Genome-Wide Identification of Regulatory Sequences Undergoing Accelerated Evolution in the Human Genome

Xinran Dong,[†,1] Xiao Wang,[†,1] Feng Zhang,[1] and Weidong Tian*[,1,2]

[1]State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai, P.R. China

[2]Children's Hospital of Fudan University, Shanghai, P.R. China

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: weidong.tian@fudan.edu.cn.

Associate editor: Patricia Wittkopp

Abstract

Accelerated evolution of regulatory sequence can alter the expression pattern of target genes, and cause phenotypic changes. In this study, we used DNase I hypersensitive sites (DHSs) to annotate putative regulatory sequences in the human genome, and conducted a genome-wide analysis of the effects of accelerated evolution on regulatory sequences. Working under the assumption that local ancient repeat elements of DHSs are under neutral evolution, we discovered that ~0.44% of DHSs are under accelerated evolution (ace-DHSs). We found that ace-DHSs tend to be more active than background DHSs, and are strongly associated with epigenetic marks of active transcription. The target genes of ace-DHSs are significantly enriched in neuron-related functions, and their expression levels are positively selected in the human brain. Thus, these lines of evidences strongly suggest that accelerated evolution on regulatory sequences plays important role in the evolution of human-specific phenotypes.

Key words: **accelerated evolution, DHS, regulatory sequence.**

Introduction

Positive selection has long been thought to be the driving force behind phenotypic distinctions between humans and our closest relatives—the chimpanzees, especially with respect to cognitive, behavioral and dietary traits (Vallender and Lahn 2004). Discovering sequence variations in the human genome that are signatures of selection is therefore of great value for understanding the evolution of human-specific phenotypes. Earlier studies have focused mainly on identifying protein-coding genes that have been subjected to positive selective pressure because sequence variations in protein-coding genes may alter their biochemical functions, making it easier to interpret the phenotypic consequences of such mutations. However, although a number of positively selected genes have been discovered in the human genome, few have been found to be involved in neuron development (Wyckoff et al. 2000; Janeway et al. 2001; Starr et al. 2003; Zhang 2003; Vallender and Lahn 2004). In addition to protein-coding genes, selection may also act on regulatory sequences, which might alter the expression pattern of target genes and lead to phenotypic changes. For example, the promoter sequence of *PDYN*, a gene that plays an essential role in regulating perception, behavior and memory, was found to be strongly positively selected (Rockman et al. 2005). It therefore seems likely, as has been proposed by King and Wilson (1975), that positive selection acting on regulatory sequence plays an important role in the evolution of human-specific phenotypes.

With the recent availability of complete sequenced primate genomes, large-scale studies have been conducted to investigate the selective pressures on regulatory sequences in the human genome. For instances, Prabhakar et al. (2006) examined the evolution rate of the conserved non-coding elements in the human genome and found 992 HACNS (human accelerated conserved non-coding sequence). Bird et al. (2007) performed a comparative analysis of vertebrate genome and identified 1,356 ANC (accelerated conserved non-coding sequence). Capra et al. combined the results of several studies to produce a list of 2,649 ncHARs (non-coding human accelerated region) (Erwin et al. 2013). Taylor et al. (2006, 2008) and Haygood et al. (2007) both carried out genome-wide studies analyzing the evolution of promoters, and the latter study discovered that genes with positively selected promoters are often involved in neural development. The aforementioned studies have confirmed that positive selection acts on regulatory sequences. However, these studies conducted thus far have focused on either conserved non-coding elements or promoters, which account for only a small fraction of the regulatory sequences in the human genome. Regulatory sequences are often located in open chromatin when they are active and are therefore sensitive to DNase I digestion (Gross and Garrard 1988; Crawford et al. 2006; Song and Crawford 2010). Recent functional genomics studies have identified millions of DNase I hypersensitive sites (DHSs) in the human genome that potentially encode regulatory sequences. Analysis of these DHSs revealed that most of the

Article

regulatory sequences are located either in intronic or intergenic regions, with <3% located within promoter regions and even fewer overlapping with conserved non-coding elements. Shibata et al. (2012) analyzed human, chimpanzee and macaque DNase-seq signals, and identified a number of DHSs that were gained or lost in the human lineage and were likely involved in the development of human-specific phenotypes. While we were preparing the submission of this work, Gittelman et al. (2015) published a study on the accelerated evolution of DHSs. In that study, they identified 524 DHSs that are under accelerated evolution in human. However, they implemented stringent filtering procedures which filtered 18 million DHSs to 113,577 conserved DHSs for investigation. Nevertheless, the aforementioned studies suggested that there are regulatory sequences under special evolutionary constraints in human, and are associated with human-specific phenotypes.

To gain a more comprehensive understanding of the effects of regulatory sequences evolution, we conducted a systematic analysis of accelerated evolution at DHSs in the human genome. We first identified its local ancient repeat elements (AREs) of each DHS and assumed they are neutrally evolving. It is worth noting that the neutral background sequences used in the previous researches were either introns (Shibata et al. 2012) or surrounding sequences (Gittelman et al. 2015), which were different from our investigation. Then, by comparing the phylogenetic trees constructed for DHSs and AREs using the corresponding orthologous sequences from human and four other primate genomes, we found ~0.44% DHS are under accelerated evolution (named ace-DHSs) in the human genome. Further analysis of ace-DHSs reveals that they tend to be open in more cell lines than background DHSs and are strongly associated with epigenetic marks of active transcription. In addition, the target genes of ace-DHSs were found to be significantly enriched in functions related to neural and immune development and tend to be highly expressed in the human brain. Thus, our results strongly support the idea that accelerated evolution on regulatory sequences plays a vital role in driving the evolution of human-specific phenotypes.

Results

Identification of DHSs under Accelerated Evolution (ace-DHSs) in the Human Genome

A total number of ~1.8 million DNase I hypersensitive sites (DHSs) in the human genome were obtained from UCSC genome browser, and were considered to be putative regulatory sequences. To identify DHSs that are under accelerated evolution (here termed ace-DHSs), we first determined the local ancient repeat elements (ARs), specifically, LINE1 and LINE2, and for each DHS in the human genome and assumed that they are neutrally evolving (Taylor et al. 2006).

We then obtained orthologous sequences of both DHSs and AREs from four primate genomes (chimpanzee, gorilla, orangutan and macaque). After several filtering steps, we selected a total number of 808,943 DHSs for investigation. Briefly, for each selected DHS we constructed phylogenetic trees for both the DHS and its ARE, and tested whether the human DHS is under accelerated evolution by comparing the two phylogenetic trees. Specifically, we used the SPH model in phyloP (Pollard et al. 2010) program to conduct both the 'sub-branch' and the 'sub-branch given the whole tree' tests for the human DHS. If the *P* values of both tests were significant (*P* value < 0.05 with fdr [Benjamini and Hochberg 1995] adjustment), then the DHS was defined as an ace-DHS. For details regarding these procedures, refer to the Materials and Methods section.

We identified a total number of 3,538 ace-DHSs (~0.44% of all DHS investigated) (supplementary table S1, Supplementary Material online). We simultaneously conducted the 'sub-branch' and the 'sub-branch given the whole tree' tests in the SPH model and considered a significant result if the adjusted *P* values (FDR) of both tests were <0. 05. Figure 1*a* shows the phylogenetic trees of an ace-DHS—DHS1020593 (Chr20: 11,590,285–11,590,675)—and its ARE (L1ME3C-6370, L1MA-3766 and 1ME3C-6371). In this example, both the 'sub-branch' and the 'sub-branch given the whole tree' tests for the human DHS produced significant *P* values (adjusted *P* value 0.03 and 0.02, respectively), indicating the DHS is under accelerated evolution compared with the ARE. To estimate the false positive rate of the identified ace-DHS, we conducted a simulation study by randomly assigning two adjusted *P* values to a DHS, and then counting the number of DHSs if both *P* values were <0.05. We repeated the simulation 1000 times, and found on an average 63 DHSs that would be considered significant in random experiment. Therefore, the false positive rate of ace-DHSs was ~1.78% (63/3,538) based on simulation.

We examined the overlap between the identified ace-DHSs and classes of conserved non-coding sequences found by several previous studies to be under accelerated evolution: HACNS (human accelerated conserved non-coding sequence) described in Prabhakar et al. (2006), ANC (accelerated conserved non-coding sequence) described in Bird et al. (2007), ncHAR (non-coding human accelerated region) described in Capra et al. (Erwin et al. 2013) and haDHS (accelerated DHSs in the human lineage) described in Gittelman et al. (2015). Odds Ratios of ace-DHS overlapping with HACNS, ANC, ncHAR and haDHS were 5.56, 2.95, 5.47 and 12.54, respectively. All overlaps were significant (*P* values of 1.8e-8, 1.4e-4, 0.05 and 2.2e-13 for overlap with HACNS, ANC, ncHAR and haDHS, respectively; fig. 1*b*). We also compared the overlaps between ace-DHSs and the human-specific DHSs found by Shibata et al. (2012), and found the overlaps were not significant, suggesting that the human-specific DHSs are not necessary under accelerated evolution.

Ace-DHSs Tend to Be More Active than Background DHSs

The DHS investigated in this study are an ensemble of DHSs generated from 124 experiments performed on 84 cell lines, each having a specific BTO ID according to the Brenda annotation system (Gremse et al. 2011). Based on the cell line data, we defined the cell line activity for a DHS, which was the proportion of experiments on a specific line by which the

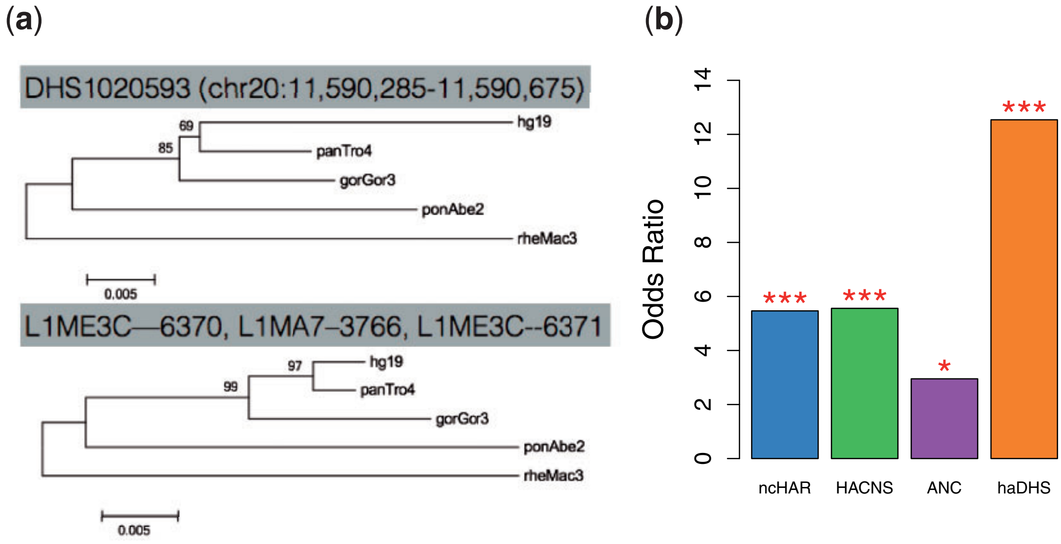


FIG. 1. (*a*) Phylogenetic tree of an ace-DHS and its local ancient repeat elements. (*b*) Bar plot for the odds ratio of the overlap between ace-DHSs and conserved non-coding sequences found to be under accelerated evolution by four other studies. A *P* value < 0.001 is indicated by '***'.

DHS was identified. If the cell line activity for a DHS is >0.5, we considered the DHS to be active in that cell line. We then counted the number of cell lines in which that DHS was active, and found that ace-DHSs tended to be active in significantly more cell lines than background DHSs. Given an ace-DHS, the average number of cell lines in which it is active is 14.3, in contrast to 9.2 for a background DHS (Wilcox test *P* value $<$ 1e-16; fig. 2*a*), indicating that ace-DHSs are generally more active than background DHSs. We further investigated the association of ace-DHS with specific cell lines. Using the cell line activity data for each ace-DHS, we clustered ace-DHSs into three groups: 375 ace-DHSs that have high activity in almost all 84 cell lines, 262 ace-DHSs that have high activity in 37 common cell lines, and the remaining 2,371 ace-DHSs that do not have a preference to specific cell lines (fig. 2*b*). Interestingly, the 37 cell lines that are associated with the second group of ace-DHS are all from differentiated tissues, such as fibroblast cells and muscle cells, whereas the remaining 47 cell lines are all stem cell like cell lines such as carcinoma, lymphocyte and stem cells (supplementary table S2, Supplementary Material online). Members of the second group of ace-DHSs likely play important roles in maintaining the differentiated state of a tissue, where the first group of ace-DHSs may be essential for basic cellular functions.

We also inspected the association of ace-DHSs with epigenetic marks. Because data are available for most epigenetic marks in the CD4 cell line, we performed the analysis using data obtained from this cell line. Compared with background DHSs, we found that ace-DHSs were significantly enriched with epigenetic marks of active transcription, whereas epigenetic marks indicating inactive transcription were significantly depleted (fig. 2*c*). For example, among most highly enriched epigenetic marks were H4K20me1, RNA Pol II, which are all well-known marks of active transcription (Kim et al. 2010; Lagha et al. 2012). The depleted marks include H3K9me2 and H3K9me3, which are significantly enriched in gene desert, imprinted domain, repeat elements, centromere, and silenced genes (Koch et al. 2007; Lee and Mahadevan 2009; Rosenfeld et al. 2009) (supplementary table S3, Supplementary Material online). It is worth noting that ace-DHSs were not statistically associated with nucleosome occupancy (supplementary fig. S2, Supplementary Material online), suggesting that they are not biased by the higher mutation rates of sequences around nucleosomes (Taylor et al. 2006, 2008; Haygood et al. 2007). We repeated the above analysis using data generated in other cell lines, and our results were similar (data not shown). Thus, ace-DHSs are not only active in more cell lines than background DHSs, but are also significantly associated with epigenetic marks of active transcription, suggesting that they tend to be more active than background DHSs.

The Target Genes of ace-DHSs Are Significantly Enriched with Neuron-Related Functions

To explore the possible phenotypic effects of ace-DHSs, we determined their putative target genes. To this end, we first classified all DHSs under investigation into two categories—local DHSs and distal DHSs—according to their relationships with coding genes. Local DHSs include those that are located in the promoter, UTR, CDS exon, or intron regions of coding genes, whereas distal DHSs are those located at least 10 kb away from any TSS of coding genes (see Materials and Methods section for details). Approximately half of the DHSs fall into the distal category. The fraction of distal ace-DHSs is similar to the fraction of background DHSs that are distal. In the previous section, we showed that ace-DHSs are strongly enriched in active epigenetic markers, implying that distal ace-DHS might tend to be enhancers. To investigate this association, we compared the overlap between distal ace-DHSs and the enhancers predicted by the program EnhancerFinder (Erwin et al. 2013). This program predicted 84,301 general enhancers in the human genome (Erwin et al. 2013), which overlap with 283 out of 1,053 distal ace-DHSs. In contrast, the predicted enhancers overlap with 53,522 out of 274,099 distal DHSs. Compared with background distal DHSs,

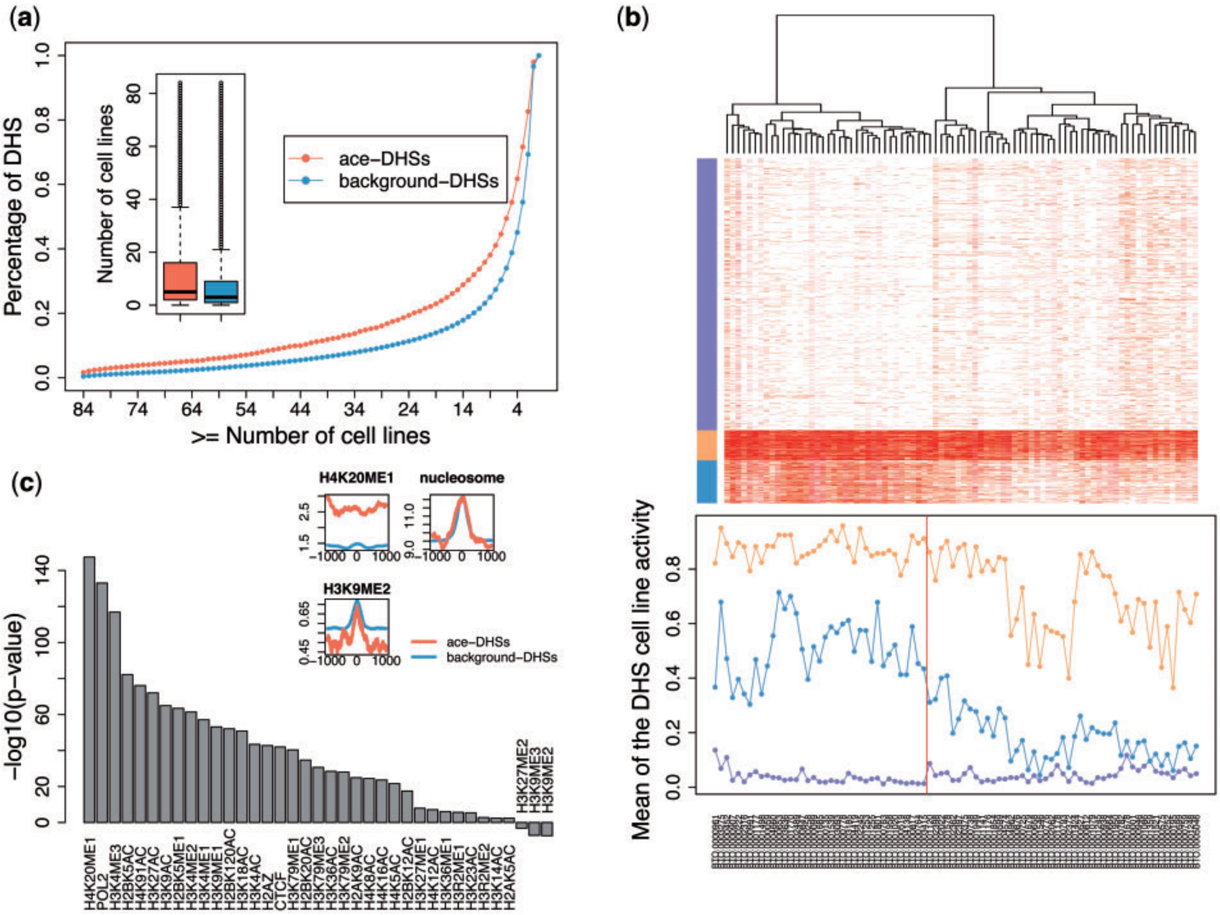


FIG. 2. (*a*) Scatterplots of the percentage of background DHSs (blue) and ace-DHSs (red) that are active in more than a given number of cell lines. Inner box shows the boxplots of the number of cell lines in which a DHS is active. (*b*) The top box shows the heat map after bi-clustering of ace-DHSs (row) and cell lines (column). The bar color to the left of the heat map corresponds to different clusters of ace-DHSs. The bottom box shows the scatterplots of the mean percentage of ace-DHSs in a given cluster that are active in a specific cell line. Each scatterplot represents a cluster of ace-DHSs and is colored the same as that in the top box. The text below the *x* axis represents the BTO ID of each cell line. (*c*) Bar plot of the −log10(*P* value) for the enrichment of epigenetic marks in ace-DHSs in comparison to that in background DHSs in the CD4 cell line. The three sub-figures on the right show the distributions of H4K20ME1, nucleosome, and H3K9me3 ChIP-Seq reads within ±1 kb of the center of ace-DHSs and background DHSs, respectively.

distal ace-DHSs have a more significant overlap with predicted enhancers (26.9% vs. 19.5% for background DHSs, with a *P* value of 4.1e-6). This finding indicated that distal ace-DHSs are more likely to be enhancers, which was consistent with our finding that ace-DHSs tend to be more active than background DHSs.

The target gene of a local DHS is simply the coding gene that overlaps with the DHS. The target gene of a distal DHS was defined as its nearest downstream gene. Here, the nearest gene approach was adopted because it was also used in other similar studies to define the target genes for distal regulatory sequences (Heintzman and Ren 2009). This approach may miss many target genes regulated by the distal regulatory sequences, because distal regulatory sequences may regulate genes from a farther distance. On the other hand, the nearest downstream gene may not also all be the true target gene. Nevertheless, for convenience we used this approach as an approximation to determining target genes. In total, we identified 18,273 and 9,805 target genes for local and distal DHSs, respectively. Among them, the number of target genes for local and distal ace-DHSs were 1,246 (6.8%) and 713 (7.3%), respectively (supplementary table S1, Supplementary Material online). We performed a GO enrichment analysis for the target genes of ace-DHSs (for details, refer to the Materials and Methods section). This analysis generated a large list of enriched GO terms, many of which were highly overlapping (i.e., sharing a substantial number of common genes), making it difficult to interpret the results. Here, we adopted a cluster-and-filter strategy to reorganize the enriched GO terms (Dong et al. 2016). Briefly, we clustered all GO terms under the Biological Process GO branch according to their relatedness (supplementary table S5, Supplementary Material online). Then, we grouped enriched GO terms according to their GO term clusters, and prepared a reduced list of enriched GO terms by selecting only the most significant GO terms from each cluster. Here, we need to emphasize that we did not discard the remaining significant GO terms. The cluster-and-filter approach was adopted simply to better display the enrichment results. All significant GO terms can be found in supplementary table S6, Supplementary Material

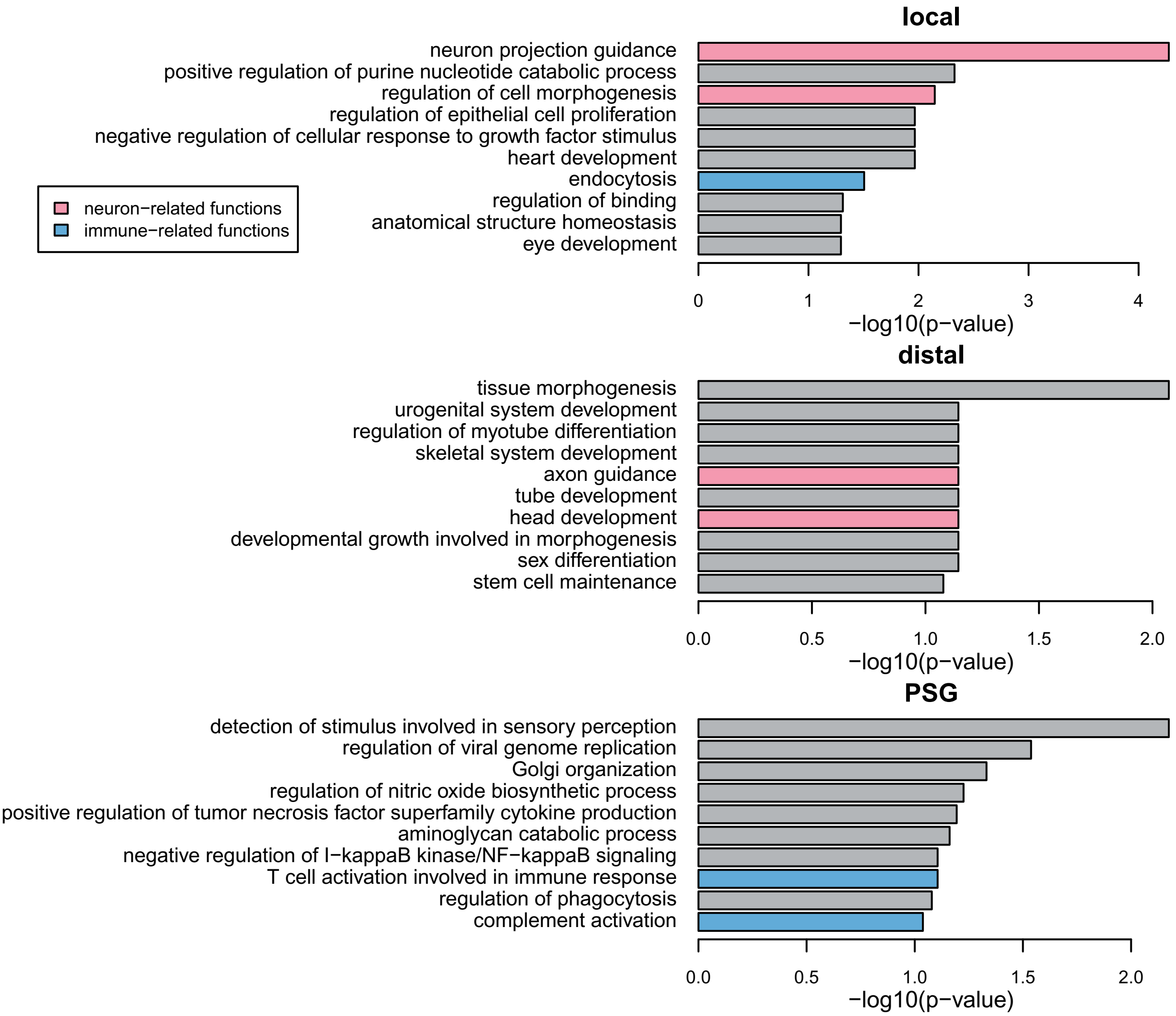
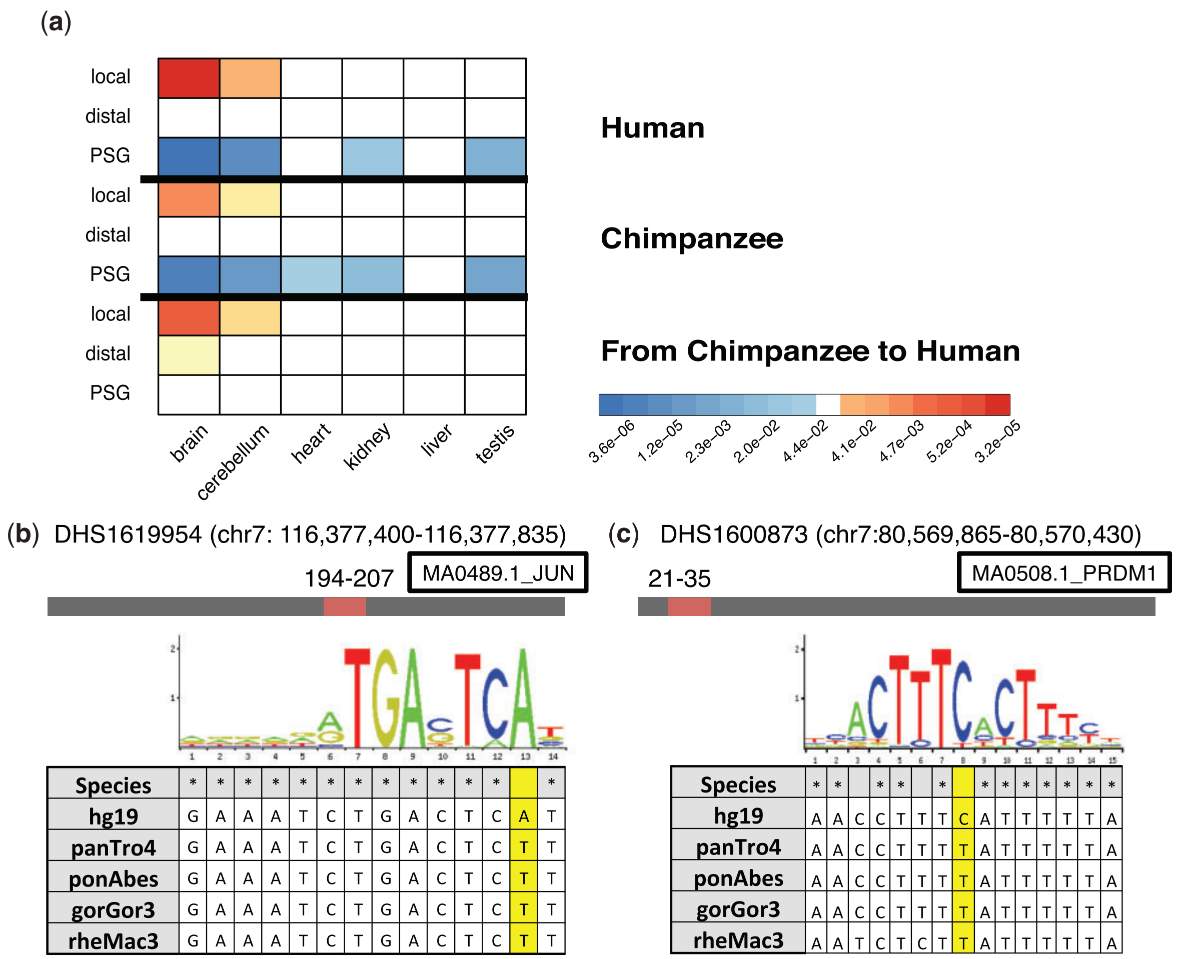


Fig. 3. The GO enrichment results for the target genes of ace-DHSs. Local, Distal and PSG represent the target genes for local ace-DHSs, the target genes for distal ace-DHSs, and positively selected genes, respectively. The enriched GO terms are organized into different clusters according to their relatedness, and the most significant GO terms of the top 10 clusters are shown in bar plots, in which the bar length corresponds to −log10(*P* value).

online. We found that the target genes of both local and distal ace-DHSs are significantly enriched with neuron and immune-related functions (fig. 3 and supplementary table S6, Supplementary Material online). In addition, the target genes of local ace-DHSs were also enriched in functions related to axon guidance and regulation of cell morphogenesis, whereas the target genes of distal ace-DHSs were also enriched with functions involved in brain development and axon guidance. In comparison, a recent study identified 304 positively selected genes (PSG) that comprises ~8.8% of all genes analyzed (Bustamante et al. 2005) (supplementary table S4, Supplementary Material online). The GO enrichment analysis of positively selected genes (PSG) revealed that they are enriched with immune-related functions, such as 'T cell activation involved in immune response', but not neuron-related functions. Thus, selective pressure on regulatory sequences can potentially affect biological process related to human-specific phenotypes, especially in neural development.

The Expression of the Target Genes of ace-DHSs Is under Positive Selection in the Human Brain

To further understand the phenotypic effects of ace-DHS, we examined the expression pattern of ace-DHS target genes in different organs. To this end, we downloaded the gene expression profile data for six organs (brain, cerebellum, heart, kidney, liver, and testis) in both human and chimpanzee that were reported by Brawand et al. (2011). Compared with target genes of the respective background DHSs, target genes of local ace-DHSs expressed at significantly higher level in the human brain and cerebellum (*P* values according to a Wilcoxon test were 3.2e-5 for brain and 6.5e-4 for cerebellum, respectively) but not in the other four organs. Similar patterns were also observed for orthologs of the target genes of ace-DHSs in chimpanzee, although they were not as significant as those observed in human. However, compared with their expression levels in chimpanzee brain, the target genes of ace-DHSs—especially those of local ace-DHSs—were significantly up-regulated in the human brain, with *P* values of 3.4e-4 and 4.1e-2 for local



Species	*	*	*	*	*	*	*	*	*	*	*	*		*
hg19	G	A	A	A	T	C	T	G	A	C	T	C	A	T
panTro4	G	A	A	A	T	C	T	G	A	C	T	C	T	T
ponAbes	G	A	A	A	T	C	T	G	A	C	T	C	T	T
gorGor3	G	A	A	A	T	C	T	G	A	C	T	C	T	T
rheMac3	G	A	A	A	T	C	T	G	A	C	T	C	T	T

Species	*	*		*	*		*		*	*	*	*	*	*	*
hg19	A	A	C	C	T	T	T	C	A	T	T	T	T	T	A
panTro4	A	A	C	C	T	T	T	T	A	T	T	T	T	T	A
ponAbes	A	A	C	C	T	T	T	T	A	T	T	T	T	T	A
gorGor3	A	A	C	C	T	T	T	T	A	T	T	T	T	T	A
rheMac3	A	A	T	C	T	C	T	T	A	T	T	T	T	T	A

Fig. 4. Gene expression analysis of the target genes of ace-DHSs. (*a*) The expression levels of the target genes of ace-DHSs in comparison to the target genes of background DHSs in six organs of human (top) and chimpanzee (middle) and the change of expression level of the target genes of ace-DHSs in six organs from chimpanzee to human (bottom). The genes in chimpanzee refer to the orthologous genes of human ace-DHSs target genes. The red color indicates that the target genes of ace-DHSs are expressed at significantly higher levels than those of background ace-DHSs or are up-regulated in human. The blue color indicates the opposite. (*b*) and (*c*) show two examples of ace-DHSs within which the transcription factor-binding site contains a mutation in the human ace-DHSs, causing increased binding affinity for the corresponding transcription factor. The content in the box is the JASPAR ID for the corresponding transcription factor.

and distal ace-DHSs, respectively (Wilcoxon test; fig. 4*a*). This pattern suggested that there is strong positive selection on the expression level of the target genes of ace-DHSs in the human brain—a possible phenotypic consequence of selective pressure on ace-DHSs. In contrast, the PSG genes were expressed at significantly lower level in nearly all organs of both human and chimpanzee, and were not differentially expressed between the respective organs of these species.

The enhanced expression of ace-DHS target genes in the human brain is likely caused by the selection on ace-DHSs. One possible mechanism by which this might occur is the mutation on ace-DHSs that potentially results in an enhanced binding affinity for transcription factors involved in the gene regulatory process. Indeed, we found many of the ace-DHS target genes up-regulated in human brain are regulated by ace-DHSs that are not only active in neuron-related cell lines, but also have at least one transcription factor-binding site (TFBS) predicted to exhibit enhanced binding affinity in human compared with other primates. Below, we presented two examples to support this hypothesis. *MET* encodes the c-Met receptor tyrosine kinase, which is essential in key social brain processes. Sequence variations in its promoter region might increase the risk of neuropsychiatric disorder, such as autism (Rudie et al. 2012). The *MET* gene expresses at high level in the human brain (quantile 98.5%), and is significantly up-regulated in the human brain compared with the chimpanzee brain. An ace-DHS (DHS1619954) is located in the intron region of MET. In the middle of this DHS, there is a JUN binding site in which the 13th position has a T->A point mutation, which significantly increases the binding affinity of JUN (fig. 4*b*). In another example, *SEMA3C* encodes a protein Semaphorin-3C, which involves in the regulation of development process and plays essential role in axon growth and guidance (Steup et al. 2000). The *SEMA3C* gene is significantly up-regulated in the human brain (quantile 96.2% by comparison between chimpanzee and human), and according to our definition of target gene of distal DHS, *3EMA3C* is regulated by the nearest distal ace-DHS (DHS1600873). This ace-DHS has a

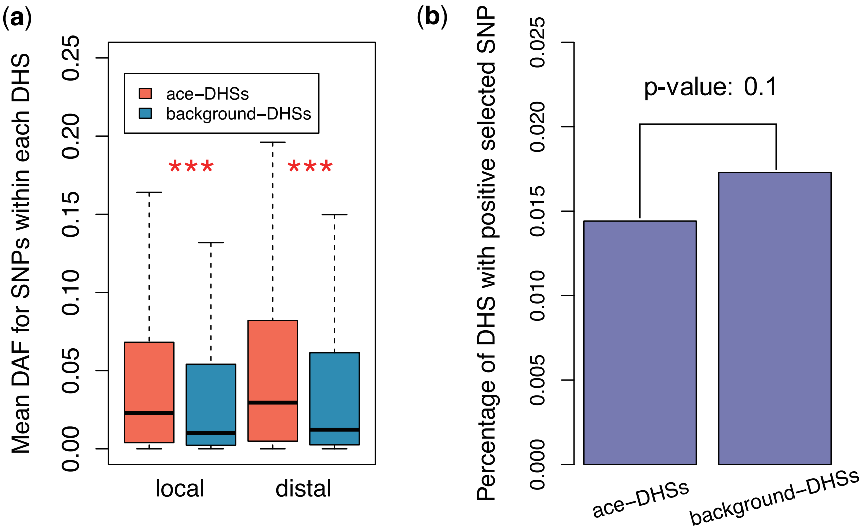


Fig. 5. (*a*) Boxplot for the mean DAF (derived allele frequency) of SNPs within ace-DHSs and within background_DHSs. Three red stars indicate that the difference of the mean DAF between ace-DHSs and background DHSs is significant, with *P* values < 0.001. (*b*) Percentage of ace-DHSs and background DHSs that contain positively selected SNPs.

PRDM1 binding site that has a T->C point mutation, leading to enhanced binding affinity for PRDM1 (fig. 4*c*). Thus, it is likely that selection on ace-DHS may occur at the TFBS in the DHS, resulting in enhanced binding affinity for transcription factors and the consequent enhanced expression of target genes, thereby leading to phenotypic changes.

The Identified ace-DHSs Are Not under Positive Selection within the Current Human Populations

The ace-DHSs identified in this study are undergoing accelerated evolution within the human lineage. However, it is not known whether they are also under selection within the current human population. Recently, the 1000 Genome Project (phase 1) has provided the DAF (derived allele frequency) for 38,248,779 SNPs (Khurana et al. 2013). We computed the mean DAF for the SNPs located within a DHS to represent the DAF of the DHS, and found that ace-DHSs have significantly higher DAF than background DHSs (fig. 5*a*), suggesting that ace-DHSs are less likely to be harmed by purifying selection. We further investigated the association of ace-DHSs with SNPs that are under positive selection within the current human populations. Li et al. (2014) has reported a list of 24,060 SNPs that are under positive selection based on a comparison of 14 population groups. The frequency of ace-DHSs that have positively selected SNP was not statistically different from that of background DHSs (fig. 5*b*), indicating that ace-DHSs are not more likely to have undergone positive selection than background DHSs within the current human populations.

Discussion

By annotating putative regulatory sequences using DNase I hypersensitive sites (DHS), we conducted a genome-wide analysis to investigate accelerated evolution on regulatory sequences in the human genome. Our analysis revealed that ~0.46% of DHS are under accelerated evolution (ace-DHSs). Although this is a small fraction, ace-DHSs may affect the expression of >7% of the genes in the genome. Though perhaps gene expression alteration caused by the selection on regulatory sequences may not be as functionally important as a change in the biochemical function of a gene through direct selection, it offers a more delicate and controllable means to gradually bring about the evolution of phenotypes. Indeed, we found that the target genes of ace-DHSs are significantly enriched with neuron-related functions and that their expression levels are significantly enhanced in the human brain. These lines of evidences provide strong support for the important role of selection on regulatory sequences in the development of human-specific phenotypes.

Several similar studies have investigated the selection on conserved non-coding sequences in the human genome (Prabhakar et al. 2006; Bird et al. 2007; Erwin et al. 2013). We have shown that ace-DHSs identified in this study share a significant overlap with the conserved non-coding sequences identified in previous studies, supporting the notion that regulatory sequences in the human genome are under selection. Since Gittelman et al. (2015) published a similar study while we were preparing the submission of our study, we made a detailed comparison between the two studies. In this study, we identified 3,538 ace-DHSs, whereas Gittelman et al. found 524 haDHS. Though the number of DHSs under accelerated evolution seems very different in the two studies, its percentage among the tested DHSs was actually quite similar (0.44% in our study, and 0.46% in Gittelman et al.'s study). On the other hand, the overlaps between ace-DHSs and haDHSs were highly significant. Therefore, our results were consistent with Gittelman et al.'s findings. The differences between our study and Gittelman et al.'s were in the followings. First of all, we implemented different filtering procedures than Gittelman et al.'s, and obtained 808,943 DHSs to test for accelerated evolution, in contrast to 113,577 conserved DHSs used by Gittelman et al. for evolutionary analysis. Secondly, the neutral background was different in the two studies. We used local LINEs within 5 kb to a DHS as the neutral background, whereas Gittelman et al. used 50 kb local sequences with filtering as the neutral background. Thirdly, the statistical tests used to evaluate the significance of accelerated evolution were different in the two studies. Fourth, the cell line activity data used by the two studies were different. Fifth, we analyzed the association of epigenetic marks with ace-DHS, whereas Gittelman et al. did not. Sixth, the target genes for distal DHSs were also defined differently in the two studies. Given all these differences, though our major conclusions were similar to Gittelman et al.'s, we provided a much larger pool of DHSs under accelerated evolution.

Generally, our study differed from those studies in the following respects. First, we analyzed the experimentally determined DHSs that represent a much more comprehensive set of regulatory sequences in the genome than the conserved non-coding sequences investigated in previous studies. Second, we used the local ancient repeat elements of each DHS as a neutral control, in contrast to the whole genome as a null model of neutral evolution used in previous studies. It has been reported that the mutation rate of genomic regions may be biased by GC conversion frequency (Marais 2003), and sequences around nucleosome tend to have higher mutation rate (Taylor et al. 2008). We showed that ace-DHSs are not statistically associated with nucleosome occupancy. We also tested the association of ace-DHSs with GC conversion frequency, and found they were not associated (supplementary fig. S1, Supplementary Material online). These indicated that the use of local ancient repeat elements as a neutral control efficiently corrects the bias caused by the higher mutation rate present at specific genomic regions. In contrast, ace-DHSs identified by using the whole genome as the neutral background were biased by GC conversion frequency (supplementary fig. S1 and table S7, Supplementary Material online), and were more frequently associated with nucleosome occupancy than ace-DHSs identified by using local AREs although the association was not statistically significant (supplementary fig. S2, Supplementary Material online). Thus, using the whole genome as a neutral control may cause biases in the results. Finally, none of the previous studies analyzed the expression of the target genes regulated by conserved non-coding sequences under selection, which we investigated extensively in this study. Thus, our study represents a more comprehensive analysis of the selection on regulatory sequences.

In this study, we showed that the expression of ace-DHS target genes is under selection in the human brain. To explain the mechanism by which the selective pressure on ace-DHSs can alter the expression pattern of target genes, we have provided several examples to show that the selection on ace-DHSs may occur at the transcription factor-binding sites within a DHS, causing enhanced binding affinity for the transcription factor, which can in turn result in an enhanced expression of target genes. Recently, Kim et al. (2010) discovered a novel class of enhancer RNA (eRNA) expressed at neural enhancers that regulate the expression of nearby genes. In our study, we found that ace-DHSs tend to be active in more number of cell lines than background DHSs, and were significantly associated with epigenetic marks of active transcription. It is therefore likely that ace-DHSs are also actively transcribed, which may produce more small RNA that might help to enhance the expression of their target genes. This process might be another mechanism by which the selection on ace-DHSs can alter the expression of their target genes. With many determined ace-DHSs, it is now possible to design experiments to test these hypotheses to further develop our understanding of the evolution of human-specific phenotypes.

Materials and Methods

Data Collection and Processing

A total number of 1,844,668 DHSs (DNase I hypersensitive sites) in the human genome were obtained from UCSC (Karolchik et al. 2009) with the track name wgEncodeRegDnaseClusteredV3. The local ancient repeat elements (AREs) of a DHS were defined as the local LINE1 (721,950 in total) and LINE2 (315,199 in total) that are within ±5 kb to the center of the DHSs in the human genome. Liftover (Hinrichs et al. 2006) was used to identify the orthologs of the DHSs and AREs in the genomes of chimpanzee, gorilla, orangutan and macaque. Firstly, a DHS and an ARE were filtered out if: (1) the DHS's length or the ARE's length was <100 bp in any genome; (2) the sequence identity between the human and primate DHSs, or between the human and primate ARE, was either <50% or equal to 100% (the alignment was performed using BLASTn; Altschul et al. 1997); (3) DHS or ARE exists in at least four species including human. Thus, 1,142,437 DHSs, 442,505 LINE1 and 207,926 LINE2 passed the three filters. Then, a DHS was filtered out if (4) its background ARE within 5 kb in human was >5 kb away from the centre of the DHSs in any of the other used primate genomes; (5) DHS's length was longer than the background ARE's or the DHS overlapped with its ARE in any genome. (6) DHSs whose local AREs (ancient repeat elements) contain the potential non-neutral LINEs (those that overlap with coding exons [GENCODE V19, plus up- and down-stream 10 bp to avoid splicing sites], promoters [GENCODE V19, transcriptional start site upstream 500 bp], simple repeats [UCSC Table browser, RepeatMasker], low complexity regions [UCSC Table browser, RepeatMasker], segmental duplications [UCSC Table browser]). Finally, 808,943 DHSs were used for the following analysis (the information about all tested DHSs and their corresponding AREs (including IDs and chromosome positions) can be downloaded from https://sourceforge.net/projects/acedhs/files/tested_DHSs.xlsx.zip/download, last accessed July 13, 2016).

Identification of DHSs under Accelerated Evolution

The multiple sequence alignments (MSAs) of DHSs and AREs were constructed by Muscle (Edgar 2004). The phylogenetic trees were constructed from the MSA using phyloFit (Siepel and Haussler 2004). phyloP (Pollard et al. 2010) was used to assess whether the DHS within the human sub-branch is under accelerated evolution, by assuming that ARE is under neutral evolution. Specifically, the SPH model (Gillies et al. 1984) in phyloP was applied, and both the 'sub-branch' and the 'sub-branch given the whole tree' tests in the SPH model were conducted. If the P values of both tests were significant (fdr adjusted P value < 0.05), then the DHS was considered an ace-DHS. In cases when the human AREs was more than twice the length of the human DHSs, a sliding window was applied to generate multiple sub-AREs from the original AREs by setting the window and step length to be the DHSs length and 10% of the DHSs length, respectively. Then, the aforementioned procedures were applied to compare the DHS with each of the sub-ARE, and the DHS was considered an ace-DHS if it was found to be significant in more than half of the cases.

Classification of DHSs

We downloaded gene annotations from UCSC (hg19, refGene), and followed a strategy similar to that of Thurman et al. (2012) to classify DHSs. A local DHS is any DHS that overlaps with a coding gene region (from 1 kb upstream of the TSS to the 3′-UTR of the coding gene). Distal DHSs are those that are located within the intergenic region and at least 10 kb away from the nearest TSS of a coding gene. The DHSs within 10–1 kb upstream to the TSS were deemed unclaissified, and were discarded.

Clustering of ace-DHSs in Cell Line Analysis

The cell line data of DHSs (Gremse et al. 2011) were downloaded from UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgEncodeVocab?type=%22cell%22, with data released before June 2012). The cell line activity of all tested DHSs can be downloaded from https://sourceforge.net/projects/acedhs/files/tested_DHSs_and_cell_line_activity.xlsx.zip/download. Bi-clustering of cell line and DHSs was first performed using hclust() in R with manhattan distance and the complete agglomeration method. The final clustering of DHSs was performed using kmeans() in R by setting $k = 3$, and the initial centers as the centers of the three groups clustered by hclust().

ChIP-Seq Data Analysis

We followed the protocol presented in Zhou et al. (2012) to process the ChIP-seq data of histone modifications, PolII and CTCF in CD4 from Lagha et al. (2012) and Barski et al. (2007). Raw ChIP-seq reads were mapped to reference genome sequences of hg19 and were averaged within ±1 kb to the center of the DHSs. t.test() in R was used to test the association of a specific mark with ace-DHSs. The P value was adjusted by Benjamini and Hochberg (1995), and the significance threshold was set to be 0.01.

Gene Ontology Enrichment Analysis

GO annotation file was downloaded from Ashburner et al. (2000) on 23 November 2014. Biological Process GO terms with a size of 30–300 genes were tested for enrichment among the target genes of a given category of ace-DHSs using fisher.test() in R. When conducting enrichment tests, we divided the tested DHSs into two categories—local and distal DHSs, and performed the analysis separately. The background set for local DHSs was the list of target genes to all local DHSs, whereas the background set for distal DHSs was the list of target genes to all distal DHSs (the target genes of all tested DHSs can be downloaded from https://sourceforge.net/projects/acedhs/files/tested_DHSs_and_target_genes.xlsx.zip/download). The significance threshold was set at 0.01. To make the GO enrichment results more interpretable, we employed a cluster-and-filter approach (Dong et al. 2016). In the clustering step, we first computed the relatedness between all pairs of GO terms under the Biological Process branch, which is defined by a Jaccard similarity that equals the number of overlapped genes/the number of union genes between two GO terms. Then, we constructed a GO term-based network in which nodes correspond to GO terms and edge weights represent the relatedness and used a network-partition algorithm called iNP (Sun et al. 2012) to partition the network into GO modules. Each GO module includes a list of GO terms that are significantly related to each other, whereas the relatedness between GO terms from different modules is low. In the filtering step, the enriched GO terms were mapped to predetermined GO modules, and the most significant GO terms from each module were selected to produce a reduced list of enriched GO terms.

Gene Expression Data Analysis

The gene expression dataset were obtained from Brawand et al. (2011), including the normalized RNA-Seq data from six organs in each of six primates, including human and chimpanzee. Wilcox.test() in R was used to determine whether the expression levels of the target genes of ace-DHSs were significantly different from those of the target genes of background DHSs in a given organ of a species. To determine whether the target genes of ace-DHSs were up-regulated in a given human organ compared with the same chimpanzee organ, the orthologous chimpanzee genes of the target genes were first identified. Then, genes were ranked according to the fold change of their expression level between chimpanzee and human, and the Wilcoxon signed-rank test was performed using wilcox.test() in R. The P value threshold was set at 0.05.

Transcription Factor-Binding Analysis

The transcription binding position weight matrixes (PWM) were obtained from JASPAR (Sandelin et al. 2004). Fimo (Bailey et al. 2009) was used with default setting to scan the sequence of ace-DHSs for candidate TFBS. The predicted TFBS was verified using the ChIP-seq data generated by the ENCODE project (Consortium 2004), and were considered putative TFBS if there existed a peak region of the corresponding TF that overlapped with the ace-DHSs.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Basic Research Program of China [2012CB316505, 2010CB529505]; the National Natural Science Foundation of China [91231116, 31071113, 30971643]; the Specialized Research Fund for the Doctoral Program of Higher Education of China [20120071110018]; the Innovation Program of Shanghai Municipal Education Commission [13ZZ006]; and X.D is a trainee of the National Institutes of Health (grant D43 TW007887). The authors thank Yuanpeng Zhou for his help with downloading files, Qingtian Gong for his help with preparing figures; and Yun Hao for suggestion on the article.

Dedications

W.T. conceived the analysis. X.D. performed the analysis. X.D., X.W., F.Z., and W.T. drafted the article. All authors read and approved the final article.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet.* 25:25.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37(suppl 2):W202–W208.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B.* 57(1):289–300.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol.* 8:R118.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478:343–348.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD. 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Consortium EP. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306:636–640.

Crawford GE, Davis S, Scacheri PC, Renaud G, Halawi MJ, Erdos MR, Green R, Meltzer PS, Wolfsberg TG, Collins FS. 2006. DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat Methods* 3:503–509.

Dong X, Hao Y, Wang X, Tian W. 2016. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Sci Rep.* 6. doi: 10.1038/srep18871.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Erwin GD, Truty RM, Kostka D, Pollard KS, Capra JA. 2013. Integrating diverse datasets improves developmental enhancer prediction. *arXiv Preprint arXiv* 1309:7382.

Gillies SD, Folsom V, Tonegawa S. 1984. Cell type-specific enhancer element associated with a mouse MHC gene, *Eβ*. 310(5978):549–554.

Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, Hawkins RD, Akey JM. 2015. Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* 25:1245–1255.

Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, Schomburg D. 2011. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* 39:D507–D513.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem.* 57:159–197.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural-and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet.* 39:1140–1144.

Heintzman ND, Ren B. 2009. Finding distal regulatory elements in the human genome. *Curr Opin Genet Dev.* 19:541–549.

Hinrichs A, Karolchik D, Baertsch R, Barber G, Bejerano G, Clawson H, Diekhans M, Furey T, Harte R, Hsu F. 2006. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* 34:D590–D598.

Janeway CA Jr, Travers P, Walport M, Capra JD, et al. 2001. Immunobiology: The Immune System in Health and Disease. 5th edition. New York: Garland Science; Part V, The Immune System in Health and Disease.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC genome browser. *Curr Protoc Bioinform.* 1.4.1–1.4.26. doi: 10.1002/0471250953.bi0104s40.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342:1235587.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465:182–187.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188:107–116.

Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U, Clelland GK, Wilcox S, Beare DM, Fowler JC, Couttet P. 2007. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.* 17:691–707.

Lagha M, Bothma JP, Levine M. 2012. Mechanisms of transcriptional precision in animal development. *Trends Genet.* 28(8):409–416.

Lee BM, Mahadevan LC. 2009. Stability of histone modifications across mammalian genomes: implications for 'epigenetic' marking. *J Cell Biochem.* 108:22–34.

Li MJ, Wang LY, Xia Z, Wong MP, Sham PC, Wang J. 2014. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res.* 42:D910–D916.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet.* 19:330–338.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20:110–121.

Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol.* 3:e387.

Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R, Zhang MQ. 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* 10:143.

Rudie JD, Hernandez LM, Brown JA, Beck-Pancer D, Colich NL, Gorrindo P, Thompson PM, Geschwind DH, Bookheimer SY, Levitt P. 2012. Autism-associated promoter variant in *MET* impacts functional and structural brain networks. *Neuron* 75:904–915.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32:D91–D94.

Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, Tewari AK, London D, Song L, Lee B-K, Iyer VR. 2012. Extensive evolutionary changes in regulatory element activity during human origins are associated with altered gene expression and positive selection. *PLoS Genet.* 8:e1002789.

Siepel A, Haussler D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*. 21:468–488.

Song L, Crawford GE. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*. 2010:pdb.prot5384.

Starr TK, Jameson SC, Hogquist KA. 2003. Positive and negative selection of T cells. *Annu Rev Immunol*. 21:139–176.

Steup A, Lohrum M, Hamscho N, Savaskan NE, Ninnemann O, Nitsch R, Fujisawa H, Püschel AW, Skutella T. 2000. Sema3C and netrin-1 differentially affect axon growth in the hippocampal formation. *Mol Cell Neurosci*. 15:141–155.

Sun S, Dong X, Fu Y, Tian W. 2012. An iterative network partition algorithm for accurate identification of dense network modules. *Nucleic Acids Res*. 40(3):e18.

Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM. 2006. Heterotachy in mammalian promoter evolution. *PLoS Genet*. 2:e30.

Taylor MS, Massingham T, Hayashizaki Y, Carninci P, Goldman N, Semple CAM. 2008. Rapidly evolving human promoter regions. *Nat Genet*. 40:1262–1263.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B. 2012. The accessible chromatin landscape of the human genome. *Nature* 489:75–82.

Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. *Hum Mol Genet*. 13:R245.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.

Zhang J. 2003. Evolution of the human *ASPM* gene, a major determinant of brain size. *Genetics* 165:2063–2070.

Zhou Y, Lu Y, Tian W. 2012. Epigenetic features are significantly associated with alternative splicing. *BMC Genomics* 13:123.