Genome-Wide Analysis of Functional and Evolutional Features of Tele-Enhancers

Di Huang * and Ivan Ovcharenko *.

* Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA

1 Corresponding author: 8600 Rockville Pike, Building 38A, Room 6S602, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA. E-mail: ovcharen@nih.gov
Running title
Function and evolution of *tele*-enhancers

Key words
enhancer, nucleotide divergence, single-nucleotide polymorphism, tissue specificity, transcription factor binding motif

Corresponding author
Ivan Ovcharenko
8600 Rockville Pike, Building 38A, Room 6S602, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20892, USA. E-mail: ovcharen@nih.gov
ABSTRACT

We investigated discriminating features between enhancers separated from their target genes by at least one intermediate gene/exon (named tele-enhancers in this study) and enhancers residing inside the target gene locus. In this study, we used genome-scale enhancer maps and whole-genome gene expression profiles to establish a large panel of heart tele-enhancers. By contrasting tele-enhancers to proximal heart enhancers located in the neighborhood of heart genes, we observed that heart tele-enhancers employ unique regulatory mechanisms based on cardiac transcription factor SRF, TEAD, and NKX-2.5, while proximal heart enhancers rely on GATA4 factors. A functional analysis shows that tele-enhancers preferentially regulate house-keeping genes and play a metabolic role during heart development. In addition, with 1) lower nucleotide divergence, 2) lower single-nucleotide polymorphism (SNP) density, and 3) smaller proportion of low derived-allele-frequency SNPs, tele-enhancers are significantly more conserved than their proximal counterparts. Similar trends have been observed in non-heart tissues and cell types, suggesting that our findings are general characteristics of tele-enhancers.

INTRODUCTION

Enhancers, which are key to the precise regulation of spatiotemporal-specific gene expression patterns, often reside at a distance from their target genes (Maston et al. 2006) and function through long-range regulatory mechanisms (GLINSKII et al. 2011; LETTICE et al. 2003; POMERANTZ et al. 2009). While some enhancers are found in the proximity of transcription start sites of the target genes, there is mounting evidence of
distant enhancers incorporated into the structure of neighboring genes or looping over intermediate unaffected genes (IRIMIA et al. 2012; LETTICE et al. 2003; PERRY et al. 2010; VISSER et al. 2012). As such, identifying the target genes of enhancers, especially enhancers that regulate genes outside of the locus they reside in (which we dubbed tele-enhancers), poses a great challenge in the research area of gene regulation. To address this challenge, many have employed evolutionary conservation of DNA sequences and focused on highly conserved enhancers (CLARKE et al. 2012; DAVIDSON et al. 2006; ENGSTRÖM et al. 2007; KIKUTA et al. 2007). Assuming that enhancers and their target genes are located within a genomic regulatory block encompassing a set of neighboring conserved genes and regulatory regions, regulatory relations between highly conserved enhancers and their target genes have been established, and transgenic models have been used to validate these regulatory relations (DONG et al. 2009; KIKUTA et al. 2007; MACKENZIE et al. 2004; NAVRATILOVA and BECKER 2009; SANYAL et al. 2012). These studies have successfully detected enhancers and their distant target genes, which are instrumental for understanding the mechanisms and evolution of gene regulation. However, these investigations focused on individual tele-enhancers. Recently, with the advancement of sequencing techniques, such as chromatin conformation capture (3C, Hi-C) (BELTON et al. 2012; DIXON et al. 2012) and chromatin interaction paired-end tagging (ChIA-PET) (LI et al. 2010; ZHANG et al. 2012), it became possible to experimentally characterize interactions between enhancers and their target genes. However, these experiments are notoriously difficult to run, and genome-wide data are available only in a few cell types. As a result, without the knowledge of a substantial number of tele-enhancers, we still have a scant genome-
wide understanding of interactions between long-range enhancers and their target genes.

In this study, we binned heart enhancers into two groups—proximal and tele-enhancer—based on their genomic location relative to the nearest heart gene, and then systematically investigated the differences between proximal and distant regulatory mechanisms of transcriptional activation in the human heart. We observed that tele-enhancers have unique biological functions—regulating housekeeping genes and playing a role in basic biological processes. Next, we demonstrated that tele-enhancer sequences feature a distinguishable binding motif signature. While both tele and proximal heart enhancers are enriched for the binding motif of cardiac transcription factors (TFs) (such as MEF2A), tele-enhancer sequences are enriched for TEAD1 and NKX-2.5 binding motifs, and are depleted of the binding motifs of the GATA4 as compared to proximal enhancers. We also demonstrated that tele-enhancers 1) are significantly more conserved than proximal enhancers, 2) display lower single-nucleotide polymorphism (SNP) density, and 3) are enriched in low derived allele frequency SNPs—all suggesting that tele-enhancers have been and are currently evolving under a stronger negative selective pressure than their proximal counterparts.

We then extended our analysis to other tissues and cell types (including fetal brain, fetal lung, HepG2, K526 HSMM, and H1-hESC cell lines), for which genome-wide enhancer maps were identified using different experimental methods, and observed similar evolutionary trends as well as sequence and functional features of tele-enhancers, suggestive of the general trends of this novel regulatory mechanism described by our study.
MATERIALS AND METHODS

Heart enhancers and genes

We used the set of 5,047 human heart enhancers identified by a p300 ChIP-Seq study on fetal heart tissue (gestational week 16; GEO data set GSE 32587) (May et al. 2012). We assigned an enhancer to the two flanking heart genes located within 500 kilobases from the enhancer. As a result, an enhancer was assigned to at most two heart genes. After discarding enhancers not associated with any heart gene and heart genes not associated with any enhancers, we obtained 3,391 enhancers linked to 1,832 heart genes. We dubbed enhancers separated from their target genes by at least one non-heart gene or exon as tele-enhancers, and the enhancers located within loci of heart genes as proximal enhancers (Figure 1).

We also generated a set of control sequences. For each enhancer, two random non-coding human DNA sequences with matching length, GC content, and repeat density were retrieved.

Enhancers and genes in other tissues and cell lines

Fetal brain enhancers were obtained from the H3K4me1 map established by ChIP-seq experiments carried out on human fetal brain tissue (GEO dataset GSM706850, human fetal brain at day 122). To ensure that the nucleotide sequences used in this study are enhancers and not proximal promoters, we discarded sequences located within 3 kb of the transcription start site of any human gene. To obtain brain genes, we collected the genes annotated to the “brain development” or its child categories in the Gene Ontology (GO) database. Additionally, we utilized gene expression profiles (Su et
and identified the top 20% of genes highly expressed in fetal brain with respect to other tissues and cell lines. After applying our method to these data, we assigned 7649 brain enhancers to 1,957 brain genes (Table S1).

Fetal lung enhancers were also obtained from the H3K4me1 map established by ChIP-seq experiments carried out on human fetal lung tissue (GEO dataset GSM706853, human fetal lung at day 101). We similarly discarded sequences located within 3 kb of any transcription start site. To the end, we linked 5996 lung enhancers to 1716 lung genes (Table S1).

In cell lines including GM12878, H1-hESC, HepG2, HSMM, HUVEC, K562, and NHEK, we used ChromHMM enhancer maps (Djebali et al. 2012; Ernst and Kellis 2012). Also, based on RNA-seq data reported in the ENCODE project (Djebali et al. 2012)), we marked the top 20% of genes highly expressed in a cell with respect to other cells as cell-specific genes. After applying our pipeline to these data (i.e., cell-specific enhancers and genes), we identified ~6000 *tele*-enhancers in each line (Table S2).

**Evaluation of enhancer-gene association**

Enhancer-promoter maps have been previously constructed using the distribution of DNase I hypersensitive sites (DHSs) across 79 distinct cell types, and the reported regulatory connections have been confirmed using 5C and ChIP-PET data (Thurman et al. 2012).

Although the distal-DHS-promoter connection map was established based on an extensive panel of cells, this map does not cover the entire regulatory-element-promoter map for all tissues and cell lines. For example, 1,538 (out of 3,391) heart enhancers
assigned to at least one heart gene could not be retrieved in this distal-DHS-promoter connection map. As such, we built the regulatory block for each gene based on distal-DHS-promoter connection map (denoted as DHS-based regulatory block in this study). Given a gene, the genomic boundaries of its DHS-based regulatory block are defined as the most distal DHSs connected to that gene. Since the enhancers located within this block more likely regulate that gene than those located outside of this block, we evaluated whether enhancers reside within the DHS-based regulatory blocks of their associated genes. We estimated the fraction of the enhancers that fell into the regulatory blocks of their associated genes, and compared this fraction with the null distribution. The null distribution was established by randomly selecting genes. In detail, given an enhancer and its associated gene (the distance between them is D), we randomly selected a gene with the regulatory block of a similar length to the tested gene (i.e., the length of the regulatory block of randomly selected genes was required to reside in the range of [0.9z, 1.1z] where z is the length of the regulatory block of the tested gene). Then we checked if the sequence located to the selected gene in a distance of D fell into the regulatory block of the selected genes. For each enhancer, we repeated this process 20 times, and used the fraction of the sequences residing in the respective regulatory blocks as expected.

**Evolutionary synteny blocks and density of evolutionary breakpoints**

Evolutionary synteny blocks (ESBs) are commonly used to identify boundaries of regulatory interactions, as regulatory elements and their target genes tend to reside within the same synteny block, if the regulatory mechanisms in question play an
important functional role. Accordingly, the density of synteny breakpoints should be reduced between regulatory elements and their target genes.

We downloaded ESBs from the ECRBase database (OVCHARENKO et al. 2005b) and evaluated the breakpoint density in the regions located between heart enhancers and the target genes assigned to them. After that, the density of synteny breakpoints was compared with expectation. Since the evolutionary breakpoints are not evenly distributed along genome (CARVER and STUBBS 1997; PEVZNER and TESLER 2003), we estimated the expected density of breakpoints “locally” instead of “globally.” That is, for each region spanning an enhancer and one of its target genes, we extended this region by 500 kilobases along both the upstream and downstream direction, and then excluded gene coding regions from the extended sequence. We used these extended regions as controls to estimate the expectation of breakpoint density.

**Control genes**

To eliminate the potential effect of loci to the function of genes, we generated control genes with a similar locus length to a given gene set. For a gene, we randomly selected five genes having a similar locus length. After repeating this process for all genes, we generated a control set for a tested gene set.

**SVM classification models and enriched binding motifs in enhancer sequences**

To discriminate enhancer sequences from controls, we designed a computation system that included two main steps. In the first step, putative TFBSs were mapped in the DNA sequence of enhancers and controls. For this, sequences were scanned using
tfSearch (OVCHARENKO et al. 2005a) with the position weight matrices (PWMs) from the TRANSFAC and JASPAR databases (SANDELIN et al. 2004; WINGENDER et al. 2001). In the second step, TFBS vectors of enhancers and controls were used to build a linear support vector machine (SVM) (CORTES and VAPNIK 1995) to discriminate between enhancers and controls. Given a training set of instances \( \{x_1, x_2, ..., x_n\} \) with associated labels \( \{y_1, y_2, ..., y_n\} \in \{-1, 1\} \), a linear SVM \( y = w^T x + b \) was built through solving the optimization problem \( \min \left( \frac{1}{2} w^T w + C \sum_i \varepsilon_i \right) \) subject to \( y_i (w^T x_i + b) \geq 1 - \varepsilon_i \) and \( \varepsilon_i \geq 0 \) (CORTES and VAPNIK 1995). In such linear SVM, a linear weight \( w_i \) is assigned to each TF binding motif. A large positive \( w_i \) indicates a binding motif that is strongly associated with the enhancers, whereas negative weights mark the binding motifs associated with the control set. We ranked motifs according to \( w_i \), i.e., the positive association with the tested enhancer set.

**Functional analysis based on GO annotation**

The enrichment of a GO functional category (\( C_i \)) for a group of genes (\( G \)) was measured according to

\[
Pr_i = \sum_{0<k<m} \frac{M \choose k}{N \choose n-k} \frac{(N-M) \choose (n-k)}{N \choose n},
\]

where \( m \) is the size of the overlap between \( C_i \) and \( G \). \( M \) and \( N \) are the size of \( C_i \) and all genes, respectively. \( n \) is the size of \( G \). To control the false positive rate, we adopted the conservative Bonferroni multiple-testing correction strategy (ABDI 2007) to adjust \( Pr_i \) as \( aPr_i = K \times Pr_i \) where \( K \) is the number of GO functional categories for the test.
**Nucleotide divergence and diversity**

To evaluate nucleotide divergence, we aligned each enhancer sequence to chimpanzee and rhesus by using axt alignment files available from the UCSC Genome Browser (http://genome.ucsc.edu/index.html).

Nucleotide difference ($d$) of a DNA region is defined as the number of nucleotide which are different across species per site. According to the three-way alignment among human, chimpanzee and macaque, we evaluated $d$ specific to each species. Given a DNA region, $d$ specific to human ($dh$) is measured as

$$dh = \frac{L_{h≠(c=r)}}{La},$$

where $La$ is the length of three-way-alignable fragment along the given region, and $L_{h≠(c=r)}$ is the number of aligned nucleotides where chimpanzee is equal to macaque, and human is the outlier with respect to chimpanzee and macaque. Nucleotide divergence specific to human ($Dh$) is then estimated by adjusting $dh$ with Juke-Cantor correction for multiple hits (JUKES and CANTOR 1969). In a similar way, the nucleotide divergence $D$ specific to chimp ($Dc$) and specific to macaque ($Dm$) were measured.

To evaluate selection neutrality within the human lineage, we defined the neutrality index (NI) in which human-specific nucleotide divergence (i.e., $Dh$) of enhancers is compared with non-human divergence (i.e., $Dr + Dc$) with respect to the neutral reference.

$$\text{Neutrality Index (NI)} = \frac{Dh/Nh}{(Dr+Dc)/(Nr+Nc)},$$

where $Nh$, $Nr$, and $Nc$ are the divergences of neutral references specific to human, rhesus, and chimp. A large NI means a large nucleotide difference during the evaluation of human (Figure 5). Also, based on $Dh$, $Dr + Dc$, $Nh$ and $Nr + Nc$, we carried out the
McDonald-Kreitman test (MK test) (McDONALD and KREITMAN 1991) to estimate the significance of nucleotide divergence.

Nucleotide diversity ($\pi$) of a region is estimated as the number of single-nucleotide polymorphisms $s$ per site in the alignable fraction of a region.

**Gene regulating mitochondrial biological processes**

After going through the Gene Ontology (GO) database (ASHBURNER et al. 2000), we collected 282 genes annotated to a mitochondrial biological process.

**Pseudogenes**

Pseudogenes, dysfunctional gene homologues (VANIN 1985), were used as a neutral reference in this study. We downloaded pseudogenes from the Pseudogene.org database (BALASUBRAMANIAN et al. 2009).

**HACNS**

HACNSs, human-accelerated conserved noncoding sequences, used in this study were from (PRABHAKAR et al. 2006)

**RESULTS**

**Identification of tele-heart enhancers and their target genes**

In this study, we employed a conservative definition of tele-enhancers, requiring a presence of an intermediate gene (at least one exon of an intermediate gene) between an enhancer and its target gene. To identify heart tele-enhancers, we first compiled a
list of developmental heart genes according to gene expression profiles and gene annotation information. After ranking all human genes according to their expression level in human heart relative to the other 78 tissues/cell types (SU et al. 2004), we selected the top 20% (2,342) of genes and observed that these genes are significantly enriched 1) in the heart development genes annotated in the Gene Ontology (GO) resource (Ashburner, Ball et al. 2000) (p-value = $2 \times 10^{-11}$, using DAVID (JIAO et al. 2012)), 2) in the neighborhood of p300 heart enhancers (hypergeometric distribution p-value = $4 \times 10^{-34}$), and 3) in the heart disease genes reported by the GeneTests database (p-value = $4 \times 10^{-13}$, Figure S1, see Supplementary Materials), suggesting a significant association between these highly-expressed genes and human heart development (see Materials and Methods). In addition, we combined these highly-expressed genes with heart development genes annotated in GO (total = 348 genes), and then established a collection of 2430 distinct heart genes. The genomic landscape of these heart genes was superimposed onto 3391 heart development enhancers that were previously identified in a p300 ChIP-seq fetal human heart tissue experiment (BLOW et al. 2010), and were located within no more than 500 kilobases from the transcription start site of a heart gene. Following a general rule postulated in 2012 by Busser et. al. (BUSSER et al. 2012), we associated each heart enhancer with the two closest heart genes—one upstream and the other downstream of the enhancer. Enhancers separated from their nearest heart gene by at least one “non-heart” gene/exon were considered as tele-enhancers (Figure 1a). In total, 3319 heart enhancers were associated with 1832 heart genes, out of which a large proportion of genes—1171 (64%)—had no proximal enhancers located on either the inside of the
gene body, or in between the gene and its two flanking genes. These findings are in line with the report that proximity is not the governing rule for enhancers regulating genes, and less than 50% of expressed genes have proximal enhancers in several cell types (including embryonic stem cell) (Zhang et al. 2013). We hypothesized that the heart genes lacking proximal enhancers were likely being regulated by tele-enhancers (Figure 1a). For simplicity, we dubbed a heart gene with at least one proximal heart enhancer GeneP, and a heart gene linked to tele-enhancer(s) only—GeneT. While genes lacking proximal enhancers may be controlled through activation of regulatory elements other than enhancers, such as locus control regions (Li et al. 2002), most genome-wide gene regulation studies are confined to promoters and enhancers for simplicity and generalization (Gaszner and Felsenfeld 2006). Therefore, we assumed that GeneTs are primarily regulated by tele-enhancers. Also, before systemically characterizing tele and proximal enhancers, we evaluated the reliability of the established assignments between genes and enhancers.

**Regulatory landscape of heart genes**

We used evolutionary synteny blocks (ESBs) to confirm regulatory relationships between the tele-enhancers and the genes they were assigned to. ESBs are a known characteristic of long-range regulatory interactions as the separation of critical regulatory elements from their target genes through chromosomal rearrangements is selected against (Ahituv et al. 2005; Mackenzie et al. 2004); ESBs have been successfully used in the past to detect target genes of long-range enhancers (Dong et al. 2009; Kikuta et al. 2007; Navratilova and Becker 2009). Accordingly, we expected
tele-enhancers and their target genes to reside within the same ESB more often than expected by chance. We utilized the defined ESBs based on the sequence alignments between human and chimpanzee, macaque, mouse, cow, and chicken (Ovcharenko et al. 2005b), and compared the density of the evolutionary break points between enhancers and their target genes with the density expected in the neighborhood of enhancers (see Materials and Methods). Our comparative results showed that the genomic space separating enhancers and their target genes exhibited a significantly lower density of evolutionary breakpoints (p-values < 10^{-5} in all cases, Figure 2a) than would be expected. As such, the enhancers, either proximal or tele-enhancers, and their associated genes were predominantly located in the same ESBs.

Additional validation of the predicted enhancer-gene relationships was obtained using a genome-wide map of enhancer-promoter associations constructed based on a DNase I comparative profiling of the human genome (Thurman et al. 2012). Although the reported DNase I map of enhancer-promoter relationships is an approximation across a large panel of cell types, which does not necessarily represent a comprehensive collection of long-range regulatory activities in the heart, it estimates the extent of regulatory interactions in gene loci. After defining a DHS-based regulatory block for each gene based on its most distal enhancers, we first noticed that GeneTs showed longer DHS-based regulatory blocks as compared to GenePs (Wilcoxon rank sum p-value = 0.06, Figure 2b). Also, we computed the fraction of the enhancers located within the regulatory block of their associated genes, and demonstrated that 82% of tele-enhancers reside within the regulatory blocks of their associated genes, which was significantly higher than expected (binomial test p-value = 3 \times 10^{-15}, Figure
Similarly, 83% of proximal enhancers were located within the regulatory blocks of their associated genes, which also was significantly higher than expected (binomial test p-value = $3 \times 10^{-11}$, Figure 2c). These results further support the established regulatory relationship between the enhancers, either proximal or tele-enhancers, and their associated genes.

Next, we examined the size of intronic and intergenic regions of heart gene loci and noticed that, on average, GeneTs featured 3.0-fold shorter intronic regions than GenePs (Wilcoxon rank sum p-value = $2 \times 10^{-26}$, Figure 2d). Similarly, the intergenic intervals flanking GeneTs were 2.78-fold shorter compared to GenePs (Wilcoxon rank sum p-value = $6 \times 10^{-64}$, Figure 2d). It is likely that the small locus size of GeneTs predisposes these genes to the acquisition of tele-enhancers. Given the recent evolutionary nature of many heart enhancers (BLOW et al. 2010), an alternative hypothesis of an intermediate gene insertion in between a heart enhancer and a GeneT is unlikely. From the evolutionary viewpoint, it has been found that the maintenance of gene function over a long evolutionary time leaves a selection signature of gene structure (VINOGRADOV 2006), and that tissue-specific genes might harbor long non-coding regions containing multiple regulatory regions, while widely-expressed genes, lacking strong intron constraints, might have been subjected to selective pressure to reduce the length of non-coding regions (EISENBERG and LEVANON 2003; POZZOLI et al. 2007). With this knowledge, and based on our finding that GeneTs showed significantly shorter intronic and intergenic spread than GenePs, we hypothesized that tele-enhancers and proximal enhancers, which respectively regulate GeneTs and GenePs, have different biological functions and undergo different evolutionary processes. To explore this hypothesis, we
next examined the functional and evolutionary features of tele-enhancers and proximal heart enhancers.

**Tele-enhancers regulate mitochondrial biological processes**

We analyzed the function of the two heart enhancer groups—proximal enhancers and tele-enhancers—according to the Gene Ontology function of their target genes (GenePs and GeneTs, respectively) (Ashburner et al. 2000). In this study, to account for the different locus lengths of GenePs and GeneTs (as discussed above), which may cause a bias in a gene function analysis, we generated controls for GenePs and GeneTs separately by randomly selecting genes with similar length intergenic and intronic regions to the tested genes (see Materials and Methods), and compared GenePs/GeneTs with the respective control genes. As expected, both proximal enhancers and tele-enhancers played an important role in biological processes related to heart development, such as heart morphogenesis, cardiac muscle differentiation, etc. (multiple testing corrected binomial test p-value < 1 \times 10^{-3}, Figure 3a and Table S3). On the other hand, each heart enhancer group featured distinct functions—proximal enhancers were strongly associated with various processes related to heart development, while tele-enhancers were likely involved in basic metabolic functions. For example, out of 18 genes associated with ventricular cardiac muscle cell differentiation (GO:0055015), 11 (79%, p-value = 4 \times 10^{-4}) genes were GeneP, while only 3 (21%) were GeneT. Similarly, among 27 genes that fell into the category of artery morphogenesis (GO:0048844), 13 genes (p-value = 1 \times 10^{-7}) had proximal enhancers, while 7 genes were GeneT (Table S3). By contrast, 73 precursor metabolites and
energy (GO:0006091) genes were categorized as GeneT, while 28 GenePs played this function. We also collected the genes taking part in the regulation of mitochondrial biological processes, and observed that GeneTs, but not GenePs, were significantly enriched for those genes (Figure 3b, 2.9% of GenePs vs 5.1% of GeneTs, binomial test p-value $= 2 \times 10^{-5}$). Our observation is supported by reports that regulation of mitochondrial processes is essential to heart development (DRENCKHAHN 2011; GOFFART et al. 2004) and has been used as a therapeutic target in heart failure (HUSS and KELLY 2005). Moreover, while both GeneTs and GenePs showed significant heart specificity as compared to the expected (i.e., respective control genes, Figure S2), GeneTs had significantly weaker heart specificity than GenePs (Figure 3c, the average of GeneTs and GenePs are 1.7 and 1.9, respectively, Wilcoxon rank-sum test p-value $= 7 \times 10^{-3}$). This further indicates that, compared with GenePs, GeneTs more likely partake in basic biological processes and display lower heart specificity.

We also looked into the function of bystander genes of tele-enhancers, i.e., the genes which are located closer to tele-enhancers than the target genes of these enhancers, but are not regulated by these enhancers during heart development. The functional Gene Ontology analysis indicated that these bystander genes were not significantly enriched for any biological function (data not shown), which further supports the established connections between tele-enhancers and their target genes.

Collectively, tele-enhancers and proximal enhancers have different biological functions. The former partake in basic biological processes, regulating mitochondrial biological processes, while the latter are more specific to heart development. Since tele-enhancers showed functions distinct from proximal enhancers, we hypothesized that
tele-enhancers and proximal enhancers are involved in different transcriptional mechanisms and are activated by different sets of TFs.

**Tele-enhancers feature distinguishable binding motif compositions**

We modified a machine learning approach that has been previously used to identify motifs specific to heart enhancers (NARLIKAR et al. 2010). This method identifies specific sequence patterns for a set of non-coding sequences relying primarily on known TF binding motifs from the TRANSFAC and JASPAR databases (MATYS et al. 2006; STORMO 2000). After mapping 981 vertebrate TF binding motifs onto enhancers and controls with similar GC content, repeat density and sequence length, we built two support vector machine (SVM) classifiers with linear kernels based on the occurrence of TF binding motifs (see Materials and methods)—one for tele-enhancers and another for proximal enhancers. In a linear SVM, TF motifs strongly associated with training enhancers received large positive weights. We selected the motifs with positive weights in either classifier and clustered these motifs based on the SVM weights, i.e., the association with enhancer classes (Figure 4a). Only 26% (30 out of 117) of TF binding motifs, in which cardiac TFs such as MEF2A were included, were shared between tele and proximal enhancers. Tele-enhancers featured positive association with the binding motifs of well-known cardiac TFs SMAD1, SRF, NKX-2.5, and TEAD and no association with the binding motifs of cardiac TFs GATA4, which were specific to proximal enhancers (Figure 4a and Table S4).

We also investigated the enrichment of TF binding motifs in tele and proximal enhancers and observed a striking difference in their motif composition (Figure 4b and
Table S5). For example, the binding motifs of NKX-2.5, TBX5, and TEAD were strongly enriched in *tele*-enhancers but not in proximal enhancers, while the binding motifs of NFAT4 and GATA4 were specific to proximal enhancers. These results indicate a set of cardiac TFs are needed for the activation of both proximal enhancers and *tele*-enhancers, and a specific transcriptional modulation by different cardiac TFs within these two groups of enhancers. The differences in transcriptional mechanisms employed by proximal enhancers and *tele*-enhancers can potentially explain the difference in the biological function and expression patterns driven by these two groups of enhancers. The slight discrepancy between TF binding motif enrichment and the weights in linear SVMs (for example, the binding motif of TBX5 was exclusively enriched in *tele* enhancer sequences while this motif was not assigned with a large weight in either *tele* or proximal enhancer SVMs) may be explained by the fact that enrichment evaluates TF binding motifs individually, while a linear SVM model estimates binding motifs collectively. Also, a high enrichment fold does not necessarily indicate that the abundance of a binding motif in enhancers is sufficient to distinguish enhancers from the rest of the non-coding genome. For example, the binding motif of TBX5, although exhibiting relatively high enrichment fold (1.3) in *tele*-enhancers, had a relatively low occurrence (0.03 per 1000 bp), which led to a small linear weight assigned to this motif in our SVM models.

**Different selective signatures of *tele*-enhancers and proximal enhancers**

It is known that the regulatory elements sharing the same cellular function and being activated in the same biological pathway evolve in concert and tend to have correlated
selective signatures (SHAPIRO and ALM 2008). We next analyzed the selective constraints imposed on heart enhancer groups. While developmental enhancers, as a whole, are under strong evolutionary constraint, the selective pressure imposed on enhancers varies greatly—phastCons, a conservation score, ranges from 0 (indicating non conservation) to 1 (perfect conservation) (BLOW et al. 2010; CLARKE et al. 2012). To analyze the selective constraint of enhancers, we used phastCons derived from 46 placental mammal sequence alignments (SIEPEL et al. 2005) and assigned the average phastCons along the tested sequence to that enhancer. Compared with proximal enhancers, tele-enhancers were more conserved (Figure S3): 15% of tele-enhancers had a phastCons greater than 0.2, while 13% of proximal enhancers reached this conservation level (binomial test, p-value = 7 \times 10^{-3}, tele vs proximal enhancers).

We next evaluated the selective pressure acting on enhancer sequences within the human lineage. After generating human-chimpanzee-macaque three-way alignments, we estimated the nucleotide divergence of enhancer sequence between any two species and evaluated the human-specific and non-human-specific divergence (Figure 5). The divergence rate of enhancer sequences was compared with the neutral reference divergence rate computed using pseudogenes. The comparative results showed that enhancers, either proximal or remote, had remarkably lower divergence levels than the control sequences (which were randomly generated along non-coding DNA with matched repeat density and GC content and the same length as enhancers) and human accelerated conserved noncoding sequences (HACNS) (Figure 6a and Table 1). Next, we evaluated the selection constraints acting on sequences along the human lineage using the neutrality index (NI). NI is defined in such a way that low NI
(<1) and high NI (>1) indicate negative and positive selection during the human lineage evolution, respectively (Figure 5). According to the NI estimates, while both proximal enhancers and tele-enhancers featured negative selection, tele-enhancers with the average NI of 0.75 evolved under stronger negative selection constraints than proximal enhancers with the average NI of 0.77 (Table 1). Also, we evaluated the patterns of selective constraints in proximal enhancers and tele-enhancers using the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991). Compared to the neutral reference, both proximal and tele-enhancers evolved under a significant negative selection pressure (p-values < 6 × 10^{-23}, Fisher’s exact test, Table 1). Moreover, the negative selection pressure acting on tele-enhancers was significantly stronger as compared to proximal enhancers (p-value = 2 × 10^{-2}, Fisher’s exact test).

Next, we applied the MK test to evaluate the selective constraints of individual enhancers, and found that, compared to control sequences, both proximal enhancers and tele-enhancers tended to be under strong negative constraints, showing less cases of positive selection and more cases of purifying selection than controls (Figure 6b). On the other hand, compared to proximal enhancers, fewer tele-enhancers were under positive selection. For example, with the p-value cutoff of 1 × 10^{-4}, proximal enhancers were more likely to be under positive selective pressure than tele-enhancers—2.5% proximal and 1.5% tele-enhancers evolved under positive selection (binomial test p-value = 4 × 10^{-3}, proximal vs tele-enhancers). In summary, the nucleotide divergence across different species indicates that tele-enhancers are under stronger negative selection than proximal enhancers.
Due to the difficulties in aligning sequences, nucleotide divergence is not as straightforward as SNPs in assessing the signature of selection of DNA regions. We therefore utilized SNP and allele frequency to evaluate the selective pressure acting on heart enhancers during modern human history. Using the genome variation data from the 1000 Genomes Project (CONSORTIUM 2010), we first observed that tele-enhancers harbor significantly less SNPs than proximal enhancers (binomial test p-value = $2 \times 10^{-3}$, Table 1). Since a shift towards lower derived allele frequency (DAF) indicates negative selection in the modern human history (GOODE et al. 2010), we also analyzed DAF distribution of SNPs in heart enhancers. The results demonstrated that 1) tele and proximal enhancers represent a 2.4% and 1.7% excess of low-DAF SNPs (i.e., SNPs having DAF<0.05) compared with neutral reference, respectively (binomial test, p-value = $4 \times 10^{-13}$ for tele-enhancers vs neutral reference, and p-value = $3 \times 10^{-6}$ for proximal enhancer vs neutral reference, Figure 6c); 2) tele-enhancers are enriched in low-DAF SNPs when compared to proximal counterpart (binomial test, p-value = $1 \times 10^{-2}$, Figure 6c).

Also, we further partitioned tele-enhancers into subgroups—intronic and intergenic tele-enhancers with a premise that genomic locations of sequences may partially determine their evolutionary patterns (HALLIGAN et al. 2004). We observed intronic tele-enhancers being under stronger selection constraint than intergenic tele-enhancers (Table S6), which is in accordance with the finding that functional intronic regions are under stronger selection pressure as compared to intergenic counterpart in mammals (DAVIDSON et al. 2009). Furthermore, we compared intronic and intergenic tele-enhancers with their proximal counterparts, and observed that tele-enhancers are
consistently more conserved than their proximal counterparts. Tele-enhancers 1) are enriched for conserved sequences (those with \( >0.20\) phastCons, Figure S3, where the weak significant p-value is partially due to the small sample pool used for statistical analysis); 2) show lower NIs; 3) have low SNP density; and 4) exhibit preference toward low-DAF SNPs (Table S6).

Finally, both the nucleotide divergence and SNP-based analyses lead to the same conclusion that while heart enhancers are under strong negative selection, the selection pressure acting on tele-enhancers (either intergenic or intronic) is even stronger than that on proximal counterparts. This is in accordance with the finding that many highly conserved enhancers are separated from their target genes by “bystander” genes (Akalin et al. 2009). Also, since tele and proximal enhancers share similar functions (for example, both these enhancer types play a role in heart development and heart morphogenesis), display common cardiac TF binding motifs (such as those of MEF2A), and exhibit common selective features (both of them are highly conserved as compared to controls), it could be expected that the evolutionary constraint difference between tele and proximal enhancers is only weakly significant (p-values are between 0.001 and 0.05) in almost all cases.

**Tele-enhancers from different tissues show consistent evolutionary and functional features**

We extended the study to other tissues, including fetal brain and lung, in an effort to analyze the generalizability of our results. In each tissue, we collected potential enhancers based on ChIP-seq experiments targeting H3K4me1, an enhancer-
associated histone mark, along with their target genes retrieved using gene expression profiles and Gene Ontology gene annotations. We observed that a large fraction of highly-expressed genes were GeneTs, i.e., the genes with no proximal enhancers. For example, 47% of brain genes and 50% of lung genes were GeneTs (Table S1).

In the case of lung and brain enhancers, similar to heart enhancers, tele-enhancers were strongly associated with the development of the corresponding tissue (multi-test corrected hypergeometric distribution p-values = 0, Tables 2 and 3 as well as S7 and S8). In addition, tele-enhancers were more strongly associated with basic processes than proximal enhancers (multi-test corrected hypergeometric distribution p-values < 1 × 10^{-3}). For example, among 23 genes regulating the response of nutrient level, which have influence on brain development (GEORGIEFF 2007), 6 genes (26%) were brain GeneTs, while none were GenePs. Also lung GeneTs, not GenePs, were significantly enriched for the genes taking part in GTP metabolic and catabolic process (those genes play an essential role in structural patterning during lung development (WAN et al. 2013)). Furthermore, GeneTs in brain displayed significantly lower relative expression than GenePs (the average of relative expressions of GeneTs and GenePs was 1.4 and 1.9; Wilcoxon rank sum test p-value = 3 × 10^{-22}). Similarly, in lung, the relative expression of GeneTs was significantly lower than that of GenePs (the average of relative expression levels of GeneTs and GenePs was 1.9 and 2.1, respectively; p-value = 2 × 10^{-4}). These suggest that GeneTs in brain and lung show relatively low tissue specificity in brain and lung, similar to GeneTs in heart.

The analysis of sequence divergence of these enhancers indicated that the enhancers activated in different tissues evolve under negative selection but at a
different degree of evolutionary constraint. With the lowest nucleotide divergence, brain enhancers were much more conserved than heart enhancers, which is consistent with previous reports (Bow et al. 2010) (Tables 1 and 4). Tele-enhancers exhibited significant lower human-specific divergence than proximal enhancers (Fisher’s exact test p-value = $5 \times 10^{-3}$ for brain, and p-value = $1 \times 10^{-2}$ for lung, Table 4), suggesting stronger negative constraints imposed on tele enhancers than on proximal counterparts across all tested tissues. Additionally, SNP-based results consistently showed that tele-enhancers harbor less SNPs than proximal enhancers across all the tested tissue. In heart and lung, the SNP density difference between tele-enhancers and proximal enhancers was significant (binomial test p-values < $4 \times 10^{-4}$, Figure 7a), while this difference in brain was not significant. Also, tele-enhancers contained more low-DAF SNPs than proximal enhancers in all tested tissues (binomial test p-value = $3 \times 10^{-2}$ for heart and lung, and p-value = $7 \times 10^{-2}$ for brain, Figure 7b).

Taken together, tele-enhancers, although having different selective signatures across different tissues, have been consistently evolving under stronger negative constraints than proximal enhancers both during modern human history and in the separation of vertebrates, indicating that the observations we obtained from the study of heart enhancers are applicable to other tissues and represent a general trend in the evolution of proximal enhancers and tele-enhancers.

**Tele-enhancers from diverse cell types show consistent evolutionary and functional features**
We also extended our study to a large panel of diverse cell types including GM12878, H1-hESC, HepG2, HSMM, HUVEC, K562, and NHEK, for which gene expression and ChromHMM enhancer maps have been previously reported (DJEBALI et al. 2012) (ERNST and KELLIS 2012). We applied our pipeline to each of these cells and identified ~6000 tele-enhancers in different cell lines (Table S2).

In agreement with our heart tele-enhancer observations, GeneTs were found to be enriched in house-keeping genes as compared to the respective GenePs in all cell types. In all cells, GeneTs showed lower relative expression than GenePs. In all cases except K256, the relative-expression difference between GeneTs and GenePs was significant (rank sum test p-values < 2 \times 10^{-3}, Figure 8a). These observations indicate that tele-enhancers likely play basic and fundamental biological roles independent of their cell-type specificity. Across all cell types, tele-enhancers featured remarkably lower human-specific divergence than their proximal counterparts (Fisher’s exact test p-values < 2 \times 10^{-3}) and neutrally evolving DNA (Figure 8b and Table S9), confirming the uniform nature of strong purifying selection acting on them. Moreover, tele-enhancers displayed a significantly lower SNP density across all cell types as compared to proximal enhancers (binomial test p-values < 7 \times 10^{-6}, Figure 8c). They also displayed a strong preference towards low-DAF SNPs—the fractions of low-DAF SNPs in tele-enhancers were those in respective proximal enhancers in all cell types (binomial test p-values < 4 \times 10^{-3}, Figure 8d).

**DISCUSSION**

Understanding chromatin looping and its role in accurately positioning long-range enhancers into the immediate proximity of their target promoters has remained one of
the most challenging problems of the post-genome era (CARTER et al. 2002; ERNST 2012; SANYAL et al. 2012; Sexton et al. 2009; West and Fraser 2005). We observed that an underappreciated, large fraction of tissue-specific genes—64% of heart genes, 47% of brain genes, and 50% of lung genes—lack proximal enhancers, being regulated by so-called tele-enhancers (enhancers that are separated from their target genes by at least one bystander gene or exon). This explicitly demonstrated that spatial proximity is not the mechanism by which enhancers communicate with target genes and a large portion of enhancers might recognize their distant partners while “skipping” bystander genes (Zhang et al. 2013). However, our knowledge of the genome-wide distribution of tele-enhancers is very limited since the previous studies of these enhancers focused on individual sequences, instead of the genome-wide map of tele-enhancers (Dong et al. 2009; NavratiLOva and Becker 2009; Sanyal et al. 2012), and the genome-wide map of chromatin interactions has yet to be experimentally established in most cell lines/tissues (Belton et al. 2012; Zhang et al. 2012). To improve our understanding of tele-enhancers, we analyzed genome-wide enhancer maps established in diversified tissues/cell types based on the activity of different biomarkers, such as transcriptional co-activator p300, enhancer-associated histone mark H3K4me1, and a combination of regulatory-related chromatin signals. After assigning enhancers to their target genes on a genome-wide scale, we established the maps of tele-enhancers in the human genome in 3 tissues (i.e., fetal heart, brain, and lung) and 7 cell types (such as GM12878, H1-hESC, K562, etc.). We then compared tele-enhancers to proximal enhancers systematically and investigated functional, regulatory, and evolutionary mechanisms specific to tele-enhancers.
We demonstrated that the genes associated with heart *tele*-enhancers (i.e., GeneTs) partake in basic biological processes, showing lower heart specificity than genes associated with proximal enhancers (i.e., GeneP). Also, heart GeneTs have significantly shorter noncoding space in their neighborhood than GenePs. These findings are in line with the “selection for economy” model stating that widely-expressed gene loci are compact due to strong pressure for shortening non-coding regions and this might explain why they rely on *tele*-enhancers for transcription activation (Eisenberg and Levanon 2003; Pozzoli et al. 2007; Vinogradov 2006). Also, heart *tele*-enhancers have the TF binding motif signature distinct from proximal heart enhancers. For example, the binding motifs of TEAD and NKX-2.5 were over-represented in *tele*-enhancers comparing to proximal enhancers while GATA4 displayed an opposite trend. This suggests that heart *tele*-enhancers, as compared to proximal counterparts, regulate distinct biological processes, and recruit different transcriptional machinery.

We also measured nucleotide divergence of heart enhancers among human, chimpanzee, and macaque, and observed that heart *tele*-enhancers displayed low human-specific divergence. Also, heart *tele*-enhancers harbored less SNPs and more likely contained low-DAF SNPs as compared to neutral reference and their proximal counterparts. All these findings consistently suggested that heart *tele*-enhancers were imposed with stronger negative selective pressure than proximal counterparts when evolving into human and during modern human history. Also *tele* and proximal enhancers showed almost identical GC content, CG site density, and ChIP-signal magnitude (Figure S4). After eliminating these possible confounding factors, we further ascertained that the genomic location of enhancers (i.e., the position relative to potential
target genes) was one of the determinant factors of functional and evolutionary signatures of heart enhancers.

We extended our study to fetal brain and lung where enhancers were identified in ChIP-seq experiments targeting H3K4me1, and a panel of 7 cell lines where enhancers were predicted according to chromatin signatures. The obtained results suggest that our results represent a common trend across different tissues and cell types.

How enhancers “travel” over intermediate regions and interact with remote core promoters to initiate transcription is one of the most enigmatic aspects of gene regulation (KLEINJAN and VAN HEYNINGEN 2005; PHILLIPS and CORCES 2009; Sexton et al. 2009). Our findings shed light on the interactions between remote enhancers and their targets, which are directly relevant to developing future strategies for analyzing tele-enhancers and understanding their role in establishing complex gene regulatory landscapes of vertebrate genomes.

FUNDING
This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

URLs

ACKNOWLEDGEMENTS
We are grateful to Radhouane Aniba and Cindy Clark for critical reading of the manuscript.

Conflict of interest statement. None declared.

REFERENCES

ABDI, H., 2007 Bonferroni and Sidak corrections for multiple comparisons in 

Encyclopedia of Measurement and Statistics, edited by N. J. SALKIND. Sage,

Thousand Oaks, CA.

AHITUV, N., S. PRABHAKAR, F. POULIN, E. M. RUBIN and O. COURONNE, 2005 Mapping cis-
regulatory domains in the human genome using multi-species conservation of 
synteny. Human Molecular Genetics 14: 3057-3063.

AKALIN, A., D. FREDMAN, E. ARNER, X. DONG, J. BRYNE et al., 2009 Transcriptional 
features of genomic regulatory blocks. Genome Biology 10: R38.


BALASUBRAMANIAN, S., D. ZHENG, Y.-J. LIU, G. FANG, A. FRANKISH et al., 2009 
Comparative analysis of processed ribosomal protein pseudogenes in four 
mammalian genomes. Genome Biology 10: R2.

BELTON, J.-M., R. P. MCCORD, J. H. GIBCUS, N. NAUMOVA, Y. ZHAN et al., 2012 Hi–C: A 
comprehensive technique to capture the conformation of genomes. Methods 58:
268-276.

BLOW, M. J., D. J. MCCULLEY, Z. LI, T. ZHANG, J. A. AKIYAMA et al., 2010 ChIP-Seq 


GOFFART, S., J.-C. VON KLEIST-RETZOW and R. J. WIESNER, 2004 Regulation of mitochondrial proliferation in the heart: power-plant failure contributes to cardiac failure in hypertrophy. Cardiovascular Research 64: 198-207.


Pomerantz, M. M., N. Ahmadiyeh, L. Jia, P. Herman, M. P. Verzi et al., 2009 The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882-884.


VISSE, M., M. KAYSER and R.-J. PALSTRA, 2012 HERC2 rs12913832 modulates human pigmentation by attenuating chromatin loop formation between a long-range enhancer and the OCA2 promoter. Genome Research.


FIGURE LEGENDS

Figure 1 Proximal and tele enhancers. (a) Schematic definition of two classes of enhancers. (b) Distribution of heart GeneTs and GenePs.

Figure 2 Heart genes and enhancers linked to them. (a) The average density of evolutionary breakpoints between enhancers and their target genes. A low density of breakpoint indicates a high probability of the enhancers and their target genes being located with the same evolutionary synteny block. ** indicate that the corresponding density is significantly low than expected, i.e., p-value < 1 × 10^{-5}. (b) Length distribution of DHS-based regulatory blocks of GeneTs and GenePs. (c) Fraction of enhancers residing within the DHS-based regulatory blocks of their associated genes. (d) Distribution of the length of non-coding regions (intergenic and intronic) of GeneTs and GenePs.

Figure 3 Function of tele and proximal enhancers. (a) Functional analysis based on GO annotation. The enrichment is measured as the ratio of the fraction of the target genes having a tested function to the expectation with matching locus (i.e., gene as well as its intergenic and intronic regions) length. The p-value is estimated using binomial test. (b) Enrichment of mitochondrial genes among GeneTs and GenePs (corresponding to tele and proximal enhancers respectively). (c) The relative expression of GeneTs and GenePs—given a gene, a low relative expression indicates a weak specificity to heart.

Figure 4 Association of TF binding motifs with different enhancer classes. (a) The weight of SVMs with linear kernel built to discriminate enhancer sequences from controls. (b) Enrichment fold of TF binding motifs in tele and proximal enhancers.
Figure 5 Schematic depiction of the human divergence, chimpanzee divergence, and macaque divergence based on three-way genome sequence alignments (see Materials and Methods). Through a comparison with neutral reference (pseudogenes in this study), the selective pressure acting on a tested region is measured by the neutrality index (NI). NI>1 indicates positive selection, while NI<1 corresponds to negative selection.

Figure 6 Divergence and Diversity of enhancers across species. (a) Human divergence (y-axis) is plotted against non-human divergence (x-axis). HACNSs are human accelerated conserved noncoding sequences. (b) Fraction of enhancers under positive and negative selective pressure. (c) The derived allele frequency spectrum of SNPs along enhancers, control sequences, neutral reference, and HACNSs. SNPs used here originate from the 1000 Human Genomes project and for 4 human populations (including AFR, EUR, ANN, and AMR).

Figure 7 SNP-based analysis of enhancers in three tissues (heart, brain, and lung). (a) Number of SNPs per kilobase of tele and proximal enhancers. (b) Fraction of SNPs with low DAF along tele and proximal enhancers. SNPs used here are detected in 1000 genomic project and for 4 human populations (including AFR, EUR, ANN, and AMR).

Figure 8 Features of tele-enhancers as compared to the respective proximal enhancers in 7 cell types. (a) Relative expression of GeneTs and GenePs. (b) Nucleotide divergence of tele and proximal enhancers. (c) Number of SNPs per kilobase. All the differences between tele and proximal enhancers are significant, i.e., binomial test p-values < 7 × 10^{-6}. (d) Fraction of SNPs having a low DAF (i.e., DAF<0.05). All
differences between *tele* and proximal enhancers are significant, i.e., p-values < $4 \times 10^{-3}$.

**TABLE LEGENDS**

*Table 1* Nucleotide divergence of heart enhancers, pseudogenes (neutral reference), control sequences and HACNs.

*Table 2* Functional analysis of proximal and *tele* lung enhancers

*Table 3* Functional analysis of proximal and *tele* brain enhancers

*Table 4* Nucleotide divergence of lung and brain enhancers according to the human-chimpanzee-macaque sequence alignment
Figure 1
Figure 3

(a) Analysis of gene expression in mitochondrial transport, ATP synthesis coupled electron transport, and other processes.

(b) Comparison of gene expression between GeneP and GeneT groups for heart genes and controls.

(c) Graph showing average relative expression levels between GeneP and GeneT for specific biological processes.
Figure 4
Neutral reference

Non-heart-specific divergence (NI<1)

Heart-specific divergence (NI>1)

\[ \text{Neutrality Index (NI)} = \frac{D_h/N_h}{(D_m+D_c)/(N_m+N_c)} \]

\( D_h \) – human divergence  
\( D_c \) – chimpanzee divergence  
\( D_m \) – macaque divergence  
\( D_c+D_m \) – non-human divergence

Figure 5
Figure 6

Graphs showing relationships between non-human-specific divergence and human-specific divergence. The graphs illustrate different categories of enhancers and their divergence patterns.
Figure 7
Figure 8
**Table 1** Nucleotide divergence of heart enhancers, pseudogenes (neutral reference), control sequences, and HACNSs.

<table>
<thead>
<tr>
<th>Sequence Region</th>
<th>Divergence (per kilobase)</th>
<th></th>
<th></th>
<th>#SNPs per kilobase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human-specific</td>
<td>Non-human-specific</td>
<td>NI</td>
<td>p-value of negative selection</td>
</tr>
<tr>
<td>Enhancers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximal</td>
<td>5.66</td>
<td>60.01</td>
<td>0.77</td>
<td>9 × 10^{-50}</td>
</tr>
<tr>
<td><em>tele</em></td>
<td>5.47</td>
<td>59.53</td>
<td>0.75</td>
<td>4 × 10^{-59}</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral reference</td>
<td>9.14</td>
<td>74.9</td>
<td>1</td>
<td>--</td>
</tr>
<tr>
<td>Controls</td>
<td>6.42</td>
<td>62.34</td>
<td>0.84</td>
<td>3 × 10^{-39}</td>
</tr>
<tr>
<td>HACNSs</td>
<td>17.24</td>
<td>29.75</td>
<td>4.75</td>
<td>1</td>
</tr>
</tbody>
</table>

HACNSs are human accelerated conserved noncoding sequences.
Table 2 Functional analysis of proximal and *tele* lung enhancers

<table>
<thead>
<tr>
<th>GOID</th>
<th>GO</th>
<th>Proximal</th>
<th></th>
<th>Tele</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Enrichment Fold</td>
<td>p-value</td>
<td>Enrichment Fold</td>
<td>p-value</td>
</tr>
<tr>
<td>GO:0030323</td>
<td>Respiratory tube development</td>
<td>14.47</td>
<td>0.00E+00</td>
<td>14.35</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0009725</td>
<td>Response to hormone stimulus</td>
<td>2.97</td>
<td>0.00E+00</td>
<td>3.14</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0035295</td>
<td>Tube development</td>
<td>6.02</td>
<td>0.00E+00</td>
<td>4.91</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0009719</td>
<td>Response to endogenous stimulus</td>
<td>2.85</td>
<td>0.00E+00</td>
<td>2.91</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0009611</td>
<td>Response to wounding</td>
<td>2.83</td>
<td>0.00E+00</td>
<td>2.49</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0030324</td>
<td>Lung development</td>
<td>15.12</td>
<td>0.00E+00</td>
<td>14.59</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0014070</td>
<td>Response to organic cyclic substance</td>
<td>3.24</td>
<td>3.42E-06</td>
<td>2.74</td>
<td>1.24E-02</td>
</tr>
<tr>
<td>GO:0048598</td>
<td>Embryonic morphogenesis</td>
<td>3.48</td>
<td>0.00E+00</td>
<td>2.19</td>
<td>1.28E-02</td>
</tr>
<tr>
<td>GO:0045596</td>
<td>Negative regulation of cell differentiation</td>
<td>4.34</td>
<td>0.00E+00</td>
<td>2.32</td>
<td>1.42E-02</td>
</tr>
<tr>
<td>GO:0035239</td>
<td>Tube morphogenesis</td>
<td>4.34</td>
<td>4.44E-13</td>
<td>2.43</td>
<td>7.02E-02</td>
</tr>
<tr>
<td>GO:0030855</td>
<td>Epithelial cell differentiation</td>
<td>3.99</td>
<td>4.69E-10</td>
<td>2.52</td>
<td>1.82E-02</td>
</tr>
<tr>
<td>GO:0008283</td>
<td>Cell proliferation</td>
<td>2.6</td>
<td>2.41E-09</td>
<td>1.97</td>
<td>2.06E-02</td>
</tr>
<tr>
<td>GO:0055093</td>
<td>Response to hyperoxia</td>
<td>3.77</td>
<td>1.00E+00</td>
<td>16.83</td>
<td>1.79E-07</td>
</tr>
<tr>
<td>GO:0007585</td>
<td>Respiratory gaseous exchange</td>
<td>2.2</td>
<td>1.00E+00</td>
<td>12.11</td>
<td>2.09E-07</td>
</tr>
<tr>
<td>GO:0055082</td>
<td>Cellular chemical homeostasis</td>
<td>1.92</td>
<td>7.66E-02</td>
<td>2.63</td>
<td>1.48E-06</td>
</tr>
<tr>
<td>GO:0046039</td>
<td>GTP metabolic process</td>
<td>1.45</td>
<td>1.00E+00</td>
<td>3.09</td>
<td>2.62E-06</td>
</tr>
<tr>
<td>GO:0006184</td>
<td>GTP catabolic process</td>
<td>1.44</td>
<td>1.00E+00</td>
<td>3.11</td>
<td>3.61E-06</td>
</tr>
<tr>
<td>GO:0006873</td>
<td>Cellular ion homeostasis</td>
<td>1.97</td>
<td>4.87E-02</td>
<td>2.61</td>
<td>8.53E-06</td>
</tr>
<tr>
<td>GO:0045730</td>
<td>Respiratory burst</td>
<td>0</td>
<td>1.00E+00</td>
<td>30.30</td>
<td>7.80E-06</td>
</tr>
<tr>
<td>GOID</td>
<td>GO</td>
<td>proximal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------</td>
<td>-----------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enrichment</td>
<td>p-value</td>
<td>p-value</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0030900</td>
<td>Forebrain development</td>
<td>14.21</td>
<td>0.00E+00</td>
<td>9.05</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:00021537</td>
<td>Telencephalon development</td>
<td>15.2</td>
<td>0.00E+00</td>
<td>11.68</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0031175</td>
<td>Neurite development</td>
<td>5.42</td>
<td>0.00E+00</td>
<td>3.66</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0007423</td>
<td>Sensory organ development</td>
<td>3.8</td>
<td>0.00E+00</td>
<td>3.39</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0048812</td>
<td>Neurite morphogenesis</td>
<td>5.77</td>
<td>0.00E+00</td>
<td>3.64</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0030182</td>
<td>Neuron differentiation</td>
<td>5.44</td>
<td>0.00E+00</td>
<td>3.52</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0007417</td>
<td>Central nervous system development</td>
<td>8.04</td>
<td>0.00E+00</td>
<td>8.76</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0021543</td>
<td>Pallium development</td>
<td>16.82</td>
<td>0.00E+00</td>
<td>11.18</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0007420</td>
<td>Brain development</td>
<td>11.09</td>
<td>0.00E+00</td>
<td>11.71</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0030902</td>
<td>Hindbrain development</td>
<td>12.72</td>
<td>0.00E+00</td>
<td>12.53</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0001764</td>
<td>Neuron migration</td>
<td>7.81</td>
<td>0.00E+00</td>
<td>4.40</td>
<td>1.16E-03</td>
</tr>
<tr>
<td>GO:0001843</td>
<td>Neural tube closure</td>
<td>8.65</td>
<td>8.66E-11</td>
<td>5.22</td>
<td>3.64E-03</td>
</tr>
<tr>
<td>GO:0043523</td>
<td>Regulation of neuron apoptosis</td>
<td>4.54</td>
<td>7.43E-10</td>
<td>3.11</td>
<td>5.82E-03</td>
</tr>
<tr>
<td>GO:0045665</td>
<td>Negative regulation of neuron differentiation</td>
<td>6.92</td>
<td>1.66E-05</td>
<td>5.80</td>
<td>7.99E-03</td>
</tr>
<tr>
<td>GO:0007611</td>
<td>Learning and/or memory</td>
<td>6.54</td>
<td>0.00E+00</td>
<td>3.13</td>
<td>8.20E-03</td>
</tr>
<tr>
<td>GO:0021696</td>
<td>Cerebellar cortex morphogenesis</td>
<td>69.24</td>
<td>0.00E+00</td>
<td>7.83</td>
<td>6.21E-02</td>
</tr>
<tr>
<td>GO:0021895</td>
<td>Cerebral cortex neuron differentiation</td>
<td>44.51</td>
<td>1.00E-10</td>
<td>8.70</td>
<td>1.21E-01</td>
</tr>
<tr>
<td>GO:0021680</td>
<td>Cerebellar Purkinje cell layer development</td>
<td>49.46</td>
<td>1.78E-12</td>
<td>8.70</td>
<td>1.21E-01</td>
</tr>
<tr>
<td>GO:0046907</td>
<td>Intracellular transport</td>
<td>1.58</td>
<td>1.36E-01</td>
<td>2.41</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>GO:0021854</td>
<td>Hypothalamus development</td>
<td>6.18</td>
<td>7.73E-01</td>
<td>46.98</td>
<td>5.86E-11</td>
</tr>
<tr>
<td>GO:0032107</td>
<td>Regulation of response to nutrient levels</td>
<td>0</td>
<td>1.00E+00</td>
<td>31.32</td>
<td>6.22E-06</td>
</tr>
<tr>
<td>Sequence Region</td>
<td>Divergence (per kilobase)</td>
<td>Selection</td>
<td>p-value against neutral ref.</td>
<td>p-value of proximal vs tele</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------</td>
<td>-----------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Human-specific</td>
<td>Non-human-specific</td>
<td>NI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brain enhancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximal</td>
<td>5.38</td>
<td>58.32</td>
<td>0.77</td>
<td>$2 \times 10^{-48}$</td>
<td>$5 \times 10^{-3}$</td>
</tr>
<tr>
<td>Tele</td>
<td>5.38</td>
<td>57.02</td>
<td>0.75</td>
<td>$8 \times 10^{-57}$</td>
<td></td>
</tr>
<tr>
<td>Lung enhancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proximal</td>
<td>5.47</td>
<td>58.15</td>
<td>0.77</td>
<td>$6 \times 10^{-50}$</td>
<td>$1 \times 10^{-2}$</td>
</tr>
<tr>
<td>Tele</td>
<td>5.30</td>
<td>57.35</td>
<td>0.75</td>
<td>$3 \times 10^{-57}$</td>
<td></td>
</tr>
</tbody>
</table>