Silencing effect of Hominoid highly conserved non-coding sequences on embryonic brain development

Morteza Mahmoudi Saber[1,2] and Naruya Saitou[1,2*]

[1]Division of Population Genetics, National Institute of Genetics, Mishima, Japan

[2]Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo, Japan

***Corresponding Author:**

Naruya Saitou

Division of Population Genetics, National Institute of Genetics

Yata 1111, Mishima, 411-8540, Japan

TEL/FAX +81-55-981-6790/6789

Email: saitounr@nig.ac.jp

ABSTRACT

Superfamily Hominoidea, which consists of Hominidae (humans and great apes) and Hylobatidae (gibbons), is well-known for sharing human-like characteristics, however, the genomic origins of these shared unique phenotypes have mainly remained elusive. To decipher the underlying genomic basis of Hominoidea-restricted phenotypes, we identified and characterized Hominoidea-restricted highly conserved noncoding sequences (HCNSs) that are a class of potential regulatory elements which may be involved in evolution of lineage-specific phenotypes. We discovered 679 such HCNSs from human, chimpanzee, gorilla, orangutan and gibbon genomes. These HCNSs were demonstrated to be under purifying selection but with lineage-restricted characteristics different from old CNSs. A significant proportion of their ancestral sequences had accelerated rates of nucleotide substitutions, insertions and deletions during the evolution of common ancestor of Hominoidea, suggesting the intervention of positive Darwinian selection for creating those HCNSs. In contrary to enhancer elements and similar to silencer sequences, these Hominoidea-restricted HCNSs are located in close proximity of transcription start sites. Their target genes are enriched in the nervous system, development and transcription, and they tend to be remotely located from the nearest coding gene. Chip-seq signals and gene expression patterns suggest that Hominoidea-restricted HCNSs are likely to be functional regulatory elements by imposing silencing effects on their target genes in a tissue-restricted manner during fetal brain development. These HCNSs, emerged through adaptive evolution and conserved through purifying selection, represent a set of promising targets for future functional studies of the evolution of Hominoidea-restricted phenotypes.

Key words: conserved noncoding sequence, accelerated evolution, Hominoidea, silencer, expression regulation

Introduction:

Identification of the molecular basis of phenotypic evolution has been an active subject of evolutionary research in the past few decades, and substantial progress has occurred in this field especially after completion of whole genome sequencing projects for different phyla and classes of organisms. As a general result, it has been suggested that an enormous number of different genes are associated with the development of phenotypic characters, and modifications in the coordination of spatial and temporal expression of these genes in developmental process must have played crucial roles in the evolution of species (Nei 2007). Complex interaction of genes associated with development forms gene regulatory networks, and they are involved in signaling pathways producing phenotypes (Davidson 2006).

The evolution of the complex systems of gene regulatory networks has occurred, at least partly, by mutational changes of the regulatory regions that control the spatial and temporal expression of surrounding coding genes. In fact, there are dozens of examples for phenotypic changes generated by mutations in cis-regulatory elements such as beaks of Darwin's finch (Abzhanov et al. 2004) and pelvic fins of stickleback fish (Wray 2007). Therefore, cis-regulatory elements are considered to play crucial roles in the evolution of phenotypic characters. This hypothesis is in fact proposed long time ago by considering the small degree of amino acid differences between closely related species with dramatic phenotypic differences (e.g., King and Wilson 1975). It is now believed that mutations in cis-regulatory elements are more important than changes in protein coding regions, because novel morphological phenotypes are often linked with changes in expression level of genes rather than changes in the encoded protein sequences. This evolutionary characteristic holds in many diverse animal phyla, and a new species emerges by mutational changes of gene regulatory networks in late stages of development (Nei 2007).

So far, several mechanisms have been proposed to explain the immense phenotypic diversity observed among organisms. At the nucleotide level, mutations including all kinds of genetic changes are the only driving force of phenotypic evolution (Nei 2007, 2013). Theoretically, the majority of mutations are evolving under neutral or nearly neutral evolution for which natural selection does not affect the spread of the mutation in the population (Kimura 1983; Saitou 2013). It is possible for a phenotype-associated mutation to be fixed in a population only through neutral evolution (Kimura 1983), However, the genetic changes contributing to the adaptively important phenotypes are mostly subject to directional selection which leads to acceleration in the speed of fixation of the mutation in the population. And later on, purifying selection operates by eliminating disrupting deleterious mutations.

To infer directional or purifying selection on a sequence, first the null model of pure neutral evolution must be examined. In protein coding sequences, the synonymous mutations serve as a convenient proxy for estimation of the neutral evolutionary rate. Then by comparing the rate of nonsynonymous substitutions (abbreviated as Ka) to that of the synonymous substitutions (Ks), the selection pattern can be inferred in protein coding regions (Nei 1987; Saitou 2013). Although this approach has proven effective in analysis of protein coding regions, however, Ka/Ks cannot be calculated for noncoding cis-regulatory elements. As alternative approach, many studies of evolution of noncoding sequences have used whole genome comparative analysis in order to identify regions with unusually slow or rapid evolutionary rates. Mainly in these studies, the sequences with significantly low nucleotide substitution rate are considered as regions under purifying selection whereas the sequences with dramatically high substitution rate in the noncoding region are considered to be under positive selection (Bejerano et al. 2004; Pollard et al. 2006; Prabhakar et al. 2006; Hettiarachchi et al. 2014; Babarinde and Saitou 2016).

Conserved noncoding sequences (abbreviated as CNSs hereafter) are the sequences under selective constraint that have been repeatedly shown to be potential cis-regulatory modules by regulating gene expression (Nobrega et al. 2003; Dermitzakis et al. 2004). Although there is still considerable uncertainty regarding the function of majority of CNSs, ranging from being enhancer elements (Babarinde and Saitou 2016) to shaping chromatin structure or structural connections between chromosomes (Dermitzakis et al. 2005), there are convincing evidence showing that these elements are under purifying selection probably due to their functional importance (Drake et al. 2006; Saber et al. 2016). Despite the difference in the methodology used in identification of CNSs, these elements consistently share some properties even in different phyla. One such property is general tendency to cluster around genes involved in development and their potential role in regulation of gene expression, especially during the embryonic stage (Benko et al. 2009; Kritsas et al. 2012). Such shared properties suggest that CNSs are potent candidates to be involved in emergence of lineage-specific phenotypes.

The superfamily Hominoidea which includes humans and apes, is one of the two living superfamilies of Catarrhini, diverged from the Old World monkey lineage around 30 million years ago (Mya) (Hedges et al. 2015; see supplementary Fig. S1, Supplementary Material online). Members of Hominoidea share unique higher brain functions (Volter and Call 2012) and structural phenotypes (Crompton et al. 2008). However, the underlying genomic elements contributing to the shared phenotypic uniqueness of Hominoidea are yet unclear. Setting neutral evolution thresholds using coding and noncoding genomic sequences following Saber et al. (2016), we identified Hominoidea-restricted highly conserved noncoding sequences that have evolved in the common ancesor of Hominoidea. Using a combination of evolutionary and statistical approaches, we showed that, in contrary to ancestral CNSs, these recently evolved

conserved elements tend to have silencing effects on their target protein coding genes during embryonic brain development.

Materials and Methods

Setting neutral evolution thresholds for identification of sequences under purifying selection

The thresholds of neutral evolution were determined using the same approach used by Saber et al. (2016). We compared the human reference genome and three outgroup primate species, namely rhesus macaque, marmoset and bushbaby, and the nucleotide substitution rates were estimated from synonymous sites of protein coding sequences and from whole genome non-repetitive noncoding sequences. The substitution rates in synonymous sites and non-repetitive noncoding sequences were calculated using genes with one-to-one orthology in human and outgroup species and whole-genome noncoding DNA sequence alignments, respectively. The mode of substitution rates in synonymous sites and non-coding DNA sequences were respectively considered as the neutral evolutionary rate in protein coding and non-coding regions of the genome (supplementary Figs. S2A-S2F, Supplementary Material online). The rate of neutral evolution in synonymous sites and non-coding sequences are similar with slight skew toward conservation in protein coding synonymous sites. This slight difference is expected due to the impact of purifying selection on some of the protein coding synonymous sites since these sites are important in mRNA stability or splicing (Chamary et al. 2006). Sequences with 100 % identity in all Hominoidea members were considered as Hominoidea-shared sequences under purifying selection.

HCNS Retrieval

For the retrieval of HCNSs, protein coding regions and repetitive sequences of *Homo sapiens*, *Pan troglodytes*, *Gorilla gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys, Macaca mulatta*, *Callithrix jacchus* and *Otolemur garnettii* genomes were first masked. The dataset resources are presented in the supplementary methods, Supplementary Material online. The human genome was used as query against other four Hominoidea and three outgroup species in the homology search using NCBI BLASTN (Altschul et al. 1997) with E-value threshold of 10^{-5}. The sequences with at least 100-bp length in all members of Hominoidea which do not have any orthologs in outgroup species with conservation level above the neutral evolution threshold were identified as Hominoidea-restricted HCNSs. Due to availability of experimental data and annotation quality, human HCNSs were used for further analysis. These series of methods essentially followed those used in Saber et al. (2016).

Derived Allele Frequency (DAF) Spectrum

The frequency of genetic polymorphisms overlapping our datasets along with the state and frequency of derived alleles were extracted from the VCF files of all the human populations examined by 1000 Genomes Project Consortium (2012). The distribution of derived allele frequencies was calculated for HCNSs and random coordinates (supplementary methods, Supplementary Material online).

HCNS–Gene Association

A proximal gene regulatory domain and a distal gene regulatory domain were defined for each protein-coding gene in the human genome using GREAT (McLean et al. 2010). When GREAT

was not applicable, a Python script was written based on the principle of GREAT and was used. A proximal gene regulatory domain was defined as the region 5 kb upstream of the transcription start site (TSS) into the promoter region and 1 kb downstream of TSS into untranslated region (UTR). The proximal regulatory domain was determined regardless of other nearby protein coding genes. The distal gene regulatory domain was also defined for each protein coding gene as 1000 kb region extended at both upstream and downstream of TSS up to the basal domain of the nearest protein coding gene. Potential target of each HCNS were determined upon its overlap with the calculated gene regulatory domains.

Selection analysis and nucleotide substitution rate estimation

Each HCNS and randomly picked coordinate in the human genome were mapped to rhesus macaque and marmoset genome sequences using UCSC whole genome alignment chain files. Sequences were then aligned using Muscle (Edgar 2004), and after alignment gaps caused by insertion and deletions were discarded, a phylogenetic tree was constructed using the Neighbor Joining method (Saitou and Nei, 1987). Proportion of nucleotide differences (p distances) were calculated for each branch within phylogenetic tree for each sequence using MEGA-CC (Kumar et al. 2012). Molecular clock was assumed and applied in calculation of nucleotide substitution rates per site per year. Insertions and deletion rates within HCNS and random coordinates for their ancestral sequences were calculated upon measuring the length difference between coordinates in the human genome and their mapped sites in rhesus macaque and marmoset genome sequences.

Gene Enrichment Test

Gene ontology analysis of HCNS target genes was conducted using a similar approach used by Babarinde and Saitou (2016). First, a list of all genes with GO terms (Atotal) was retrieved from Ensembl biomart build 75. Then a list of HCNS potential target genes with GO terms (AHCNS) was prepared. Genes were represented in AHCNS according to the frequency of HCNSs targeting them. For each GO term, the number of HCNS target genes (THCNS) and total number of genes (Ttotal) associated with GO term was counted.

The GO enrichment was calculated as:

$$\text{GO Enrichment } = \frac{T_{HCNS} \times A_{total}}{T_{total} \times A_{HCNS}}.$$

Boneferroni-corrected empirical p-value was calculated based on 10^5 replicates using $\chi 2$ test.

Results

Identification of Hominoidea-restricted HCNSs

Pairwise whole-genome homology searches were conducted on coding sequence masked and repeat-masked genomes of human, chimpanzee, gorilla, orangutan, gibbon, rhesus macaque, marmoset and bushbaby (see Materials and Methods) to identify HCNSs shared only by Hominoidea. The human genome sequences were used as queries. HCNSs in this study were defined as noncoding sequences in the human genome at least 100bp long with absolute identity in chimpanzee, gorilla, orangutan and gibbon that do not have orthologous sequences in rhesus macaque, marmoset and bushbaby with conservation level above the neutral evolution (see Materials and Methods section). In order to eliminate the erroneously identified HCNSs, happening due to occasional misalignment of HCNSs with the non-conserved paralogs rather than conserved orthologs in outgroup species that occurs due to BLASTN software errors, each

HCNS in human genome was also individually mapped to rhesus macaque and marmoset genomes using whole genome alignment data (supplementary methods, Supplementary Material online). If an HCNS had conserved orthologs in rhesus macaque or marmoset genomes regardless of being repetitive, it was discarded. This approach (see diagram at supplementary Fig. S3, Supplementary Material online) identified 679 HCNSs uniquely shared by five members of Hominoidea. These sequences and their coordinate information on the human genome GRCH37 are provided in a supplementary material file, Supplementary Material online, in FASTA format.

Functional analysis of HCNSs

Sequences under functional constraint are expected to have lower derived mutations in the population. Using the genomic polymorphisms available in phase 3 of the 1000 genome project for humans, and great apes genome project for chimpanzee, gorilla and orangutan, the measured frequency of polymorphisms overlaid on HCNSs revealed that HCNSs have significantly lower non-eliminated mutations compared to random expectations (Fig. 1A), indicating the existence of functional constraint on these elements.

To confirm that the lower evolutionary rates in HCNSs are not because of their location on mutational cold spots, we conducted derived allele frequency analysis. For regions under purifying selection, derived alleles (mutants) would not be able to fix in the population and tend to remain at low frequencies, and this leads to the excess of low-frequency derived alleles. Analysis of derived allele frequency spectra for HCNSs (Fig. 1B) did reveal that the 679 HCNSs have significantly higher proportions of low-frequency derived alleles compared to random coordinates. We also showed that the selection constraint acting on HCNSs is similar to that of protein coding sequences (CDSs) and vista enhancer elements, and is significantly

higher than those for random coordinates, for human genome regions under accelerated evolution (HARs) (Pollard et al. 2006) and for CNSs under accelerated evolution in human (HACNs) (Prabhakar et al. 2006) (Supplementary Fig. S12, Supplementary Material online). We checked the conservation level of HCNS and their upstream and downstream flanking regions in genomes of humans and great apes, and they showed that non-neutral conservation of HCNSs is extended to neither upstream nor downstream regions (Fig. 1C). These results clearly demonstrate that HCNSs are different from their up/downstream flanking regions and from random coordinates regarding the action of purifying selection, and prove the existence of functional constraint on these sequences.

Evolution of HCNSs

Having confirmed that HCNSs are under purifying selection, we then asked how these sequences have evolved in the common ancestor of Hominoidea. This important question is mostly unanswered in studies of conserved noncoding sequences except for our previous study on Hominidae-restricted HCNSs (Saber et al. 2016). This is probably due to the algorithms used for identification of these conserved elements that limits the possibility of identification of CNS ancestral and orthologous sequences in closely related species. Setting and using the neutral evolution threshold for identification of HCNSs provide the opportunity for identification of HCNS orthologs in closely related species, and also makes it feasible to analyze and characterize the evolutionary changes occurred at HCNS ancestral sequences during the evolution of common ancestor of Hominoidea. For this investigation, we first mapped each of Hominoidea HCNSs to genome data of the two closest species, namely, rhesus macaque and marmoset.

Out of 679 HCNSs, 364 (53.6%) could be mapped in the rhesus macaque genome and 352 (51.8%) could be mapped to the marmoset genome. Out of the mapped sequences in rhesus macaque and marmoset genomes, 203 (29.9%) were shared. We then aligned the sequences and calculated the evolutionary distances between sequences for each mapped HCNS (See materials and methods). By constructing phylogenetic tree using the mapped HCNSs in rhesus macaque and marmoset genomes, the evolutionary distances (branch lengths) of Hominoidea ancestral sequences were calculated. The same analysis was also performed for random coordinates with the same size but ten times higher in number compared to HCNSs. The total average nucleotide substitution rate at Hominoidea ancestral HCNSs sequences is 2.38 times higher than that of random coordinates (Fig. 2A). Nucleotide substitution rates from evolutionary distances at Hominoidea HCNS ancestral sequences revealed a bimodal graph with one mode at 9E-10 that is approximated to the single mode of the nucleotide substitution distribution for random coordinates, and the second mode at 2.8E-10, that indicates accelerated nucleotide substitution rate at HCNS ancestral sequences for a portion of HCNSs (Fig. 2B). These 81 HCNSs whose ancestral sequences experienced accelerated evolution are designated in the supplementary file, Supplementary Material online, mentioned before.

Acceleration of the nucleotide substitution rate observed at Hominoidea HCNSs ancestral sequences was computed using 30% of total HCNSs successfully mapped in rhesus macaque and marmoset genomes. To confirm this result using a higher proportion of HCNSs, the pairwise evolutionary distances between human HCNSs and their orthologous sequences in the rhesus macaque genome were computed using 53.6% of the total HCNSs. Similar to the nucleotide substitution rate pattern in Hominoidea HCNS ancestral sequences, a bimodal nucleotide substitution rate was revealed with first mode at 1E-09, that is equal to nucleotide substitution rate at neutrally evolving random sequences (Fig. 2C), and the second mode at 1.7E-09. The average nucleotide substitution rate between HCNSs and their orthologs in rhesus

macaque is also 1.68 times higher than that of random coordinates (Fig. 2A). These results in combination give a strong evidence for existence of accelerated nucleotide substitution rate in HCNS ancestral sequences for at least a portion of HCNSs.

To investigate other evolutionary forces contributing to the formation of HCNSs, we also probed the rates of insertions and deletions at HCNS ancestral sequences. To this end, the average length differences of HCNSs and their orthologs in rhesus macaque and marmoset genomes were calculated. The same analysis was also performed for random coordinates of the same size and ten times higher in number than HCNSs. Sixty percent of random sequences mapped to rhesus macaque and marmoset genomes have experienced no insertions nor deletions during the evolution of the common ancestor of Hominoidea, however, only 17% of HCNSs showed the same characteristic (Supplementary Fig. S4, Supplementary Material online). On the other hand, the proportion of HCNSs with orthologous sequences in rhesus macaque and marmoset genomes with length difference above ten nucleotides is significantly higher than that of random coordinates (Supplementary Fig. S4, Supplementary Material online). In summary, these results indicate that HCNS ancestral sequences have been under accelerated evolution with respect to nucleotide substitution, insertion and deletion, especially at the common ancestor of Hominoidea which have led to the formation of these conserved elements and then strong purifying selection started to operate to keep these elements in Hominoidea genomes.

Examination of HCNS distribution

Having established that HCNSs emerged through adaptive evolution and conserved by purifying selection, we then analyzed their genomic distributions. We asked whether HCNSs are preferentially located close to protein coding genes. To answer this, we defined the

proximal regulatory domain (1kb downstream and 5kb upstream of TSS) and the distal domain (1000 kb downstream and upstream of TSS up to the proximal domains of close by protein coding genes) for each protein coding gene in the human genome (see Materials and Methods section). Fig. 3 shows that HCNSs are enriched in close proximity of transcription start sites, especially at distance between 5 to 50 kb and underrepresented at distances farther than 50 kb. There is no significant difference at distances less than 5kb. LincRNAs have similar distribution as random coordinates and experimentally verified enhancer elements are located at significantly farther distances compared to HCNSs, lincRNAs and random coordinates. Genomic distribution analysis using GREAT genomic regions enrichment annotation tool (McLean et al. 2010) also revealed enrichment of HCNSs at distance range of 5-50 kb at upstream and downstream regions compared to random coordinates and vista enhancers (Visel et al. 2007) (Supplementary Fig. S5, Supplementary Material online). Silencer elements are also enriched in close proximity of TSSs at distance ranges of <50kb and underrepresented at farther distances similar to distribution pattern of HCNSs (Fig. 3 and Supplementary Fig. S5, Supplementary Material online). These results confirm nonrandom distribution pattern of HCNSs within the genome, and demonstrate similarities in genomic locations of HCNSs to silencer elements.

Features of HCNS-target genes

Proving nonrandom distribution of HCNSs, we then investigated the properties of HCNS target genes. Conserved noncoding sequences have been previously reported to be enriched in close proximity of genes involved in development, transcription and nervous system (McEwen et al. 2009; Takahashi and Saitou, 2013; Matsunami and Saitou, 2013; Babarinde and Saitou 2013; Saber et al. 2016). Gene ontology analysis confirmed that the same enrichment pattern exists

for Hominoidea HCNSs. These HCNSs also tend to be underrepresented in proximity of genes involved in defense and immunity (Fig. 4A), which was also observed for mammalian oder-specific HCNSs (Babarinde and Saitou 2013). Unique distribution and pattern of enrichment in gene functional categories of HCNS-associate genes suggest that conserved non-coding sequences are likely to be involved in evolution of gene expression especially in the tissue of fetal brain, because genes involved in transcription regulation, development and nervous system are mainly expressed at this stage.

To investigate the hypothesis that HCNSs associated genes have unique expression pattern in human tissues following the GO enrichment prediction, RNA-Seq data of human tissues from the Roadmap Epigenome project (Kundaje et al. 2015) were retrieved and analyzed. Average RPKM (reads per kilobase per million mapped reads) score for all HCNSs target genes along with target genes of random coordinates and vista enhancer elements were calculated across human tissues. As expected, HCNS target genes have unique expression pattern in embryonic brain, however, the expression of HCNS target genes in fetal brain is significantly lower than not only compared to experimentally verified vista enhancer element but also compared to random expectations (Fig. 4B). For further confirmation, RNA-seq data of human tissues (Necsulea et al. 2014) were retrieved and analyzed. The results of this analysis consistently revealed the same expression pattern across all the investigated tissues (Supplementary Fig. S6, Supplementary Material online). These results give clear evidences for association of HCNSs with lower gene expression in their proximal genes during the development of embryonic brain.

Do HCNS target genes have unique features in terms of genomic distribution and gene structure? Genes associated with conserved noncoding sequences are expected to have larger proportion of noncoding sequences due to the action of evolutionary forces to prevent loss of these potential-regulatory elements (Babarinde and Saitou, 2016). Therefore, we would expect

HCNS target genes to have larger noncoding proportions compared to the genes not targeted by HCNSs. Analysis of HCNS target gene structure confirmed this hypothesis; HCNS target genes have considerably higher proportion of noncoding sequences (93.79%) compare to the whole genome average of genes not targeted by HCNSs (86.56%) as shown in Fig. 5A.

To figure out whether HCNS target genes have a nonrandom distribution in the genome, we analyzed the distance between HCNS associated coding genes and their proximal coding genes. Median distances to upstream and downstream flanking genes showed that HCNS target genes are located dramatically farther away from the nearest coding genes compared to the whole genome average of genes not associated with HCNSs (Fig. 5B). These results indicate that HCNS target genes are unique not only in their structure but also in their location throughout the genome.

Epigenomic characterization of HCNSs

Analyzing features of HCNS-associated genes, we have shown that HCNS-target genes have significantly lower expression at fetal brain, the tissue in which HCNSs are expected to be in their most active form according to GO analysis. If HCNSs are associated with lower expression in their target genes, we would also expect epigenomic markers for active enhancer elements such as H3k4me1 (Creyghton et al. 2010, Akhtar-Zaidi et al. 2012) to be depleted in HCNSs, especially in fetal brain. To investigate this hypothesis, we analyzed the chip-seq data from roadmap epigenome project. Human tissues, for which chip-seq data are available, were classified into four categories: fetal brain, other fetal tissues, adult brain and other adult tissues. As shown in Fig. 6A, the lowest signal for H3k4me1 in fetal brain was found for HCNSs while the highest signal was found to be for vista enhancer elements. LincRNAs were also shown to

have no significant difference from random coordinates in any of the tissue categories (Fig. 6A).

The signal pattern for H3K4me3, the epigenomic mark for active promoter elements (Cain et al. 2011), is similar to that of H3k4me1 in the aspect of HCNSs having the lowest signal in fetal brain and other fetal tissues, however, the difference is also visible in adult brain and other adult tissues for H3K4me3. The observed high signal intensities of lincRNAs for H3K4me3 could be explained by their transcription rate and proximity to protein-coding genes (Babarinde and Saitou 2016).

Enhancer RNAs or eRNAs represent a class of non-coding RNAs bidirectionally transcribed from enhancer elements, and the level of eRNA expression correlates positively with expression levels of the target genes (Kim et al. 2010; Li et al. 2013), suggesting tissue-specific expression of eRNAs. Based on this hypothesis, we would expect HCNSs to have similar tissue-specific eRNA expression pattern as that of HCNS-associated genes. To investigate this hypothesis, uniformly processed whole-genome RNA-seq data were retrieved for seven human tissues (pancreas, thymus, spleen, lung, ovary, brain and fetal brain) from Roadmap epigenome project and nucleotide-wise average expressions were computed for HCNSs, random coordinates and vista enhancer elements. Vista enhancer sequences are expected to have high eRNA expression levels due to their verified enhancer function in embryonic brain, therefore, these element could be considered as positive control in eRNA expression analysis. As expected, HCNSs have lower eRNA expression levels compared to vista enhancer elements, most significantly at embryonic brain tissue (Supplementary Fig. S7, Supplementary Material online).

Programmed shielding of particular genes from the action of enhancers, done by insulators, has many applications during development and transcriptional repressors CTCF are

important for insulator activity and enhancer blocking (Herold et al. 2012). Analysis of the enrichment of CTCF binding sites shows that HCNSs are enriched in CTCF binding sites compared to random coordinates (Supplementary Fig. S11, Supplementary Material online). These results indicate a potential role of HCNSs as insulator elements and give further evidence for the likely role of these highly conserved regions as tissue-specific silencers.

Evolutionary importance of HCNS-mediated down-regulation

Evolution of non-coding regulatory elements within the genome leading to modifications in the gene expression levels is believed to mainly underlie the phenotypic differences across species. In the study of the evolution of gene expression levels in mammalian organs, it was shown that after divergence from rhesus macaque, the gene expression levels of many genes of Hominoidea have evolved drastically, especially at the nervous system (Brawand et al. 2011). HCNSs are among the strong candidates underlying these expression switches, and this hypothesis is supported by the analysis of the data by Brawand et al. (2011), which showed that multiple targets of HCNSs are among the genes with significant expression downregulation specifically at the common ancestor of Hominoidea.

Literature searches provide insights into potential functional implications of expression shifts for a number of HCNS target genes which have uniquely undergone down-regulation in the nervous system of the Hominoidea common ancestor. For instance, *growth hormone receptor (GHR)* gene, encoding transmembrane receptor protein for growth hormone and *mammalian aryl hydrocarbon receptor (AHR)* gene, a ligand-activated transcription factor mediating the toxic effects of dioxins and related compounds, have been shown to be crucial for the proper development of the nervous system (Qin and Powell-Coffman 2004; Waters and Blackmore 2011). Multiple target genes of HCNSs down-regulated in the common ancestor of

Hominoidea play key roles in growth and wiring of neurons in the nervous system. Among them, *Docking Protein 6 (Dok6), Semaphorin 3E (Sema3e)* and *Collagen Type V Alpha 2 (COL5A2)* are noticeable. *Dok6*, a novel Dok-4/5-related adaptor molecule, is highly expressed in the developing central nervous system and regulates Ret-mediated axonal projection (Crowder et al. 2004). *Sema3e* is a member of the semaphorin family, which is an important group of proteins controlling cell migration and axonal growth cone guidance (Roth et al. 2009). *COL5A2*, a member of collagen adhesion molecules expressed in the vertebrate nervous system, is also involved in neural circuit formation (Fox 2008). Autism is a complex genetic disorder, developing as a result of several factors including the collective misregulation of multiple genes inside brain (Purcell et al. 2001). *Neuroligin-4, X-linked (NLG4X)*, functioning as splice site-specific ligands for beta-neurexins involved in the formation and remodeling of central nervous system synapses, and *Eighty-five Requiring 3A (EFR3A)*, a critical component of a protein complex required for the synthesis of the phosphoinositide PtdIns4P with variety of functions at the neural synapse, are yet other targets of HCNSs involved in pathogenesis of autism spectrum disorders (ASDs) (Gupta et al. 2014; Jamain et al. 2003). During the mammalian evolution, these two genes are also specifically downregulated during the evolution of common ancestor of Hominoidea, however, the exact phenotypic effects of these downregulations in the context of nervous system are yet to be known. These results in total indicate the potential role of HCNSs in down-regulating the expression of the specific genes in the nervous system of Hominoidea common ancestor, and these HCNSs may underlie the evolution of unique developmental and behavioral characteristics of Hominoidea.

Discussion

Using a computational approach, we have identified 679 highly conserved noncoding genomic elements only shared by all members of Hominoidea. The strong purifying selection acting on HCNSs further indicates the functionality of these conserved sequences as it proves the constant action of natural selection to eliminate mutations occurring within these sequences. The potent purifying selection acting on HCNSs also demonstrates the critical functional importance of these conserved elements in the evolution and adaptions of Hominoideas.

Mammalian conserved noncoding elements have been proposed to be classified into two groups with different modes of evolution; the first mode group consists of CNSs where a single parameter can model the nucleotide substitution rate throughout the phylogeny (Kim and Pritchard 2007) and the second mode group departures from the basic model with speed-ups and slow-downs on particular branches (Kim and Pritchard 2007; Doan et al. 2016). HCNSs we analyzed in this study mainly follow the evolutionary pattern of the latter group and have experienced accelerated nucleotide substitution rate (Fig. 2) along with the accelerated rate of insertions and deletions (Supplementary Fig. S4, Supplementary Material online) in the common ancestor of Hominoidea. This is followed by strong selective constraint which has led to absolute conservation of these elements in the superfamily Hominoidea.

We previously found that 31% (164 out of 527) of ancestral sequences of Hominidae-specific HCNSs experienced accelerated evolution (Saber et al. 2016). In this study, 40% (81 out of 203 HCNSs) of Hominoidea-restricted HCNSs ancestral sequences showed the number of nucleotide substitutions that were significantly higher than those of purely neutrally evolving sequences ($p < 0.01$; See Fig. 2B and the supplementary file, Supplementary Material online). If these HCNSs are responsible for Hominidae-specific and Hominoidea-specific macroscopic phenotypes, their ancestral sequences were real targets of positive selection, which is classically considered to be the main force of evolution. Leung et al. (2015) showed that even one-nucleotide change may influence enhancer activity. It is therefore possible that an ancestral

HCNS, which was not previously conserved, gradually acquired some enhancer activity step by step through a series of nucleotide substitutions, and eventually that sequence became indispensable for a group of organisms such as Hominoidea or Hominidae, and now we recognize it as an HCNS.

It has been argued that many of the reported human accelerated regions (HARs) are likely to be simply a result of GC biased gene conversions (Galtier and Duret 2007). One of the main characteristics of GC biased gene conversions in mammalian genomes is an excess of AT→GC transitions which leads to high contents of GC in regions affected by biased gene conversion (Duret and Galtier 2009). GC-content analysis of Hominoidea-restricted HCNSs, however, demonstrated that not only these sequences are not GC-rich but in contrary they are GC-poor sequences (Supplementary Fig. S8, Supplementary Material online). Another phenomenon which might interfere with proper calculation of the evolutionary rate in ancestral HCNSs is that HCNSs might occasionally align not with orthologs but with paralogs in outgroup species; however, we have aimed to minimize this effect by using whole-genome global alignments in addition to making use of repeat-unmasked genome sequences. These results, in total, provide strong evidence that the significant portion of HCNSs is the result of adaptive evolution in the common ancestor of Hominoidea.

HCNSs are overrepresented in close proximity of protein coding genes, in distance range of 5 to 50 kb from the transcription start sites. The significant non-random and contradictory genomic distribution of HCNSs with respect to distance from transcription start sites compared to verified enhancers suggests that HCNSs identified in this study are not enhancer elements. On the other hand, the distribution pattern of intergenic silencer elements in the human genome clearly indicates that in contrary to enhancer elements, the silencer elements along with HCNSs tend to be located in proximity of transcription start sites. The genomic location pattern of Hominoidea HCNSs is also contradictory to the pattern of

distribution of CNSs shared by amniotes reported by Babarinde and Saitou (2016). This discrepancy could be due to the difference in functionality of old CNSs which evolved more than 300 million years ago in amniotes and serving the enhancer role in diverse variety of species, in contrast to young HCNSs emerged less than 30 million years ago that are functional only in Hominoidea.

Hominoidea-restricted HCNSs do possess unique enrichment pattern regarding active enhancer epigenomic marker (H3K4me1) and also active promoter epigenomic marker (H3K4me3). These HCNSs show depletion in tissue-specific manner especially in fetal brain compared to lincRNA and random coordinates while vista enhancer elements revealed to be significantly enriched with this marker in fetal brain as expected (Fig. 6A). Analysis of H3K4me3 promoter marker enrichment, while again showed significant depletion for HCNSs, however, the depletion showed no tissue-specificity (Fig. 6B). These results are consistent with previous studies (Leung et al. 2015; Andersson et al. 2014) and reflects more tissue-restricted mode of function of enhancers compared to promoters. These results indicate that HCNSs are not serving their roles as lincRNA nor enhancers. Analysis of transcription level of HCNSs further indicates tissue-specific silenced nature of these elements in fetal brain by showing that HCNSs produce significantly less enhancer RNAs not only compared to vista enhancer elements but also to that of random coordinates (Supplementary Fig. S7, Supplementary Material online).

It has been previously reported that Hominoidea members such as human and chimpanzee possess heterogeneous lineage-specific immune response (e.g. Barreiro et al. 2010), and on the other hand, they share similar physiological and anatomical brain characteristics (e.g. Bailey and Geary 2009; Volter et al. 2012). These observations could be explained by our findings as Hominoidea-restricted HCNSs are shown to be enriched in proximity of genes involved in the nervous system but depleted for immunity and defense (Fig.

4A). It has also been suggested that modifications in temporal and spatial gene expression during development play crucial role in the evolution of species (Nei 2007). The results of our analysis further corroborates this hypothesis by showing that target protein coding genes of HCNSs are enriched for developmental process and do possess significantly modified expression pattern within the tissue of fetal brain, which could be involved in evolution of family-specific unique intellectual characteristics observed in Hominoidea.

These results, in total, strongly suggest that HCNSs are imposing tissue-restricted silencing effects on their proximal genes that are involved in embryonic brain development. Similar characteristics regarding the distance to protein coding genes transcription start site, eRNA production and expression of target genes were also found for highly conserved noncoding sequences restricted to humans and great apes identified by Saber el al. (2016) (Supplementary Fig. S9, Supplementary Material online).

Hominoidea-restricted 679 HCNSs discovered in this study and Hominidae-restricted 1658 HCNSs found by Saber et al. (2016) also showed a similar characteristic in terms of chromosomal distribution. Comparison of chromosome-wide density of Hominoidea and Hominidae restricted HCNSs is shown in Supplementary Fig. S10, Supplementary Material online. Chromosome sizes were retrieved from NCBI website (https://www.ncbi.nlm.nih.gov/genome/). The HCNS density of Y chromosome is very low for both Hominoidea and Hominidae, while that of chromosome 19 is high for both groups. Because of these two chromosome values, HCNS density between Hominoidea and Hominidae are positively correlated. However, Y chromosome is known to be rich in short repeat sequences, and the positive correlation disappears if we eliminate Y chromosome values. This indicates that both Hominoidea and Hominidae restricted HCNSs are more or less evenly distributed among chromosomes; 0.1-0.4 and 0.4-0.8 per megabase for Hominoidea and

Hominidae restricted HCNSs, respectively. Hominidae-restricted HCNSs showed the highest density (1.2 per megabase), however, no clear clustering was observed (data not shown).

The Hominoidea-restricted HCNSs identified in this study emerged less than 30 Mya, and these young HCNSs are different from ancestral CNSs which have emerged more than 300 million years in three different aspects: 1) genomic distribution (Enrichment of young HCNSs in proximity of TSSs vs. depletion of ancestral CNSs in vicinity of TSSs), 2) enrichment of epigenomic markers (depletion of young HCNSs in H3K4me1 enhancer marker vs. enrichment of old ancestral CNSs in H3K4me1) and 3) expression pattern of target genes (significantly lower expression of young HCNS target genes in fetal brain vs. dramatically higher expression of ancestral CNS-associated genes in fetal brain). These results clearly indicate heterogeneous age-dependent characteristics of conserved noncoding sequences. It has also been shown that while ubiquitous transcription factor binding sites in human are GC-rich, the tissue specific transcription factor binding sites are GC-poor (Hettiarachchi and Saitou, 2016). The significantly low GC-content of Hominoidea-restricted HCNSs suggests that these sequences may be functioning as tissue-specific transcription factor binding sites. This hypothesis is in line with our finding that suggests strong tissue-specific function of Hominoidea HCNSs.

Although the silencing effect of HCNSs is deducible from their characteristics and also expression dynamics of the HCNS-associated genes, however, the mechanism by which the repression is being implemented and the functional importance of such effects is yet to be explored through experimental analysis. The Hominoidea-restricted HCNSs, therefore, represent a set of promising targets for future studies of the evolution of Hominoidea-restricted phenotypes.

Acknowledgments

The authors thank Dr. Nilmini Hettiarachchi, for her valuable suggestions and comments.

This work was partially supported by foreign student fellowship from Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan to M. M. S. and by a grant-in-aid for scientific research from MEXT of Japan to N.S. Some of the analyses were performed on National Institute of Genetics supercomputer.

Literature cited

1000 Genomes Project Consortium 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491:56-65.

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science. 305:1462-1465.

Akhtar-Zaidi B, et al. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. Science. 336:736-739.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Andersson R, et al. 2014. An atlas of active enhancers across human cell types and tissues. Nature. 507:455-461.

Babarinde IA, Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biol Evol. 5:2330-2343.

Babarinde IA, Saitou N. 2016. Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics. Mol Biol Evol. 33:1807-1817.

Bailey DH, Geary DC. 2009. Hominid brain evolution: Testing climatic, ecological, and social competition models. Human Nature. 20:67-79.

Barreiro LB, Marioni JC, Blekhman R, Stephens M, Gilad Y. 2010. Functional Comparison of Innate Immune Signaling Pathways in Primates. PLoS Genet. 6:e1001249.

Bejerano G, et al. 2004. Ultraconserved elements in the human genome. Science. 304:1321-1325.

Benko S, et al. 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet. 41:359-364.

Brawand D, et al. 2011. The evolution of gene expression levels in mammalian organs. Nature 478(7369):343-348.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene Expression Differences Among Primates Are Associated With Changes in a Histone Epigenetic Modification. Genetics. 187:1225-1234.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet. 7:98-108.

Creyghton MP, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 107:21931-21936.

Crompton RH, Vereecke EE, Thorpe SK. 2008. Locomotion and posture from the common hominoid ancestor to fully modern hominins, with special reference to the last common panin/hominin ancestor. J Anat. 212:501-543.

Crowder RJ, Enomoto H, Yang M, Johnson EM, Milbrandt J. 2004. Dok-6, a Novel p62 Dok family member, promotes Ret-mediated neurite outgrowth. J Biol Chem. 279(40):42072-42081.

Davidson EH. 2006. Gene Regulatory Networks: The Roots of Causality and Diversity in Animal Evolution, The Regulatory Genome (Burlington: Academic Press), 187-240.

Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat Rev Genet. 6:151-157.

Dermitzakis ET, et al. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. Genome Res. 14:852-859.

Doan RN, et al. 2016. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell. 167:341-354.e312.

Drake JA, et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38:223-227.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285-311.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

Fox MA. 2008. Novel roles for collagens in wiring the vertebrate nervous system. Curr Opin Cell Biol. 20(5):508-513.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 23:273-277.

Gupta A. R., et al. 2014. Rare deleterious mutations of the gene EFR3A in autism spectrum disorders. Mol Autism. 5:31.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol. 32:835-845.

Herold M., Bartkuhn M., Renkawitz R. 2012. CTCF: insights into insulator function during development. Development 139:1045-1057

Hettiarachchi N, Kryukov K, Sumiyama K, Saitou N. 2014. Lineage-specific conserved noncoding sequences of plant genomes: their possible role in nucleosome positioning. Genome Biol Evol. 6:2527-2542

Hettiarachchi N, Saitou N. 2016. GC content heterogeneity transition of conserved noncoding sequenecs occurred at the emergence of verterbrates. Genome Biol Evol. doi:10.1093/gbe/evw219

Jamain S., et al. 2003. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. Nat Genet. 34 (1):27-29

Kim SY, Pritchard JK. 2007. Adaptive Evolution of Conserved Noncoding Elements in Mammals. PLoS Genetics. 3:e147.

Kim T-K, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature. 465:182-187.

Kimura M. 1983. The Neutral theory of molecular evolution. Cambridge: Cambridge University Press.

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

Kritsas K, et al. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. Genome Res. 22:2455-2466.

Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics. 28:2685-2686.

Kundaje A, et al. 2015. Integrative analysis of 111 reference human epigenomes. Nature. 518:317-330.

Leung D, et al. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature. 518:350-354.

Li W, et al. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature. 498:516-520

Necsulea A, et al. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 505:635-640.

Matsunami M. and Saitou N. 2013. Vertebrate paralogous conserved non-coding sequences may be related to gene expressions in brain. Genome Biol Evol. 5:140-150.

McEwen GK, et al. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 5:e1000762.

McLean CY, et al. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 28:495-501.

Nei M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei M. 2007. The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA. 104:12235-12242.

Nei M. 2013. Mutation-driven evolution. Oxford University Press.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. Science. 302:413.

Pollard KS, et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature. 443:167-172.

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

Purcell A.E., et al. 2001. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. Neurology 57 (9):1618-1628.

Qin H. and Powell-Coffman J. A. 2004. The Caenorhabditis elegans aryl hydrocarbon receptor, AHR-1, regulates neuronal development. Dev Biol. 270 (1):64-75.

Roth, L., et al. 2009. The many faces of semaphorins: from development to pathology. Cell Mol Life Sci. 66 (4):649-666.

Saber MM, Adeyemi Babarinde I, Hettiarachchi N, Saitou N. 2016. Emergence and Evolution of Hominidae-Specific Coding and Noncoding Genomic Sequences. Genome Biol Evol. 8:2076-2092.

Saitou N. 2013. Introduction to evolutionary genomics. Springer.

Saitou N. and Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406-425.

Takahashi M. and Saitou N. (2012) Identification and characterization of lineage-specific highly conserved noncoding sequences in mammalian genomes. Genome Biol Evol. 4:641-657.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 35:D88-92.

Volter CJ, Call J. 2012. Problem solving in great apes (Pan paniscus, Pan troglodytes, Gorilla gorilla, and Pongo abelii): the effect of visual feedback. Anim Cogn. 15:923-936.

Waters M. J. and Blackmore D. G. 2011. Growth hormone (GH), brain development and neural stem cells. Pediatr Endocrinol Rev. 9 (2):549-53.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8:206-216.

Figure legends

Fig. 1. HCNSs are under functional constraint in Hominoidea genomes. (A) HCNSs have lower proportions of non-eliminated mutations compared to their upstream and downstream flanking regions and random coordinates. (B) HCNSs have higher ratio of low-frequency derived allele polymorphisms indicating the action of purifying selection (Chi square P value <0.001). (C) Conservation in HCNS flanking regions is equal to the whole genome average (error bars are 95% CI).

Fig. 2. Nucleotide substitution rates at ancestral sequences of Hominoidea-restricted HCNSs. (A) Color-coded phylogenetic tree of simians, representing nucleotide substitution rate ratio to neutrally evolving sequences. Nucleotide substitution rates during the evolution of common ancestor of Hominoidea (α) for ancestral sequences of HCNSs is 2.37 times higher than those under the pure neutral evolution. (B) Distribution of nucleotide substitution rates for HCNS ancestral sequences at branch α of panel (A) compared to those for neutrally evolving random coordinates provide an evidence for accelerated evolution in Hominoidea HCNS ancestral sequences. (C) Distribution of nucleotide substitution rates for HCNS ancestral sequences at branches α and β of panel (A) compared to those for neutrally evolving random coordinates also showed a similar tendency as in panel (B).

Fig. 3. Nonrandom distribution of HCNSs in the human genome. HCNSs are enriched in close proximity of protein coding genes, especially at distance range of 5-50 kb from transcription start sites. Random coordinates ten times the number of HCNSs but with the same size were used. Chi square P values are <0.0001 for pairwise comparison of HCNSs with random coordinates, vista enhancers and lincRNAs.

Fig. 4. Enrichment of HCNS-target genes. (A) Gene ontology enrichment analysis of HCNS target genes. (B) Expression enrichment of HCNS target genes across human tissues. ns (non-significant); ***P value < 0.001 (Mann–Whitney U test).

Fig. 5. Unique features of HCNS target genes. (A) HCNS target genes have significantly higher proportion of noncoding sequences than genes with no associated HCNSs (Mann–Whitney U test P values <0.0001). (B) Genes associated with HCNSs tend to be located in isolation, far away from their upstream and downstream protein coding genes compared to genes with no associated HCNSs (Mann–Whitney U test P values <0.00001).

Fig. 6. Hominoidea-restricted HCNS are depleted in enhancer and promoter epigenomic markers. (A) HCNSs have remarkably weaker signal for H3K4me1 (enhancer) compared to random coordinates and lincRNAs. The difference is most significant in the fetal brain tissue. Experimentally verified vista enhancer elements were used as positive control. (B) HCNSs also possess weaker signal for H3K4me3 (promoter) than random coordinates, lincRNAs and vista enhancer elements. Pattern of signals are relatively uniform across all tissues consistent with weak tissue-specificity of promoters. The error bars show the 95% CI.

Figure 1.

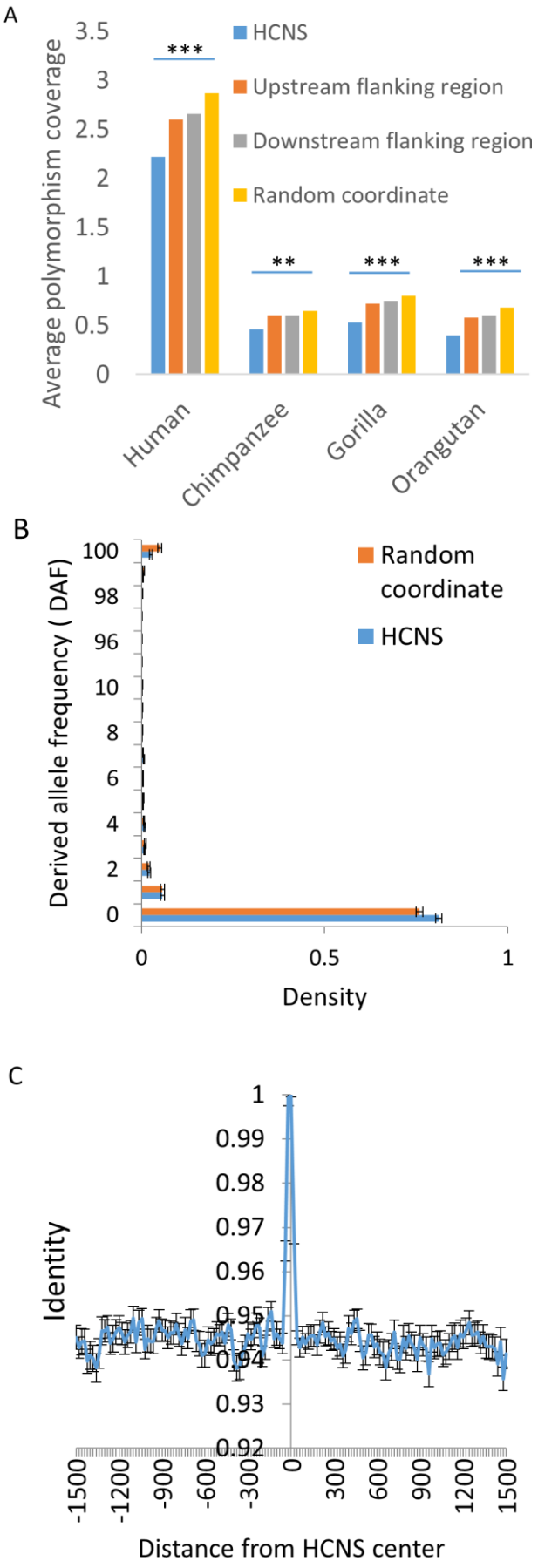

A
Average polymorphism coverage
3.5
3
2.5
2
1.5
1
0.5
0
HCNS
Upstream flanking region
Downstream flanking region
Random coordinate
***
**
***
***
Human
Chimpanzee
Gorilla
Orangutan
B
Derived allele frequency (DAF)
100
98
96
10
8
6
4
2
0
Random coordinate
HCNS
0
0.5
1
Density
C
Identity
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
-1500
-1200
-900
-600
-300
0
300
600
900
1200
1500
Distance from HCNS center

Figure 2.

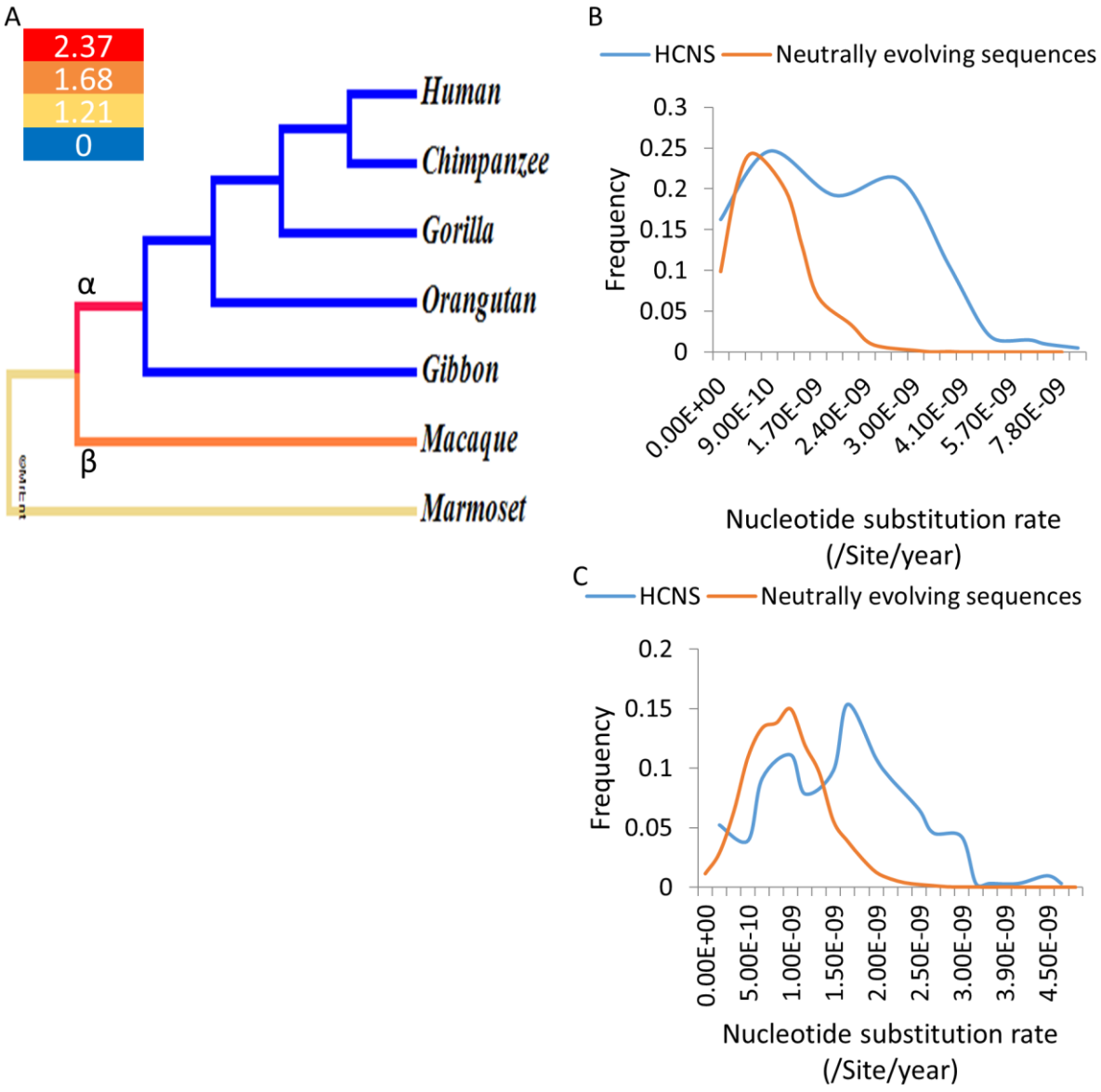

A
2.37
1.68
1.21
0
Human
Chimpanzee
Gorilla
Orangutan
Gibbon
Macaque
Marmoset
α
β
B
HCNS
Neutrally evolving sequences
Frequency
0.3
0.25
0.2
0.15
0.1
0.05
0
0.00E+00
9.00E-10
1.70E-09
2.40E-09
3.00E-09
4.10E-09
5.70E-09
7.80E-09
Nucleotide substitution rate (/Site/year)
C
HCNS
Neutrally evolving sequences
Frequency
0.2
0.15
0.1
0.05
0
0.00E+00
5.00E-10
1.00E-09
1.50E-09
2.00E-09
2.50E-09
3.00E-09
3.90E-09
4.50E-09
Nucleotide substitution rate (/Site/year)

Figure 3.

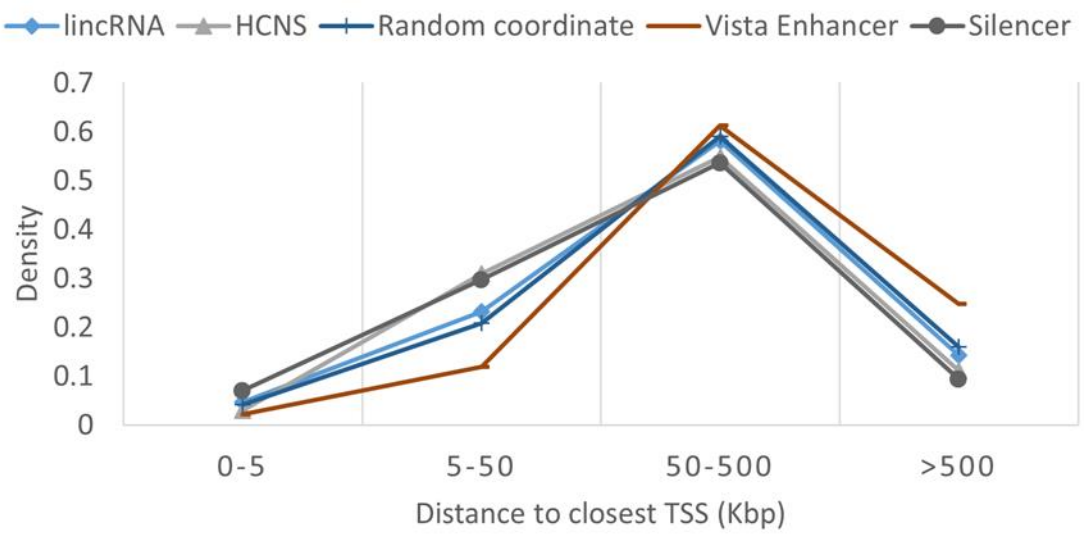

lincRNA
HCNS
Random coordinate
Vista Enhancer
Silencer
Density
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0-5
5-50
50-500
>500
Distance to closest TSS (Kbp)

Figure 4.

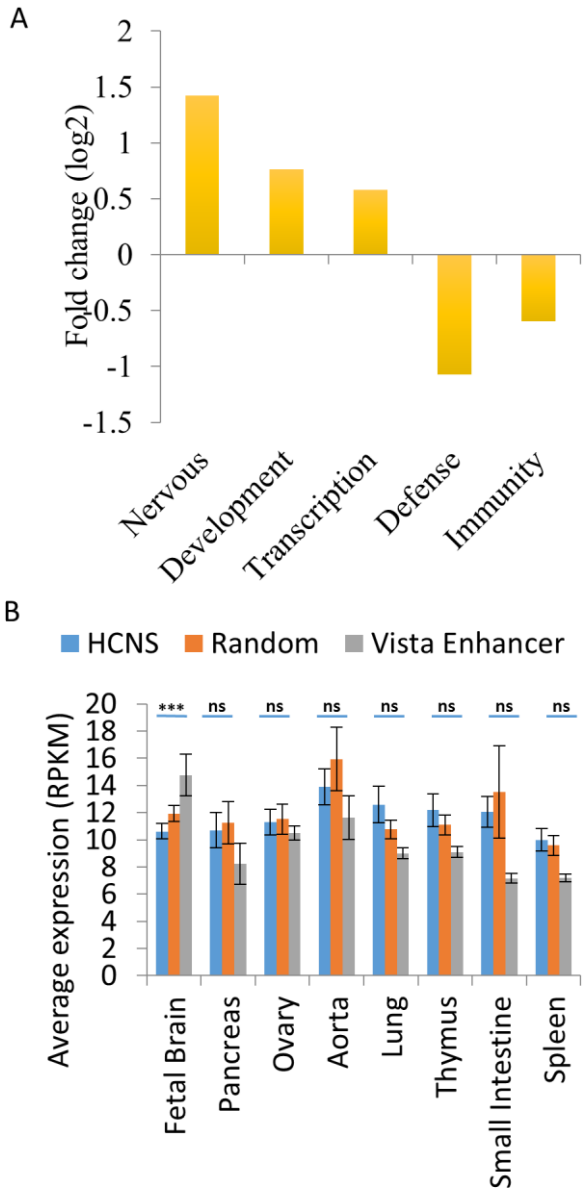

A
Fold change (log2)
2
1.5
1
0.5
0
-0.5
-1
-1.5
Nervous
Development
Transcription
Defense
Immunity
B
HCNS
Random
Vista Enhancer
Average expression (RPKM)
20
18
16
14
12
10
8
6
4
2
0
***
ns
ns
ns
ns
ns
ns
ns
Fetal Brain
Pancreas
Ovary
Aorta
Lung
Thymus
Small Intestine
Spleen

Figure 5.

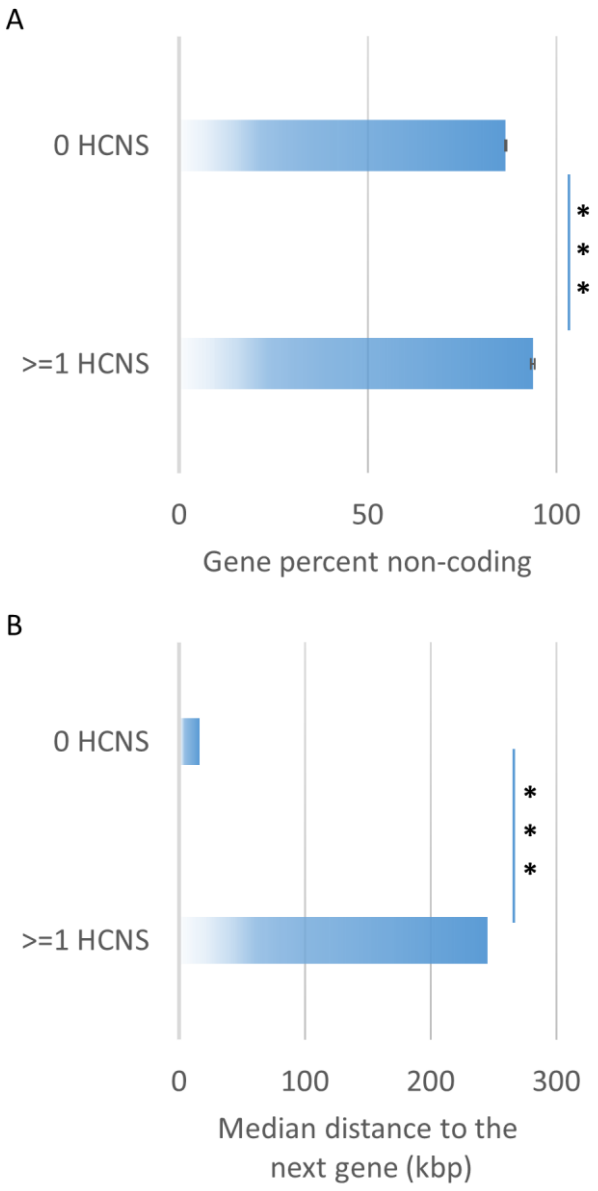
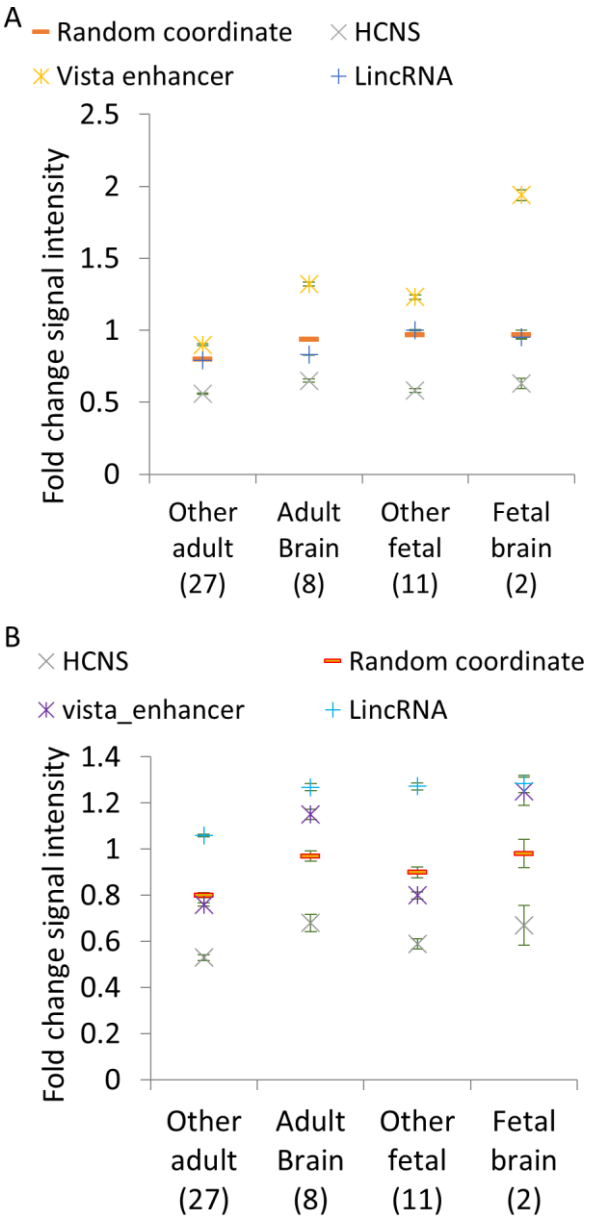

A
0 HCNS
>=1 HCNS
*
*
*
0
50
100
Gene percent non-coding
B
0 HCNS
>=1 HCNS
*
*
*
0
100
200
300
Median distance to the next gene (kbp)

Figure 6.


A
Random coordinate
HCNS
Vista enhancer
LincRNA
Fold change signal intensity
2.5
2
1.5
1
0.5
0
Other adult (27)
Adult Brain (8)
Other fetal (11)
Fetal brain (2)
B
HCNS
Random coordinate
vista_enhancer
LincRNA
Fold change signal intensity
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Other adult (27)
Adult Brain (8)
Other fetal (11)
Fetal brain (2)