSs. Third, the preferential locations of cTFBSs at 5'-ends and 3'-ends of a concatenated CDS (Figure 1f) suggest that cTFTBs tend to encode amino acids at N-terminus and C-terminus of a protein. These termini probably have less crucial functions than those in the middle of the whole peptide. Fourth, 5'-uTFBSs tend to be located at the middle of 5'-UTRs (Figure 1f), and avoid the two ends where 5'-UTR function-related sequences are encoded, such as transcription start sites (TSSs), ribosome entry sites and upstream open-reading frames (65,81). Finally, 3'-uTFBSs tend to appear at the two ends of 3'-UTRs (Figure 1f), and avoid the middle where 3'-UTR function-related sequences might be coded, such as polyadenylation sites, miRNA response elements and RBP(RNA binding protein) binding sites (66). Therefore, certain level of prevalent dual-use of exons is possible.

The next question is, when only 4.1% of the human genome code for exons (82) and the remaining 95.9% are NESs that can be potentially used to encode CRMs, why are sparse exons exploited to encode TFBSs in CRMs? Our finding that active cTFBSs, 5'-uTFBSs and 3'-uTFBSs tend to be in close physical proximity to distal promoters in TADs may provide an explanation. When chromatins are folded into relatively conserved 3D structures (70-72), linearly distal target genes, promoters and enhancers are brought in close physical proximity in compartments such as TADs that typically span around 1 Mbp (73,74,83). Since functionally

related genes tend to cluster together in linear DNA(71,84), it is highly likely that there is not enough NESs in close proximity to some promoters to function as their enhancers duo to space constraints in a TAD(85). In such a scenario, a few nucleotides in a CDS or UTR in close proximity to a promoter may well likely be opted for cTFBSs, if they code for less important amino acids such as those in some loops, termini of proteins, or a less critical part of UTRs. In this regard, it seems that dual-use of some exons is unavoidable, and nature chooses less critical exons for eTFBSs, thereby avoiding the dilemma of evolving a sequence for two unrelated functions.

Our finding that non-degenerate positions tend to be under strongly purifying selection, while degenerate positions tend to be subject to either strongly purifying selection or strongly positive selection, might suggest a parsimonious scenario of how a stretch of less critical codons evolve into a cTFBS: When no NESs are in close proximity to a (potential) promoter, nature might choose physically available codons whose non-degenerate positions match important positions in a TFBS to be evolved into. As these non-degenerate positions would be subject to double purifying selections, they are more conserved than non-CRMC codons as we demonstrated (Figures 4a, 4b, S2a, S2b). At the same time, nature chooses the non-degenerate positions of the codons to either mutate to the desired nucleotides or remain the same if they match the desired nucleotides in the TFBS to be evolved into. In the former case, these non-degenerate positions would be under positive selection, while in the latter case they would be subject to purifying selection, as we observed (Figures 4c and S2c). Moreover, a similarly parsimonious scenario of how a stretch of less critical nucleotides in a UTR that are in close physical proximity to a (potential) promoter evolve into a uTFBS can be envisaged based on our finding that uTFBS positions tend to be subject to either strongly purifying selection or strongly positive selection (Figures 5 and S3).

SUPPLEMENTARY DATA

Supplementary data are available online.

FUNDING

This work was supported by National Science Foundation (DBI-1661332 to Z.S. and DBI-2051491 to J.G); and National Institutes of Health (R15GM132846 to J.G).

CONFLICT OF INTEREST

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We would like to thank Dr. Way Sung for stimulating discussions and comments on the manuscript.

REFERENCES

1. Pennisi, E. (2004) Searching for the genome's second code. *Science*, **306**, 632-635.
2. Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol*, **13**, 418.
3. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
4. Narlikar, L. and Ovcharenko, I. (2009) Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic*, **8**, 215-230.
5. Negre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R. *et al.* (2011) A cis-regulatory map of the Drosophila genome. *Nature*, **471**, 527-531.
6. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116-120.
7. Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147-151.
8. Neznanov, N., Umezawa, A. and Oshima, R.G. (1997) A regulatory element within a coding exon modulates keratin 18 gene expression in transgenic mice. *J Biol Chem*, **272**, 27549-27557.
9. Tumpel, S., Cambronero, F., Sims, C., Krumlauf, R. and Wiedemann, L.M. (2008) A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *Proc Natl Acad Sci U S A*, **105**, 20077-20082.
10. Barthel, K.K. and Liu, X. (2008) A transcriptional enhancer from the coding region of ADAMTS5. *PLoS One*, **3**, e2184.
11. Dong, X., Navratilova, P., Fredman, D., Drivenes, O., Becker, T.S. and Lenhard, B. (2010) Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. *Nucleic Acids Res.*, **38**, 1071-1085. doi: 1010.1093/nar/gkp1124. Epub 2009 Dec 1076.
12. Birnbaum, R.Y., Clowney, E.J., Agamy, O., Kim, M.J., Zhao, J., Yamanaka, T., Pappalardo, Z., Clarke, S.L., Wenger, A.M., Nguyen, L. *et al.* (2012) Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.*, **22**, 1059-1068. doi: 1010.1101/gr.133546.133111. Epub 132012 Mar 133522.
13. Li, M., Zhao, H., Wei, J., Zhang, J. and Hong, Y. (2015) Medaka vasa gene has an exonic enhancer for germline expression. *Gene*, **555**, 403-408.
14. Hirsch, N. and Birnbaum, R.Y. (2015) Dual Function of DNA Sequences: Protein-Coding Sequences Function as Transcriptional Enhancers. *Perspectives in biology and medicine*, **58**, 182-195.
15. Lang, G., Gombert, W.M. and Gould, H.J. (2005) A transcriptional regulatory element in the coding sequence of the human Bcl-2 gene. *Immunology*, **114**, 25-36.
16. Lampe, X., Samad, O.A., Guiguen, A., Matis, C., Remacle, S., Picard, J.J., Rijli, F.M. and Rezsohazy, R. (2008) An ultraconserved Hox-Pbx responsive element resides in the

coding sequence of Hoxa2 and is active in rhombomere 4. *Nucleic Acids Res*, **36**, 3214-3225.
17. Yang, J.Q., Remmers, E.F. and Marcu, K.B. (1986) The first exon of the c-myc proto-oncogene contains a novel positive control element. *EMBO J*, **5**, 3553-3562.
18. Hurt, M.M., Bowman, T.L. and Marzluff, W.F. (1991) A common transcriptional activator is located in the coding region of two replication-dependent mouse histone genes. *Mol Cell Biol*, **11**, 2929-2936.
19. Farnham, P.J. and Means, A.L. (1990) Sequences downstream of the transcription initiation site modulate the activity of the murine dihydrofolate reductase promoter. *Mol Cell Biol*, **10**, 1390-1398.
20. Mous, J., Stunnenberg, H., Georgiev, O. and Birnstiel, M.L. (1985) Stimulation of sea urchin H2B histone gene transcription by a chromatin-associated protein fraction depends on gene sequences downstream of the transcription start site. *Mol Cell Biol*, **5**, 2764-2769.
21. Hoeben, R.C., Fallaux, F.J., Cramer, S.J., van den Wollenberg, D.J., van Ormondt, H., Briët, E. and van der Eb, A.J. (1995) Expression of the blood-clotting factor-VIII cDNA is repressed by a transcriptional silencer located in its coding region. *Blood*, **85**, 2447-2454.
22. McLellan, A.S., Kealey, T. and Langlands, K. (2006) An E box in the exon 1 promoter regulates insulin-like growth factor-I expression in differentiating muscle cells. *Am J Physiol Cell Physiol*, **291**, C300-307.
23. Chiquet, M., Mumenthaler, U., Wittwer, M., Jin, W. and Koch, M. (1998) The chick and human collagen alpha1(XII) gene promoter--activity of highly conserved regions around the first exon and in the first intron. *Eur J Biochem*, **257**, 362-371.
24. Chen, H.P., Lin, A., Bloom, J.S., Khan, A.H., Park, C.C. and Smith, D.J. (2008) Screening reveals conserved and nonconserved transcriptional regulatory elements including an E3/E4 allele-dependent APOE coding region enhancer. *Genomics*, **92**, 292-300.
25. Bhatia, S. and Kleinjan, D.A. (2014) Disruption of long-range gene regulation in human genetic disease: a kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Hum Genet.*, **133**, 815-845. doi: 810.1007/s00439-00014-01424-00436. Epub 02014 Feb 00435.
26. Tayebi, N., Jamsheer, A., Flottmann, R., Sowinska-Seidler, A., Doelken, S.C., Oehl-Jaschkowitz, B., Hulsemann, W., Habenicht, R., Klopocki, E., Mundlos, S. *et al.* (2014) Deletions of exons with regulatory activity at the DYNC1I1 locus are associated with split-hand/split-foot malformation: array CGH screening of 134 unrelated families. *Orphanet journal of rare diseases*, **9**, 108.
27. Lango Allen, H., Caswell, R., Xie, W., Xu, X., Wragg, C., Turnpenny, P.D., Turner, C.L., Weedon, M.N. and Ellard, S. (2014) Next generation sequencing of chromosomal rearrangements in patients with split-hand/split-foot malformation provides evidence for DYNC1I1 exonic enhancers of DLX5/6 expression in humans. *J Med Genet*, **51**, 264-267.
28. Birnbaum, R.Y., Patwardhan, R.P., Kim, M.J., Findlay, G.M., Martin, B., Zhao, J., Bell, R.J., Smith, R.P., Ku, A.A., Shendure, J. *et al.* (2014) Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genet*, **10**, e1004592.
29. Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M. and Stark, A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074-1077.

30. Liu, Y., Yu, S., Dhiman, V.K., Brunetti, T., Eckart, H. and White, K.P. (2017) Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol*, **18**, 219.
31. Wang, X., He, L., Goggin, S.M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M. and Kellis, M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun*, **9**, 5380.
32. Stergachis, A.B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E.M., Akey, J.M. *et al.* (2013) Exonic transcription factor binding directs codon choice and affects protein evolution. *Science.*, **342**, 1367-1372. doi: 1310.1126/science.1243490.
33. Yadav, V.K., Smith, K.S., Flinders, C., Mumenthaler, S.M. and De, S. (2016) Significance of duon mutations in cancer genomes. *Sci Rep*, **6**, 27437.
34. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110-121.
35. Xing, K. and He, X. (2015) Reassessing the "duon" hypothesis of protein evolution. *Mol Biol Evol.*, **32**, 1056-1062. doi: 1010.1093/molbev/msu1409. Epub 2015 Jan 1012.
36. Agoglia, R.M. and Fraser, H.B. (2016) Disentangling Sources of Selection on Exonic Transcriptional Enhancers. *Mol Biol Evol*, **33**, 585-590.
37. Kwasnieski, J.C., Fiore, C., Chaudhari, H.G. and Cohen, B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res*, **24**, 1595-1602.
38. Kleftogiannis, D., Kalnis, P. and Bajic, V.B. (2016) Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform*, **17**, 967-979.
39. Dogan, N., Wu, W., Morrissey, C.S., Chen, K.B., Stonestrom, A., Long, M., Keller, C.A., Cheng, Y., Jain, D., Visel, A. *et al.* (2015) Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics & chromatin*, **8**, 16.
40. Arbel, H., Basu, S., Fisher, W.W., Hammonds, A.S., Wan, K.H., Park, S., Weiszmann, R., Booth, B.W., Keranen, S.V., Henriquez, C. *et al.* (2019) Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proc Natl Acad Sci U S A*, **116**, 900-908.
41. Ni, P. and Su, Z. (2021) Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. *NAR Genom Bioinform*, **3**, lqab052.
42. Visel, A., Minovitsky, S., Dubchak, I. and Pennacchio, L.A. (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res*, **35**, D88-92.
43. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*, **46**, D1062-d1067.
44. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.
45. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84-98.
46. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*, **35**, D760-765.
47. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res*, **47**, D729-d735.
48. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

49. Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680-682.
50. Wang, S., Li, W., Liu, S. and Xu, J. (2016) RaptorX-Property: a web server for protein structure property prediction. *Nucleic Acids Res*, **44**, W430-435.
51. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A. and Batzoglou, S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol.*, **6**, e1001025. doi: 1001010.1001371/journal.pcbi.1001025.
52. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D. *et al.* (2018) The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res*, **46**, D762-d769.
53. Paulsen, J., Rodland, E.A., Holden, L., Holden, M. and Hovig, E. (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res*, **42**, e143.
54. Spitz, F. and Furlong, E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, **13**, 613-626.
55. Erceg, J., Saunders, T.E., Girardot, C., Devos, D.P., Hufnagel, L. and Furlong, E.E. (2014) Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genet*, **10**, e1004060.
56. Lin, M.F., Kheradpour, P., Washietl, S., Parker, B.J., Pedersen, J.S. and Kellis, M. (2011) Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res*, **21**, 1916-1928.
57. Peng, T., Zhai, Y., Atlasi, Y., Ter Huurne, M., Marks, H., Stunnenberg, H.G. and Megchelenbrink, W. (2020) STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol*, **21**, 243.
58. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature.*, **409**, 860-921.
59. Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S.F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C. *et al.* (2018) Evolutionary determinants of genome-wide nucleotide composition. *Nature ecology & evolution*, **2**, 237-240.
60. Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordan, R. and Rohs, R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.*, **39**, 381-399. doi: 310.1016/j.tibs.2014.1007.1002. Epub 2014 Aug 1014.
61. Caceres, E.F. and Hurst, L.D. (2013) The evolution, impact and properties of exonic splice enhancers. *Genome Biol*, **14**, R143.
62. Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T. and Ast, G. (2006) Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. *Mol Cell*, **22**, 769-781.
63. Savisaar, R. and Hurst, L.D. (2018) Exonic splice regulation imposes strong selection at synonymous sites. *Genome Res*, **28**, 1442-1454.
64. Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J. and Nickerson, D.A. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods.*, **7**, 250-251. doi: 210.1038/nmeth0410-1250.
65. Hinnebusch, A.G., Ivanov, I.P. and Sonenberg, N. (2016) Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413-1416.
66. Mayr, C. (2019) What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol*, **11**.
67. Kim, R. and Guo, J.T. (2010) Systematic analysis of short internal indels and their impact on protein folding. *BMC structural biology*, **10**, 24.
68. Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306-1311.

69. Belton, J.M., McCord, R.P., Gibcus, J.H., Naumova, N., Zhan, Y. and Dekker, J. (2012) Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*.
70. McArthur, E. and Capra, J.A. (2021) Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet*, **108**, 269-283.
71. Krefting, J., Andrade-Navarro, M.A. and Ibn-Salem, J. (2018) Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol*, **16**, 87.
72. Rowley, M.J., Nichols, M.H., Lyu, X., Ando-Kuri, M., Rivera, I.S.M., Hermetz, K., Wang, P., Ruan, Y. and Corces, V.G. (2017) Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Mol Cell*, **67**, 837-852.e837.
73. de Laat, W. and Duboule, D. (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, **502**, 499-506.
74. Furlong, E.E.M. and Levine, M. (2018) Developmental enhancers and chromosome topology. *Science*, **361**, 1341-1345.
75. Buratowski, S. (2009) Progression through the RNA polymerase II CTD cycle. *Mol Cell*, **36**, 541-546.
76. Pavesi, A., Vianelli, A., Chirico, N., Bao, Y., Blinkova, O., Belshaw, R., Firth, A. and Karlin, D. (2018) Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One*, **13**, e0202513.
77. Ning, Q., Li, Y., Wang, Z., Zhou, S., Sun, H. and Yu, G. (2017) The Evolution and Expression Pattern of Human Overlapping lncRNA and Protein-coding Gene Pairs. *Sci Rep*, **7**, 42775.
78. Savisaar, R. and Hurst, L.D. (2017) Estimating the prevalence of functional exonic splice regulatory information. *Hum Genet*, **136**, 1059-1078.
79. Savisaar, R. and Hurst, L.D. (2016) Purifying Selection on Exonic Splice Enhancers in Intronless Genes. *Mol Biol Evol*, **33**, 1396-1418.
80. Kleftogiannis, D., Kalnis, P. and Bajic, V.B. (2015) DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res*, **43**, e6.
81. Leppek, K., Das, R. and Barna, M. (2018) Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol*, **19**, 158-174.
82. Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M.C. and Vitale, L. (2016) GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. *Database : the journal of biological databases and curation*, **2016**.
83. Gorkin, D.U., Leung, D. and Ren, B. (2014) The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell*, **14**, 762-775.
84. Thévenin, A., Ein-Dor, L., Ozery-Flato, M. and Shamir, R. (2014) Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. *Nucleic Acids Res*, **42**, 9854-9861.
85. Sun, F., Chronis, C., Kronenberg, M., Chen, X.F., Su, T., Lay, F.D., Plath, K., Kurdistani, S.K. and Carey, M.F. (2019) Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. *Mol Cell*, **73**, 250-263.e255.
86. Kleinjan, D.A., Bancewicz, R.M., Gautier, P., Dahm, R., Schonthaler, H.B., Damante, G., Seawright, A., Hever, A.M., Yeyati, P.L., van Heyningen, V. *et al.* (2008) Subfunctionalization of duplicated zebrafish pax6 genes by cis-regulatory divergence. *PLoS Genet*, **4**, e29.
87. Neznanov, N., Thorey, I.S., Ceceña, G. and Oshima, R.G. (1993) Transcriptional insulation of the human keratin 18 gene in transgenic mice. *Mol Cell Biol*, **13**, 2214-2223.

88. Thorey, I.S., Ceceña, G., Reynolds, W. and Oshima, R.G. (1993) Alu sequence involvement in transcriptional insulation of the keratin 18 gene in transgenic mice. *Mol Cell Biol*, **13**, 6742-6751.
89. Nonchev, S., Vesque, C., Maconochie, M., Seitanidou, T., Ariza-McNaughton, L., Frain, M., Marshall, H., Sham, M.H., Krumlauf, R. and Charnay, P. (1996) Segmental expression of Hoxa-2 in the hindbrain is directly regulated by Krox-20. *Development*, **122**, 543-554.
90. Nonchev, S., Maconochie, M., Vesque, C., Aparicio, S., Ariza-McNaughton, L., Manzanares, M., Maruthainar, K., Kuroiwa, A., Brenner, S., Charnay, P. *et al.* (1996) The conserved role of Krox-20 in directing Hox gene expression during vertebrate hindbrain segmentation. *Proc Natl Acad Sci U S A*, **93**, 9339-9345.
91. Maconochie, M., Krishnamurthy, R., Nonchev, S., Meier, P., Manzanares, M., Mitchell, P.J. and Krumlauf, R. (1999) Regulation of Hoxa2 in cranial neural crest cells involves members of the AP-2 family. *Development*, **126**, 1483-1494.

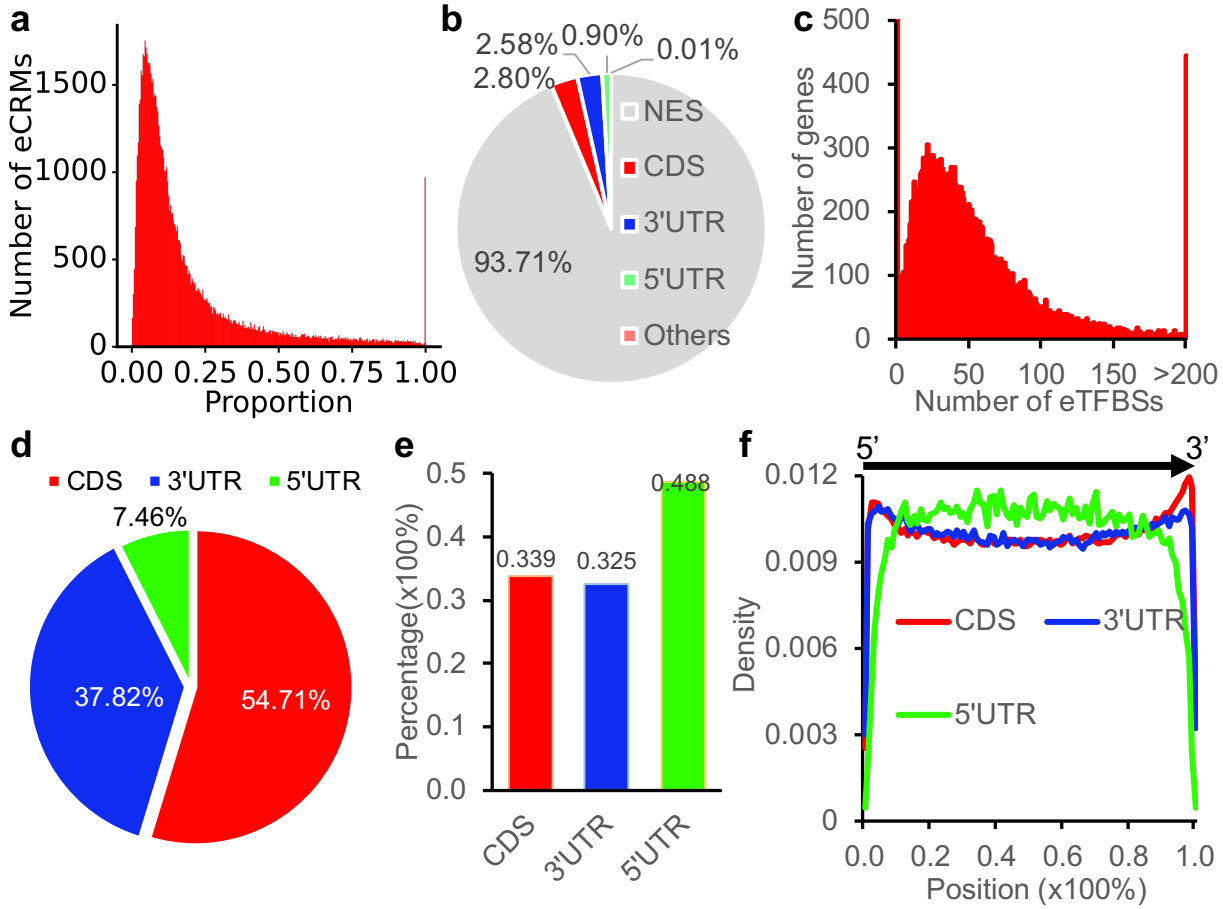

Figure 1. Properties of the predicted eCRMs and eTFBSs in the human genome. **a.** Number of eCRMs with different proportions of exonic sequences. **b.** Proportions of the predicted TFBS positions falling in NESs, CDSs, 5'-UTRs, 3'-UTRs or other sequences (with both 5'-UTR and 3'-UTR annotations). **c.** Number of genes containing different numbers of predicted eTFBSs. **d**. Percentages of eTFBS positions falling in CDSs, 5'-UTRs or 3'-UTRs. **e**. Percentage of eTFBS positions in the total length of CDSs, 5'-UTRs or 3'-UTRs. **f**. Distribution of eTFBSs along CDSs, 5'-UTRs and 3'-UTRs of genes from the 5'-end to the 3'-end as indicated by the horizontal arrow. A CDS is concatenated by all coding exons of along a gene.

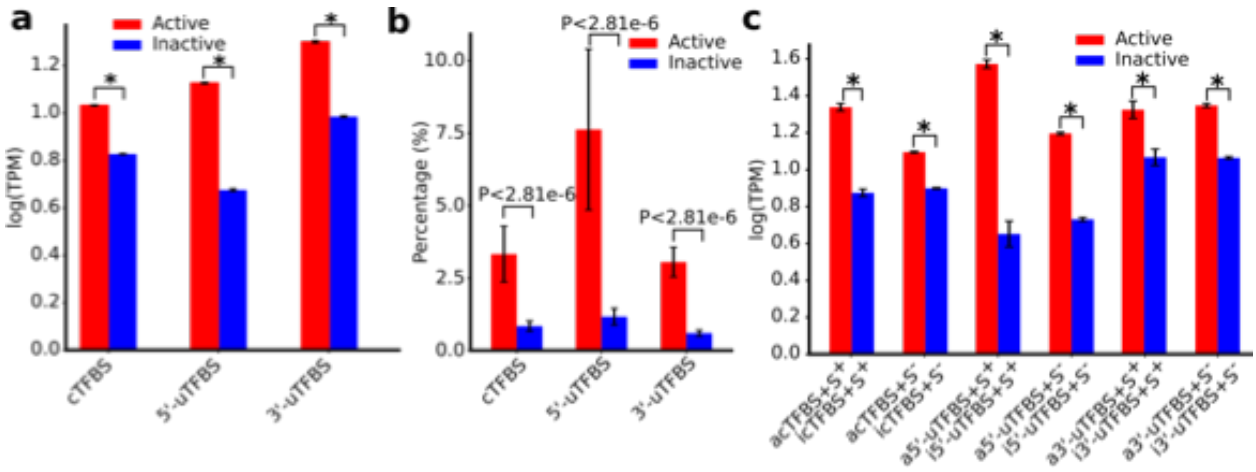

Figure 2. Active eTFBSs are involved in enhanced transcription of nearby genes. **a.** Transcription levels of the closest genes to active eTFBSs in comparison with those of the closest genes to inactive eTFBSs in 19 cell/tissue types. **b.** Percentages of active eTFBSs and inactive eTFBSs overlapping ATAC-STARR-seq peaks in five cell lines. **c.** Transcription levels in the five cell lines of genes closest to 1) active eTFBSs that overlap STARR-seq peaks (aeTFBSs+S^+), 2) inactive eTFBSs that overlap STARR-seq peaks (ieTFBSs+S^+), 3) active eTFBSs that do not overlap STARR-seq peaks (aeTFBSs+S^-), and 4) inactive eTFBSs that do not overlap STARR-seq peaks (ieTFBSs+S^-). The data are shown in mean±SE. *$p<2.23\times10^{-302}$. p-values were computed using the Mann Whitney U test (a, c) or χ^2 test (b).

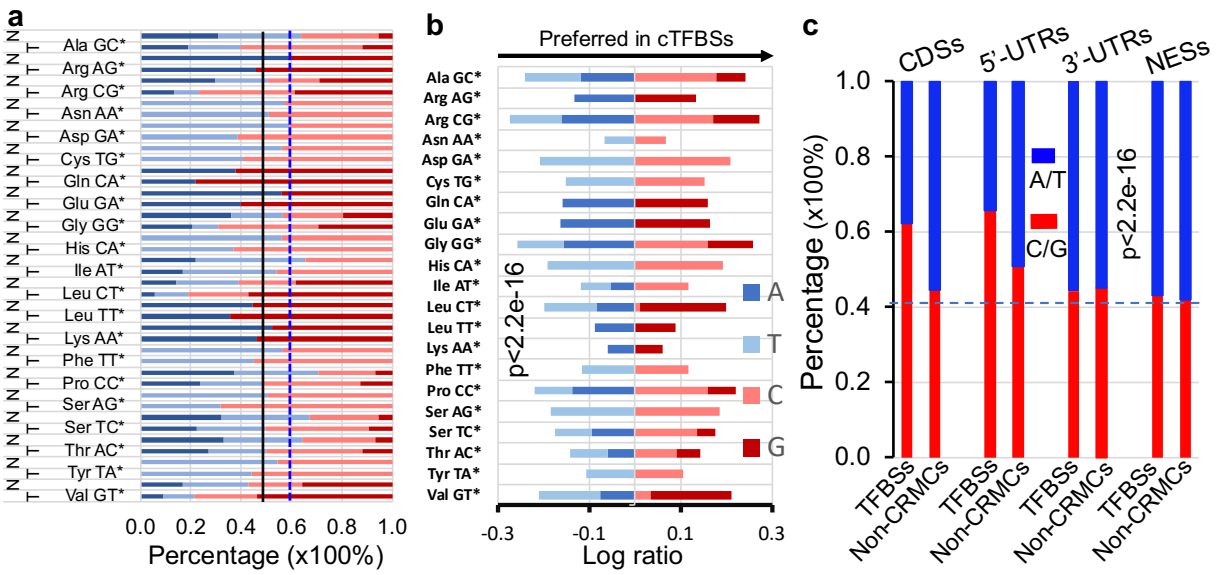

Figure 3. Biased distribution of A/T and C/G in predicted TFBSs and non-CRMCs. **a.** Nucleotide frequencies at the third positions of the 21 synonymous codon sets in cTFBSs (T) and in non-CRMC CDSs (N). The dotted vertical line indicates the neutral expectation (59%) of A/T contents in the genome. **b.** Log ratio of A/T and C/G contents at the third positions of the 21 synonymous codon sets in the cTFBSs over the those in non-CRMC CDSs. The p-value was computed using the χ^2 test. **c.** A/T and C/G contents in the TFBSs in CDSs (degenerate positions at the third positions of the synonymous codons), 5'-UTRs, 3'-UTRs, and NESs in comparison with those in their non-CRMC counterparts. The dotted horizontal line indicates the neutral expectation (41%) of C/G contents in the genome.

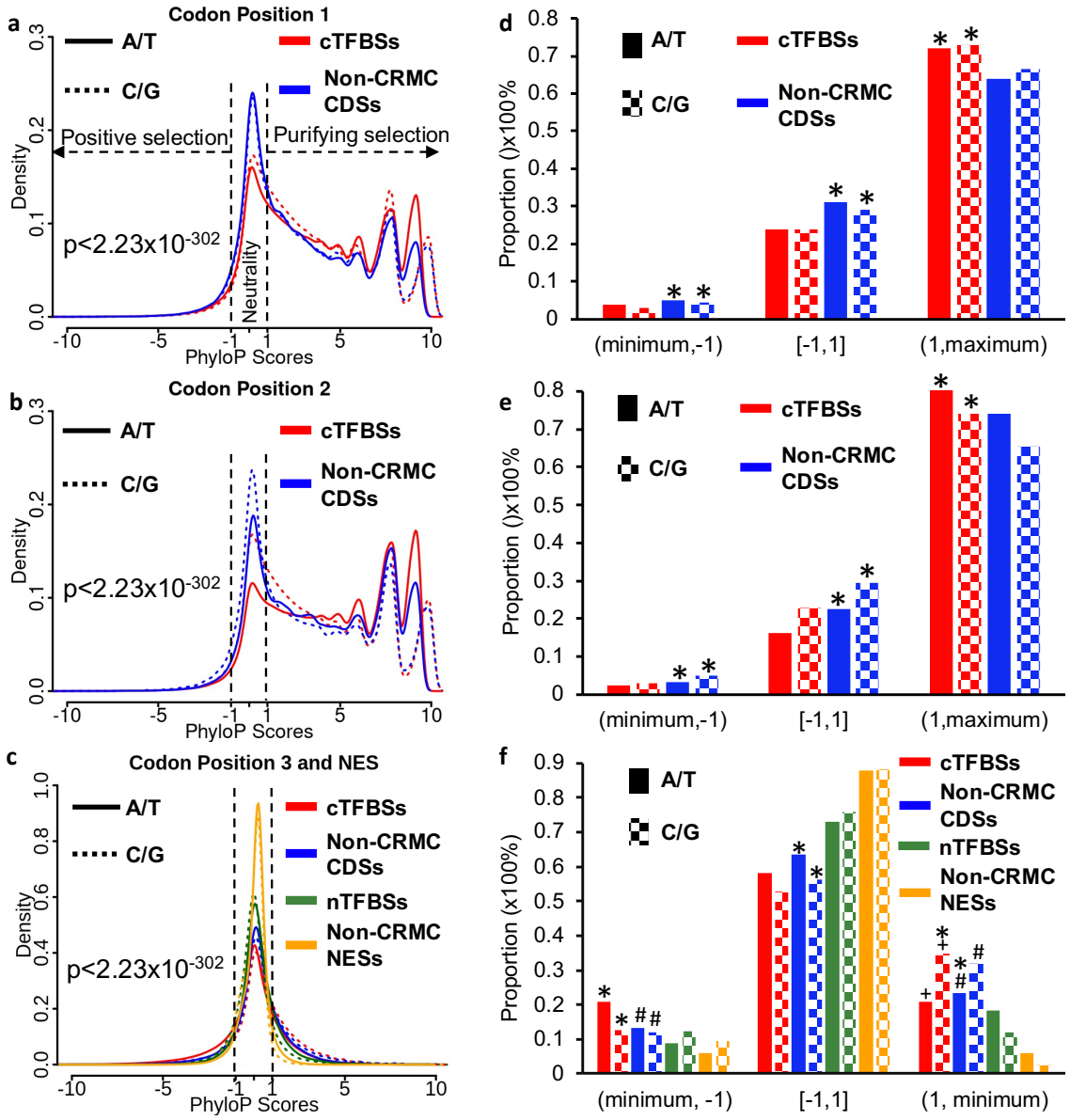

Figure 4. Comparison of phyloP scores of the predicted TFBSs in different types of sequences. **a-c**: Distributions of phyloP scores of A/T and C/G at the first nondegenerate codon positions in cTFBSs and non-CRMC CDSs (**a**), at the second nondegenerate codon positions in cTFBSs and non-CRMC CDSs (**b**), and at the degenerate third codon positions, in comparison with those in nTFBSs and non-CRMC NESs. Comparisons of density curves were done using the K-S test. **d-f:** Proportions of A/T and C/G under positive selection [(minimum, -1)], purifying selection [(1, maximum)], or evolutionarily neutral [[-1, 1]] at the first nondegenerate codon positions in cTFBSs and non-CRMC CDSs (**d**), at the second nondegenerate codon positions in cTFBSs and non-CRMC CDSs (**e**), and at the degenerate third codon positions in cTFBSs and non-CRMC CDSs, in comparison with those in nTFBSs and non-CRMC NESs (**f**). *$p < 2.2\times10^{-16}$, comparisons between degenerate positions in cTFBSs and in non-CRMC CDSs using the χ^2 test. +$p < 2.2\times10^{-16}$, comparisons between degenerate positions in cTFBSs and nTFBSs using the χ^2 test. #$p < 2.2\times10^{-16}$, comparisons between degenerate positions in non-CRMC CDSs and NESs using the χ^2 test

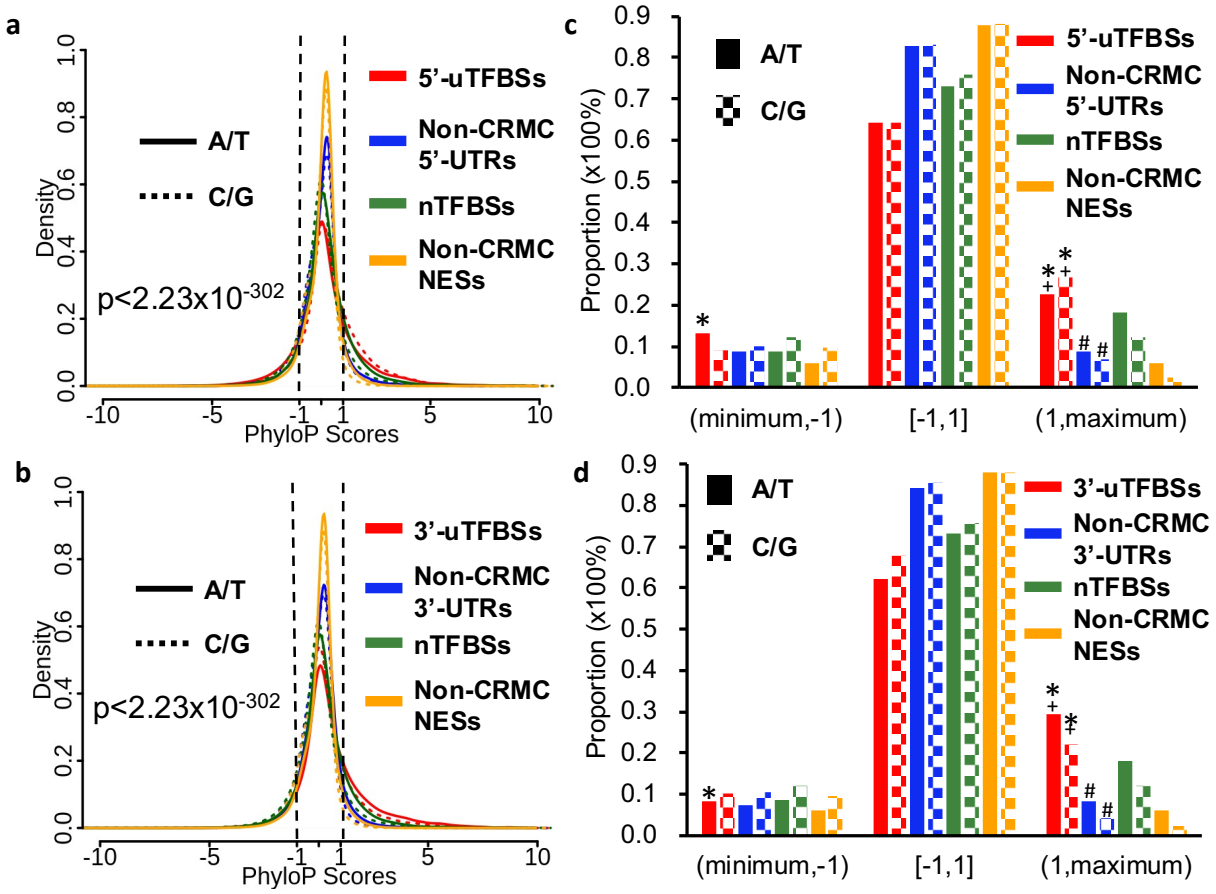

Figure 5. Comparison of phyloP scores of the predicted uTFBSs, non-CRMC UTRs, nTFBSs and non-CRMC NESs. **a-b:** Distributions of phyloP scores of A/T and C/G positions in 5'-uTFBSs (**a**), 3'-uTFBSs (**b**) and their non-CRMCs counterparts, in comparison with those in nTFBSs and non-CRMC NESs. Comparisons of density curves were done using the K-S test. **c-d**: Proportions of the positions that are under positive selection [(minimum, -1)], purifying selection [(1, maximum)], or selectively neutral [[-1, 1])], in 5'-UTRs (**c**) and 3'-UTRs (**d**) and their non-CRMC counterparts, in comparison with those in nTFBSs and non-CRMC NESs. *p < 2.2×10^{-16}, comparisons between uTFBSs and non-CRMC UTRs using the χ^2 test. +p < 2.2×10^{-16}, comparisons between uTFBSs and nTFBSs using the χ^2 test. #p < 2.2×10^{-16}, comparisons between non-CRMC UTRs and NESs using the χ^2 test.

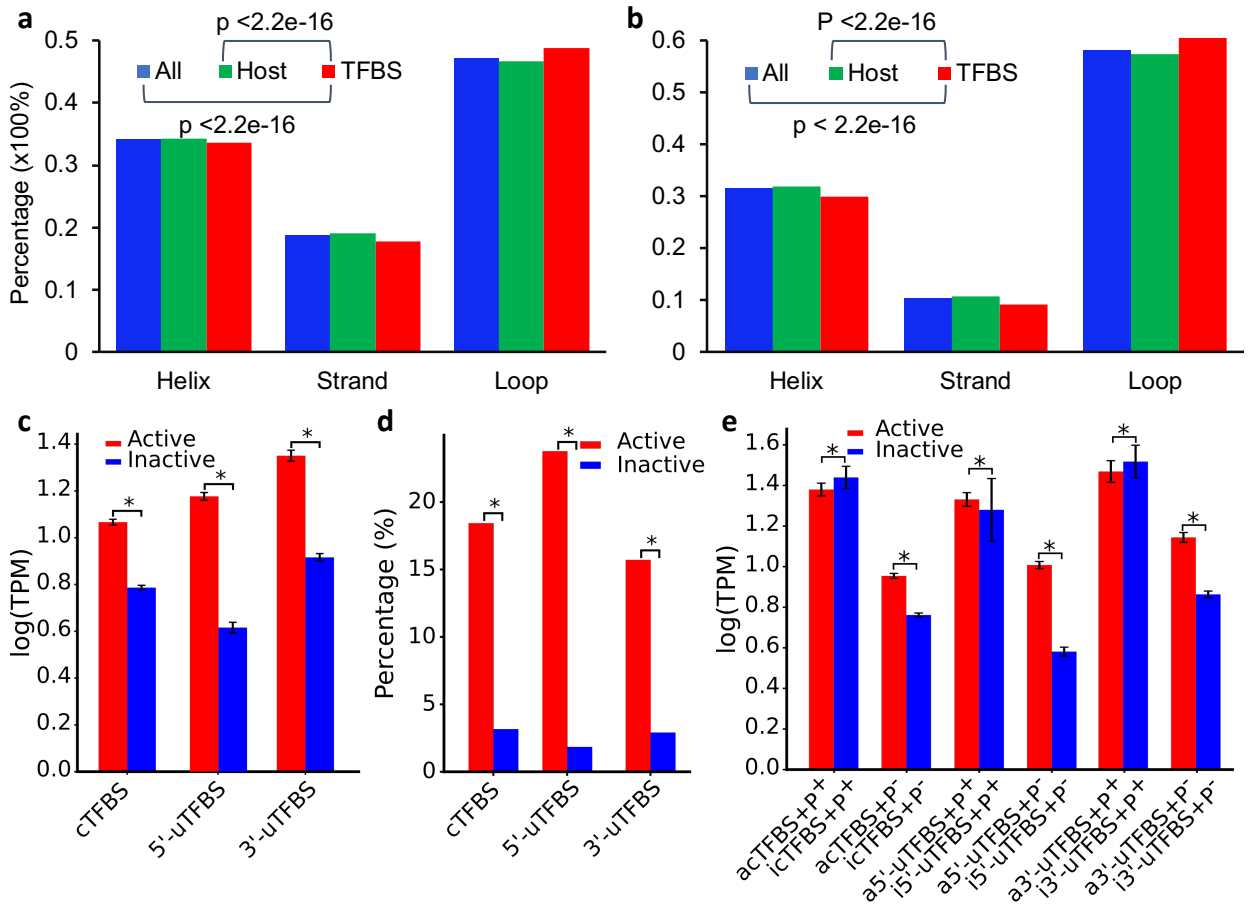

Figure 6. Preference of eTFBSs for secondary structural types in proteins and in physical proximity to distal promoters. **a.** Proportions of cTFBS-encoded amino acids mapped to known loops, helices, and strands in comparison with those in the host proteins and in all known non-redundant protein structures. P-values were computed using the χ^2 test**. b.** Proportions of cTFBS-encoded amino acids mapped to the predicted loops, helices, and strands in comparison with those in the host proteins and in all predicted secondary structures. **c.** Transcription levels of closest genes to active eTFBSs in comparison with those of closest genes to inactive eTFBSs in K562 cells. **d**) Percentage of active eTFBSs and inactive eTFBSs that interact with non-own promoters in K562 cells. **e.** Transcription levels in K562 cells of genes 1) closest to promotors that are in close proximity to active eTFBSs (aeTFBSs+P^+), 2) closest to promotors that are in close proximity to inactive eTFBSs (ieTFBSs+P^+), 3) closest to active eTFBSs that are not in close proximity to promoters (aeTFBSs+P^-), and 4) closest to inactive eTFBSs that are not in close proximity to promoters (ieTFBSs+P^-). * The data in c, d and e are shown in mean±SE. $p<2.23\times10^{-302}$. P-values were computed using the χ^2 test (a, b, d) Mann Whitney U test (c, e).