Genome Biology and Evolution Advance Access published June 11, 2016 doi:10.1093/gbe/evw132

Emergence and evolution of Hominidae-specific coding and noncoding genomic sequences

Morteza Mahmoudi Saber[1,2], Isaac Adeyemi Babarinde[3,2] , Nilmini Hettiarachchi[3,2] and Naruya Saitou[2,1,3,*]

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

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

[3]Department of Genetics, School of Life Science, Graduate University for Advanced Studies (SOKENDAI), Mishima, 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

Family Hominidae, which includes humans and great apes, is recognized for unique complex social behavior and intellectual abilities. Despite the increasing genome data, however, the genomic origin of its phenotypic uniqueness has remained elusive. Clade-specific genes and highly conserved noncoding sequences (HCNSs) are among the high-potential evolutionary candidates involved in driving clade-specific characters and phenotypes. On this premise, we analyzed whole genome sequences along with gene orthology data retrieved from major DNA databases to find Hominidae-specific genes and HCNSs. We discovered that Down syndrome critical region 4 (DSCR4) is the only experimentally verified gene uniquely present in Hominidae. DSCR4 has no structural homology to any known protein and was inferred to have emerged in several steps through LTR/ERV1, LTR/ERVL retrotransposition, and transversion. Using the genomic distance as neutral evolution threshold, we identified 1,658 Hominidae-specific HCNSs. Polymorphism coverage and derived allele frequency analysis of Hominidae-specific HCNSs showed that these HCNSs are under purifying selection, indicating that they may harbor important functions. They are overrepresented in promoters/untranslated regions, in close proximity of genes involved in sensory perception of sound and developmental process, and also showed a significantly lower nucleosome occupancy probability. Interestingly, many ancestral sequences of the Hominidae-specific HCNSs showed very high evolutionary rates. This suggests that new functions emerged through some kind of positive selection, and then purifying selection started to operate to keep these functions.

Key words: Hominidae, conserved noncoding sequences, lineage-specific genes, DSCR4, accelerated evolution

INTRODUCTION

Family Hominidae which includes humans, chimpanzees, bonobos, gorillas and orangutans is one of the two living families of ape superfamily Hominoidea (see supplementary figure S1, Supplementary Material online). Taxonomically this family belongs to the order Primates. All members of this family have large brains, well-known for their complex social behavior and intellectual abilities. Facial expressions and intricate vocalization play a pivotal role in their behavior. Apart from humans, other species of this family have also shown signs of problem solving (Völter and Call 2012), the phenotype which have not been observed in other closely related species.

The disproportionately enlarged frontal cortex is believed to be mainly responsible for the uniqueness of human cognitive specialization. Several studies comparing human with non-primates and non-Hominidae primates like baboon showed unique disproportionately enlarged frontal cortex in humans (McBride et al. 1999). However, by investigating frontal cortex of several primate species, including all extant hominoids using magnetic resonance imaging (MRI), Semendeferi et al. (2002) showed that human frontal cortex is not disproportionately larger than that of other great apes. Their findings clearly showed disproportionately enlarged frontal cortex to be a unique shared characteristic of the Hominidae family members and a distinctive feature compared to the rest of the species.

Despite the increasing genome data in the past decade, the genetic factors that contribute to the phenotypic uniqueness of Hominidae have remained elusive. Phenotype is the result of a collective network of genes along with other regulatory elements. Recent completion of the whole genome sequencing and gene annotation projects for a diverse variety of species, including the Hominidae family members and their closely related species has provided a strong foundation for comparative genomics analysis of lineage-specific characteristics.

So far, to identify the sequences underlying lineage specific phenotypes within the Hominidae family, the majority of the studies have focused on detecting signatures of positive selection on humans using comparative genomics or genetic variation data produced by the International HapMap Project, Perlegen or 1000 Genomes Project. More than 20 genome-wide scans for positive selection have been performed on the human genome. Although the signals are not generally consistent, strongest signatures of positive selection were found to be on genes involved in host-pathogen interaction, immune response, reproduction (especially spermatogenesis), and sensory perception (Sabeti et al. 2006). Kosiol et al. (2008) studied signatures of positive selection on human-chimpanzee common ancestor as well as the common ancestor of Catarrhini, in which only 7 and 21 genes showed signs of positive selection, respectively. Positively selected genes in that study were also involved in immune response, reproduction and sensory perception. To date, positive selection on protein-coding genes has received the most attention as potential drivers of unique properties observed across the Hominidae family. However, there are other important aspects of the evolution of lineage specific phenotypes which have so far been undervalued in Hominidae studies.

Clade-specific conserved noncoding sequences and clade-specific novel genes are high-potential evolutionary candidates, which may have been involved in driving clade specific

phenotypes. New genes have been revealed to be involved in the evolution of new molecular and cellular functions, developmental processes, sexual dimorphism and phenotypic diversity across species (Chen et al. 2013). Examining the evolutionary period of vertebrates provided evidence for accelerated new gene origination in the recent evolution of hominoids (Zhang et al. 2010). By analyzing expression profiles of human, chimpanzee and macaque, Bleckman et al. (2008) reported that taxonomically restricted genes may play a role in enabling organisms to adapt to changing environmental conditions. If the same scenario holds for clade-specific genes, it implies that the acquisition of new genes by the common ancestor of a particular clade may have played an important role in the development of adaptive novel clade-specific complex biochemical processes.

In addition to genes, conserved noncoding sequences (CNSs) have also been reported to determine lineage specific characteristics. Eight percent of the human genome is speculated to be presently subject to negative selection and likely to be functional (Rands et al. 2014). CNSs are regions within the genome that are evolutionarily conserved despite not coding for proteins. To date, there have been numerous studies on the general features (Babarinde and Saitou 2016; Harmston et al. 2013; Mikkelsen et al. 2007; Pennacchio et al. 2006) and evolutionary dynamics (Faircloth et al. 2012; Lowe et al. 2011; Pennacchio et al. 2006) of conserved noncoding sequences, nearly all of which have proceeded to assign regulatory functions to these conserved genomic elements. CNSs have been reported to be linked to human disease (Visel et al. 2009). In stickleback, loss of a conserved noncoding sequence containing a transcriptional enhancer regulating the pleiotropic Pitx1 gene led to major phenotypic change (loss of pelvic spines) (Chan et al. 2010). In several studies in animals (Babarinde and Saitou 2013; Hiller et al. 2012;

Takahashi and Saitou 2012) and plants (Hettiarachchi et al. 2014), CNSs are also proposed to be involved in lineage specific phenotypes.

Lineage specific duplication is yet another driving force of evolutionary changes across species. Studies of gene family evolution indicate that duplication events are enriched in primates and especially within ancestral branch leading to human and African great apes (Marques-Bonet et al. 2009). Although Hominidae ancestor do not show a strong burst in in duplication or deletion events as the common ancestor of African great apes, duplication and deletion might underlie some of the unique Hominidae lineage-specific traits. Analysis of duplication and deletion activities within Hominidae family have been conducted elsewhere (Marques-Bonet et al. 2009), and since the objective of this study is to discover Hominidae specific unique genomic elements; we did not analyze duplication/deletion events in this paper.

Here we explored the unique genomic elements underlying phenotypes restricted to the Hominidae family by identifying Hominidae-specific orphan genes and Hominidae-specific (HS) highly conserved noncoding sequences. We analyzed whole genome sequences along with gene expression and orthology data retrieved from databases to identify Hominidae specific genes. We also analyzed Hominidae family members' whole genomes along with those of gibbon, rhesus macaque and marmoset to discover Hominidae-specific highly conserved noncoding sequences. Because of the short divergence time between Hominidae members and other closely related species, we used stringent thresholds for identifying HS orphan genes and HCNS to minimize type I error. We found that there is a low proportion of HS protein-coding gene to HS putative regulatory HCNSs, suggesting a likely stronger contribution of regulatory elements than novel genes in defining Hominidae-clade specific phenotypes.

MATERIALS AND METHODS

Retrieving genome sequence and annotations

The human genome annotation was obtained from Gencode 19 (Encyclopedia of genes and gene variants) project (Harrow et al. 2012). For the rest of the species, genomic gene sets were retrieved from Ensembl release 75 FTP website. The repeat masked genome sequences of simians were retrieved from the Ensembl genome database. The genome nucleotide count used for identification of HS HCNSs in chimpanzee; gorilla and orangutan genomes are 2,902,322,413, 2,860,568,349 and 3,091,708,170, respectively. All the genomes are at least 5.6X coverage. The genomic coding coordinates were masked from genome sequences.

Hominidae-specific genes

Homology search for detection of homologous genes: Phylostratigraphic analysis of gene age has been shown to be prone to erroneous gene age underestimation and substantially influenced by length of the encoded protein and its rate of evolution (Moyers and Zhang 2015). Young genes have been shown to be subject of weaker purifying selection (Cai and Petrov 2010) and encode shorter proteins (Wolf et al. 2009). Such characteristics of young genes have made accurate identification of Hominidae specific genes challenging. In the study of Hominoid-specific de novo genes by Xie et al (2012) six novel genes were found to be restricted to human, chimpanzee and orangutan, however, none of them could be identified as ape-specific protein coding gene using phylostratigraphic analysis of current DNA, protein and orthology databases

(see supplementary table S1, Supplementary Material online). To minimize false positive results due to BLAST software limitations (Moyers and Zhang 2015) strict thresholds were used for identification of young genes restricted to Hominidae family.

Experimentally verified human genes derived from Gencode project version 19 were selected as reference and searched against the other three Hominidae members' genes using Ensembl Compara pipeline. Intersection of these three groups represents Hominidae shared genes. Using the same strategy, pairwise orthologous genes were identified between human and all non-Hominidae species available in Ensembl (see supplementary table 2, Supplementary Material online). The genes shared by Hominidae that are not present in outgroup species were searched in INPARANOID (Ostlund et al. 2010), TreeFam (Schreiber et al. 2014), PhylomeDB (Huerta-Cepas et al. 2011) and OrthoDB (Waterhouse et al. 2013) orthology prediction databases and the genes with orthologs in non-Hominidae members were discarded. NCBI MegaBlast was recruited to search the remaining gene sequences in Genbank, EMBL, DDBJ, PDB and RefSeq database. NCBI BlastP was also used to search HS protein-coding genes in UniprotKB database. Any of the gene queries with hits more than 70% coverage and 50% identity in non-Hominidae members was discarded. Summary of Hominidae specific gene detection pipeline is depicted in Figure 1.

Identifying the evolutionary origin of Hominidae specific genomic elements: To identify evolutionary processes leading to the emergence of HS protein coding gene, its orthologous sequences in Hominidae closely related species, namely gibbon, rhesus macaque and marmoset, along with mouse were retrieved from pairwise whole genome lastZ alignments. The whole gene multiple sequence alignment of HS protein-coding gene was constructed using a combination of MISHIMA (Kryukov and Saitou 2010) and Mcoffee (Notredame et al. 2000). Neanderthal

sequence homologous to HS protein coding genes were retrieved as short read alignments (Prufer et al. 2014) and were analyzed using SAMtools. At each position the nucleotide with highest average mapping quality and base quality score were chosen to construct HS protein coding gene in Neanderthal. Shotgun sequencing data of bonobo and baboon genome homologous to HS protein coding were respectively retrieved from NCBI (Prufer et al. 2012) and The Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) and analyzed using Biopython.

To understand whether transposable elements have been involved in the evolution of Hominidae-specific gene, its exonic sequences were searched against transposable elements alignments and hidden Markov models of such elements using Repeatmasker and Dfam database (Wheeler et al. 2013). Analysis of the contribution of transposable elements in formation of human's whole-genome protein coding exons (retrieved from Pfam database) was done using UCSC Galaxy.

Analysis of Selection: Codon-wise and nucleotide-wise analysis of selection using the method described by Haygood et al. (2007) was performed using HyPHY software (Pond et al. 2005). Analysis of selection in human populations was conducted on the 1000 genomes data (Pybus et al. 2014). EDAR and SLC24A5 genes were used as reference for measuring the significance of positive selection. Ectodysplasin A receptor coded by EDAR gene has been shown to be under positive selection in Asian populations (Bryk et al. 2008). Solute carrier family 24 member 5 coded by SLC24A5 affects skin pigmentation and has undergone positive selection in European populations (Lamason et al. 2005). Analysis of selection on these two genes using three classes of population variation based tests, namely allele frequency spectrum (Tajima's D test), linkage

disequilibrium structure (EHH test) and population differentiation (XP-CLR test) (Chen et al. 2010) showed evidence of positive selection within these genes along with their flanking regions.

We measured signals of positive selection on HS protein coding gene along with its upstream and downstream flanking region using Tajima's D test, EHH test and XP-CLR test. Signals of positive selection on EDAR and SLC24A5 genes were used as positive control and were compared with that of HS protein-coding gene.

Hominidae-specific highly conserved non-coding sequences

Setting the percent identity threshold of sequences under purifying selection and neutrally evolving sequences: Since the main objective of this study is to identify Hominidae-unique genomic elements evolved in Hominidae common ancestor ~ 16-19 million years ago (mya) (see supplementary figure S1, Supplementary Material online), it is quite important to accurately differentiate between sequences that are under actual selective constraint and those that just did not have sufficient time to accumulate mutation. This fact is quite important due to the short evolutionary distance of 3.1 mya between the emergence of the closest outgroup species used in this study which is gibbon and the emergence of the most distant member of Hominidae family, orangutan (see supplementary figure S1, Supplementary Material online). To minimize the probability of false positives due to short divergence time, we set the threshold as 100% similarity in conservation and 100bp in length for the identification of sequences under purifying selection.

For accurately determining the threshold for neutral evolution, we compared protein coding sequences' synonymous site variation rate with that of noncoding genomic divergence

rate between species; we considered the former as the depiction of neutral evolution rate in coding sequence and the latter as the neutral evolution rate in noncoding sequences. We retrieved the d_s values of one2one (with one to one correspondence in Ensembl biomart) orthologous protein coding genes for the human genes against gibbon, rhesus macaque and marmoset. To deal with the issue of unreasonably high d_s values we discarded 1% of outliers at the high end and constructed the distribution plot of d_s values. Mode of the plot was considered as the neutral evolution threshold in coding sequences. For setting the neutral evolution rate within noncoding sequences, after running pairwise noncoding blast search, we constructed the distribution plot of sequence divergence values. Mode of the plot was considered as the threshold of neutrally evolving sequences within noncoding sequences.

Homology search for noncoding sequences: After masking coding sequences in each genome, we searched for sequences that are under selective constraint in the Hominidae family. To this end, we used human genome as reference query because of its high quality and availability of genome information, and used BLASTN 2.2.25+ (Altschul et al. 1997) to run whole genome pairwise homology search. The thresholds used were E-value of 10^{-5} and database size of 3×10^9. The E-value cutoff of 10^{-5} with 100bp size minimum length was proven to be efficient thresholds for identification of conserved noncoding sequences within primates (Babarinde and Saitou 2013). Non-chromosomal sequences (such as mitochondrial genome, unmapped DNA and variant DNA) were excluded. In the case of overlapping hits, only the longest hit was retained. Sequences under purifying selection within Hominidae family which had no homologs with conservation level above the neutral evolution threshold in outgroups were assigned as HS HCNS. In order to prevent erroneous identification of HS HCNSs as a result of repeatmasker software errors, UCSC netted chained files were used to map each HS HCNS in gibbon, rhesus

macaque and marmoset unmasked genomes. HS HCNSs with conserved orthologous in interspersed repeats were also discarded.

Evolutionary origin of HS HCNSs: To investigate the evolutionary origins of HS HCSNs, we mapped each of human HS HCNSs to gibbon and rhesus macaque genome sequences and aligned using ClustalW (Thompson et al. 1994). These alignments were concatenated and blocks with gaps were removed. Genetic distances were calculated using MEGA version 6 (Tamura et al. 2013).

Single Nucleotide Polymorphism and Derived Allele Frequency Analyses: We retrieved the final release of phase 3 variant set of 1000 Genomes project. For chimpanzee, gorilla and orangutan, genome variation data were retrieved from the Great Ape Genome Project (Prado-Martinez et al. 2013). Pyliftover and UCSC netted chain files were used to lift chimpanzee, gorilla and orangutan's HCNS coordinates to the corresponding human hg18 coordinates. For each species we retrieved and combined Single Nucleotide Polymorphism (SNP) and insertion/deletion variation data and since the variations were mapped to human genome, we filtered out all variations with allele frequency of 1.0. For each of the three Great Apes, we generated random sequences with the same number and size as HS HCNSs in each species and investigated the coverage of variation in HCNSs and random sequences. For derived allele frequency analysis, we retrieved SNP frequency data of the Yoruba population of Nigeria, from the International HapMap project. The ancestral alleles of SNPs overlapping the HS HCNSs or random sequences were determined using pyliftover and chimpanzee sequence.

HS HCNS flanking region conservation level: We extracted HS HCNSs along with 2000bp upstream and downstream flanking sequences and aligned the sequences using ClustalW. For

each alignment, we made sliding windows of 50bp and step size of 20bp starting from 30bp inside the CNSs and calculated the percent identity in each window. We then calculated the average of the percent identity for each window.

Nucleosome Occupancy Probability: Kaplan et al. (2009) developed a probabilistic model of sequence nucleosome preferences. Considering dinucleotide signals along with favored and disfavored pentamer sequences in known nucleosome, this model produces a nucleosome occupancy score for each nucleotide of the subject sequence. Using version 3 of the nucleosome position prediction program, nucleosome occupancy probability for HS HCNSs were calculated considering 4000 bp region at upstream and downstream starting from the center of the HS HCNSs. The average nucleosome occupancy probability was calculated for each nucleotide site of the total 8,000 bp along the length of sequences. The same procedure was carried out for random sequences of the same number and same size. Statistical significance was calculated using t-test for HCNS sites scores.

To confirm lower nucleosome occupancy of HS HCNSs, we retrieved genome binding/occupancy profiling data derived by high throughput sequencing and MNase-seq nucleosome positioning experiments from ENCODE/Stanford/BYU using UCSC (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhNsome/Gm12878Sig.bigWig). Average nucleosome occupancy score for HS HCNSs and flanking regions were calculated considering 4000 bp region at upstream and downstream starting from the center of the HS HCNSs.

H3K9 methylation is the mark of heterochromatin regions. To further confirm the underrepresentation of HS HCNSs within heterochromatin regions, we retrieved H3K9me

mapping data from ENCODE project and analyzed HS HCNS overlap with H3K9me histone mark compared to random sequences.

Genomic distribution: We retrieved the annotations of the human genome from Gencode project and parsed each gene into regions, intersecting over alternative transcripts and splices, so that what are termed 'UTR' and 'intronic sites' are such sites with respect to all known transcripts and splices. We defined promoter region as the region within 1000bp upstream of a transcription start site. We then found HCNSs that are located on UTR, promoter, intronic and intergenic regions. We also calculated the fractions of UTRs, introns, promoters and intergenic sequences in the human whole-genome. Chi square test was used to analyze the significance of fraction differences.

Gene ontology analysis: We retrieved the coordinates of protein-coding genes from Gencode project. For HS HCNSs, we retrieved the list of genes found upstream and downstream of each HCNS. The gene that lies closest to a particular HCNS was considered as the likely target gene. If a HCNS was found inside a gene (including introns and UTR), the gene in which it resides was considered as the likely target gene. The likely target gene is with respect to the human reference genome. We checked the functional analysis of HS HCNSs using Panther 9.0 (Mi et al. 2010). P-value corrected for multiple testing using Bonferroni correction was calculated. Unless otherwise stated, all scripts used for these analyses were written by one of us using Python(Van Rossum 2007) or R (R development core team 2010) and are available upon request.

Tissue-specificity of HS HCNSs: To investigate whether HS HCNSs do have unique properties in tissue-specific manner, we retrieved Dnase, chipseq and histone modification data for all tissues from Epigenome roadmap project (http://www.roadmapepigenomics.org/data/). The

average score was calculated for each 400-bp window along the length of HS HCNS and flanking regions for the total of 18,500 bp. Standard error value for each window was calculated using SciPy (http://www.scipy.org/).

RESULTS

Down syndrome critical region of 4, Hominidae specific Orphan gene

By analyzing the DNA, protein and orthology databases, Down syndrome critical region of 4 (DSCR4) gene, on chromosome 21 discovered by Nakamura et al. (Nakamura et al. 1997) via EST mapping, was found to be the only annotated HS protein coding gene. DSCR4 is an experimentally known gene, present in Ensembl, VEGA and consensus CDS protein set (CCDS) databases, and codes one known 117 amino-acid residue long polypeptide, one putative 127 amino-acid residue long polypeptide and a single 79 nonsense mediated decay transcript. However, although the 117 known amino acid long transcript is annotated in chimpanzee and orangutan, this transcript is missing in gorilla genome annotation. Close examination of the gorilla genome sequence revealed that the 117 amino acid long transcript could be constructed using gorilla-human orthologous sequences (see supplementary figure S2a, Supplementary Material online); but it was not annotated due to limitations in the annotation algorithm. Analysis of Neanderthal and bonobo genome sequence homologous to human DSCR4 sequence showed that the complete ORF could be successfully constructed in these two genomes indicating potential expression of DSCR4 in all members of Hominidae whose genomes have been sequenced (see supplementary figure S2a, Supplementary Material online);. Although the expression data for placental tissue where DSCR4 is mainly expressed is not available for great

apes, expression analysis has detected DSCR4 polyadenylated RNA in bonobo and chimpanzee testis as well as gorilla testis and heart (Brawand et al. 2011).

Proteins are generally composed of one or more functional domains. Combination of existing domains within a protein provides insights into the function of the protein. PFam database (Finn et al. 2014) contains high quality, manually curated protein domain entries named PFam-A along with automatically generated domain entries produced by Automatic Domain Decomposition Algorithm (ADDA) named Pfam-B. Searching DSCR4 protein sequence within PFam showed no signs of homology to any known or predicted protein domain family. Examining uniprotKB database also revealed no homology to any existing protein sequence in any species other than Hominidae family. However, significant homology was found with yet uncharacterized proteins in all other members of Hominidae.

No experimental 3D structure analysis has been undertaken for DSCR4 protein, nor were there any experimental structures with >90% sequence identity to DSCR4 in protein 3D databases. However, the secondary structure of DSCR4 based on Chou and Fasman algorithm (Chou and Fasman 1974) suggests the existence of potential α-Helices and β-Sheets (see supplementary figure S3a, Supplementary Material online). Constructing the 3D structure by protein model portal (Arnold et al. 2009) and I-TASSER (Zhang 2008) also showed evidence for the existence of α helices and β sheets in the protein coded by DSCR4 (see supplementary figure S3b and S3c, Supplementary Material online).

Analyzing High coverage short-read data of gibbon genome (Carbone et al. 2014) revealed that DSCR4 exon 3 coding sequence is partially missing in all sequenced gibbon individuals. This result indicates the possibility of lineage-specific deletion in gibbon genome

sequence orthologous to human DSCR4 gene. This observation is the sole reason for our shift to macaque genome as template for evolutionary analysis of the origin of DSCR4 gene (Figure 2a).

DSCR4 is separated by a 92 bp sequence from the DSCR8 gene. The 92-bp separator sequence is part of a bidirectional promoter which initiates transcription from both of these genes. While DSCR4 is limited to Hominidae family, DSCR8 is present in Hominoidea, old world and new world monkeys. Multiple sequence alignment of DNA sequences corresponding to DSCR4/DSCR8 gene and their shared promoter in Hominidae family along with closely related species and mouse suggests multi-step evolution of DSCR4.

Movement and accumulation of transposable elements (TE) have been a major force shaping the genes of almost all organisms (Feschotte and Pritham 2007). Investigating the role of TEs in evolution of human protein coding genes revealed 1.1% of all human protein coding exons to be at least partly derived from TEs (see supplementary figure S4, Supplementary Material online). TEs also played a major role in the evolution of DSCR4/DSCR8 genes. The first three exons of both DSCR4 and DSCR8 genes have been derived at least partly from transposons (Figure 2c).

By analyzing pairwise whole genome alignment data of Amniote lastZ (http://www.ensembl.org/info/genome/compara/analyses.html), evolution of DSCR4 could be mainly classified in three evolutionary periods. Period 1) LTR79 retrotransposition that took place in the common ancestor of mammals more than 100 million years ago. This transposition formed DSCR4's exon 3 ancestral sequences. Period 2) During the evolution of common ancestor of primates at 29-45 million years ago, three independent retrotranspositions by MLT2C1, LTR16A and LTR9 led to the formation of DSCR4's exon 2, exon 1 along with

DSCR4/8 shared bidirectional promoter (see supplementary TableS3, Supplementary Material online). Analysis of the core promoter region of DSCR4/8 bidirectional promoter also reveals that the DSCR4/8 bidirectional promoter region has retrotransposed and activated at this period (see supplementary figure S5, Supplementary Material online). Period 3) The final ORF-enabling mutation was a GC transversion at DSCR4 exon 3 that formed the stop codon TGA (Figure 2b, supplementary figure S2b, Supplementary Material online). This transversion, which took place in the common ancestor of Hominidae 15-19 million years ago, completed the formation of the DSCR4 gene.

Analysis of selection

The 117 amino acid long, experimentally known transcript of DSCR4 along with its orthologous sequences in other Hominidae species were used to examine signatures of selection. Codon-wise analysis of selection using Hyphy package showed no statistically significant signs of selection on any of the codons. Nucleotide-wise examination of selection also didn't reveal any positively selected sites in promoter region. Population-based tests of selection (Tajima's D, XP-EHH and XP-CLR) also showed no consistent sign of selection in any of European, Chinese or African populations (see supplementary figure S6, Supplementary Material online). Analysis of the nonsynonymous and synonymous substitution rate also didn't reveal evidence for strong purifying selection on this gene. These results are consistent with previous findings stating that young genes are subject to weaker purifying selection (Cai and Petrov 2010).

Highly conserved noncoding sequences

Gorilla diverged from the common ancestor of Homo and Pan Genera 8.8 mya and later Homo and Pan diverged about 6.9 mya. The common ancestor of Hominidae diverged from Hylobatidae 18.8 mya and 3.1 million years later, orangutan, the most distant member of Hominidae family emerged (Hedges et al. 2015). Such short divergence times within family members and between Hominidae family and phylogenetically close species have made discerning HS functional noncoding sequences under purifying selection from neutrally evolving sequences a challenging objective.

The crucial parameter for identifying linage-specific CNSs in closely related species is the nucleotide identity threshold of the sequences evolving neutrally and sequences evolving under purifying selection. Due to short divergence times, false positive results are of high concern and thresholds are set in a way that takes special care of type I errors. Since the majority of noncoding DNA sequences are assumed to be under neutral evolution (Kimura 1983; Saitou 2014), we considered the mode of the noncoding sequence alignment plot as the neutral evolution threshold. To verify the authenticity of this threshold, we also analyzed the neutral substitution rate in protein coding sequences. We constructed the synonymous substitution rate plot between human and three closest outgroup species, namely, gibbon, rhesus macaque and marmoset (see supplementary table S2 for scientific nomenclature, Supplementary Material online). In several studies, a number of synonymous sites in protein coding genes have been shown not to be strongly following neutral fashion. Some of these synonymous sites have been shown to be under weak selection constraint (Chamary et al. 2006) and may affect mRNA stability or splicing. On this premise, it is expected that protein coding's ds-based neutral divergence plot to have similar distribution as the noncoding sequence identity-based plot but with a weak skew towards the conserved end. This pattern was indeed observed in pairwise

comparison of human and all three outgroups (see supplementary figure S7 and S8, Supplementary Material online) which suggests that our thresholds for neutrally evolving sequences are accurate. We filtered out all HS HCNS with orthologous sequences in any outgroups with divergence levels lower than neutral evolution threshold. Using this strategy we identified 1,658 HS HCNSs (HS HCNS coordinates, sequences and multiple alignment results are provided in supplementary material files, Supplementary Material online, in FASTA format). Length distributions of Hominidae specific HCNSs are shown in supplementary figure S9 (see supplementary figure S7, Supplementary Material online). Probability analysis using whole genome blast hits frequency data (see supplementary figure S10, Supplementary Material online) showed that the frequency of sequences meeting all these conditions by chance is 3.88×10^{-8}. Since the total number of pairwise blast hits in each of reference genomes pairs are much less than 3.88×10^{8}, it is extremely unlikely for HS HCNSs to be only cases of the outliers of neutral evolution.

Functional analysis of HS HCNSs

Genetic variation is a suitable indicator of selective constraint on a sequence. We investigated the frequency of SNPs, deletions and insertions overlaid on the HS HCNSs in human and great apes using 1000 genome and great apes genome project data. The frequency of polymorphisms (SNP density per site: 2.4E-2, 8.6E-3, 5.3E-3 and 5.0E-3 for human, chimpanzee, gorilla and orangutan, respectively) in HS HCNSs are significantly lower than that of random sequences of the same number and same size (2.9E-2, 1.2E-2, 8.5E-3 and 7.5E-3) in all members of the Hominidae family (Figure 3a).

Derived allele frequency (DAF) analysis is another test of functionality of a sequence. Purifying selection is considered as the main evolutionary force to prevent conserved noncoding sequences from accumulating mutations. We found a higher proportion of HS HCNSs having lower derived alleles than random expectation. This suggests that HS HCNSs are under purifying selection (Figure 3b). At the level of DAF<0.1, HS HCNS showed a significant excess of rare-derived polymorphisms compared to random expectations (fisher test p-value: 0.004) and by comparing all categories we noticed a significant shift in HS HCNS polymorphisms' allele frequency toward rare allele frequencies (chi squared p=0.001).

Are HS HCNSs located on local mutation cold spot regions in all the Hominidae family? To address this question, we checked the conservation level of the HS HCNS flanking regions. Figure 3c shows the pattern of conservation within HS HCNSs with up to 1770bp up- and down-stream flanking regions. For random sequences, unfiltered alignments of at least 2000 bp long were used. The conservation plot indicates that only the HS HCNSs are highly conserved, indicating that they are under the strong constraints, relative to their flanking regions. We also investigated genetic variation frequency at upstream and downstream regions within the same length of each HS HCNS that do not overlap with known coding sequences. The genetic variations at HS HCNS up- and down-stream flanking regions are not significantly different from random noncoding sequences. However, their variation was significantly higher than that in HS HCNSs (Figure 3a). These results indicate that HS HCNSs are not located in mutation cold spots.

Evolutionary origin of HS HCNSs

How did HS HCNSs emerge? We need to compare outgroup species sequences of HS HCNSs to answer this question. Using whole genome mapping data, 32% (527) of HS HCSNs were mapped to gibbon and rhesus macaque genomes while the rest could not be mapped to these outgroup species genomes. We thus examined 527 multiple alignments of three sequences (HS HCNS, gibbon and rhesus macaque; see HS_HCNS_alignments.txt, Supplementary Material). Length size distribution analysis revealed that average length difference of the mapped sequences in gibbon and rhesus macaque genomes from HS HCNSs is significantly higher than that of random sequences (See supplementary figure S11, Supplementary Material online).

We also estimated substitution rates (/site/year) at three branches (α, β, γ) of Figure 4a for HS HCNSs orthologous and ancestral sequences using mapped gap-removed alignments. Divergence time estimates shown in Supplementary Figure S1 are used for rate estimations. We are particularly focused on branch on branch α of the phylogenetic tree shown in Figure 4a, because this branch corresponds to the common ancestor of Hominidae after divergence of the common ancestor of Hominidae and Hylobatidae (gibbons). The mean rate of nucleotide substitution at branch α was 5.5 x 10^{-9} (Figure 4a), which is five times higher than that (1.1 x10^{-9}) of the neutrally evolving genomic regions. Interestingly, the substitution rates for branches β and γ (2 x 10^{-9} and 1.9 x 10^{-9}, respectively) were also higher than the neutral rate (Figure 4a).

A very high mean substitution rate for branch α of Figure 4a suggests the existence of positive selection at this branch followed by purifying selection in the later Hominidae lineages. We therefor examined the distribution of substitution numbers at branch α for those 527 HS HCNSs, as shown in blue bars of Figure 4b. Red bars of Figure 4b are corresponding to distribution of 1,658 randomly chosen sequences, which are considered to be under pure neutral

evolution. Distribution patterns of blue and red bars are clearly different, and a total of 97 (18% of 527) HS HCNSs showed the rates higher than 0.02, the largest branch length value observed for some purely neutral genomic regions. This suggests that at least 18% of HS HCNSs experienced some kind of positive selection which enhanced their substitution rates.

We found that 527 HS HCNSs were orthologous both to gibbon and rhesus macaque sequences. However, there were 1,001 HS HCNSs whose orthologs were found only in gibbons. In this case, without rhesus macaque, we cannot distinguish branches α and β. Yet, if the average of these two branches again showed elevated substitution rates, our finding based on only 527 HS HCNSs can be strengthened. In fact, the mean substitution rate for branches α and β combined for 1,001 HS HCNSs was 2.3 $x10^{-9}$(Figure 4c),. If we subtract the contribution of branch β from this rate, we obtain the new substitution rate estimate (3.9 $x10^{-9}$) for branch α. This value is slightly lower than that 5.5 x 10^{-9} for 527 HS HCNSs, but still more than three times higher than the neutral rate. This confirms an elevated nucleotide substitution rate at branch α. Branch length distribution of those 1,001 HS HCNSs is shown as blue bars in Figure 4c with random regions shown in red bar. Branch lengths for HS HCNSs are clearly shifted to larger ones compared to purely neutral ones.

These results indicate that insertions and deletions along with accelerated evolution in the common ancestor of Hominidae are the main evolutionary changes leading to the formation of HS HCNSs. We examined neighboring genes of these HS HCNSs with very high substitution rate for branch α of Figure 4a. Supplementary table S4 of supplementary Material online lists these genes. NADPH oxidase (NOX) 3 is a member of the NOX/dual domain oxidase family with 50 fold overexpression in inner ear. Nox3 is indispensable gene in formation of otoconia within inner ear (Paffenholz et al. 2004). Sall3 is a member of splat gene family. Mutations in

members of this family have been associated with several congenital disorders (Sweetman and Munsterberg 2006). ABCD4 is a member of the superfamily of ATP-binding cassette (ABC) transporters involved in peroxisome biogenesis and adrenoleukodystrophy (ALD) disorder (Matsukawa 2011).The cocaine- and amphetamine-regulated transcript peptide (CARTPT) is involved in reward and feeding behavior and function as a psychostimulant (Lohoff et al. 2008). TPRXL and MAGEA1 which are involved in embryonic development are among the likely target genes of highly conserved HS HCNSs.

Genomic distribution of HS HCNSs

We investigated the genomic location of each HCNS to examine whether there is any general trend in the distribution of HS HCNSs. HS HCNSs were categorized into four classes: intergenic, intronic, UTR and promoteric. Distribution of HS HCNS within these categories is shown in Table 1. Their distribution significantly differs between HS HCNSs and rest of the genome (P value = 2.2E-16, Chi Square test). The fraction of HS HCNSs residing in introns, UTR and promoter regions of the human genome are significantly higher than those of the whole genome. The fractional increments are especially prominent in UTR (more than three times higher than the whole genome fraction) and promoter regions (almost two times higher than the whole genome). The increased proportions of HS HCNSs within UTR and promoter regions are consistent with previous findings of the genomic distribution of CNSs in primates (Takahashi and Saitou 2012; Babarinde and Saitou 2013) who reported the notably increased fraction of CNSs in UTR and promoter regions.

Prediction of nucleosome positioning

Nucleosome positioning with respect to DNA plays a crucial role in transcription regulation. Packing DNA in nucleosomes can limit the accessibility of the sequences and low nucleosome occupancy is considered as an important feature of transcription factor binding site (TFBS) (Miele et al. 2008; Schones et al. 2008). We computed the nucleosome position probability of HS HCNSs and their flanking regions using the nucleosome prediction probability algorithm developed by Kaplan et al. (2009), 4000bp region from the center of each HS HCNS at both upstream and downstream. A clear drop in nucleosome occupancy was observed directly overlapping with the center of HCNSs indicating the possibility of nucleosome depletion within the HS HCNS regions (Figure 5a). The nucleosome occupancy probability within HCNS regions was significantly lower than the random expectations (p value = 4.149E-40, t-test). This result was further confirmed using experimental genome occupancy profiling data derived by high throughput sequencing and MNase-seq nucleosome positioning experiments (Figure 5b). Analysis of H3K9me heterochromatin mark also revealed significant underrepresentation of HS HCNSs within H3K9me-marked heterochromatin regions (Figure 5c). Babarinde and Saitou (2013) discussed the possibility of their low-GC mammalian CNSs as also found for HS HCNSs (see supplementary figure S12, Supplementary Material online) to nucleosome occupancy, and Kenigsberg and Tanay (2013) found a similar nucleosome positioning pattern in Drosophila CNSs.

Gene ontology analysis

We considered the closest genes to HS HCNSs as the likely target gene on the premise that regulatory elements reside in close proximity with the gene they regulate, and examined the enrichment of biological process of HS HCNSs using PANTHER. Ninety seven percent of HS HCNSs are located within 1Mb of their nearby protein coding gene, the range in which most of gene regulatory elements are located (See supplementary figure S13, Supplementary Material online). This observation is significantly different from the random expectation (p value: 1e-05, empirical chi-square test; for null distribution, see supplementary figure S14, Supplementary Material online). However, growing number of human genetic conditions are being found resulting from mutations in regulatory elements located more than 1 Mb away from the gene they regulate (Ghiasvand et al. 2011; Symmons and Spitz 2013). As a result, the possibility of the remaining 3% of HS HCNSs being regulatory elements of their nearby genes could not be ruled out.

Table 2 shows the top categories in which HS HCNSs are enriched. The gene enrichment analysis for the HS HCNS target genes indicate that sensory perception of sound has the highest fold enrichment among significantly overrepresented biological function categories. PCDH15 and cdh19 are two auditory critical genes located in close proximity of HS HCNSs. Protocadherin 15 (PCDH15) mutations in which causes inherited deafness called usher1F syndrome (Sotomayor et al. 2012) is the likely target gene of two HS HCSNs found in this study. PCDH15 plays a crucial role in mechanotransduction that is important for sound characterization in the inner ear. Cadherin 19 (cdh19) is another likely target gene of two HS HCSNs, down-regulation of which has been linked to the development of cholesteatoma, an expanding destructive epithelial lesion within the middle ear (Klenke et al. 2012). Analysis of the gene ontology of random sequences did not show any enriched category of biological functions.

Consistent with previous analysis of conserved noncoding sequences (Babarinde and Saitou 2013;Takahashi and Saitou 2012), genes involved in developmental process are also mainly enriched as likely target genes of HS HCNSs (Table 2). Fox and Sox gene families play critical roles in the process of development. The FOX gene family genes are involved in developmental processes, organogenesis and speech acquisition (Hannenhalli and Kaestner 2009). Several members of this family, including FOXD4L2, FOXD4L5, FOXD4L4, FOXK1, FOXE1, FOXR2, FOXI2, FOXG1 and FOXN3 are in close proximity with HS HCNSs. Among the other likely target genes of HS HCNSs are SOX1, SOX5 and SOX11. These genes are members of the SOX gene family that is also involved in regulating several crucial aspects of development (Prior and Walter 1996).

Brawand et al. (2011) analyzed the evolution of gene expression in mammalian organs and identified numerous genes with expression switch on the branch connecting great apes and macaque. These reported genes are the target of 158 HS HCNSs (see supplementary Table S5, Supplementary Material online) with enrichment of the expression in cerebellum that is associated with language processing, learning, addiction and motor functions (Strick et al. 2009). This result is significantly different from random expectation (p value: 0.00767, empirical chi-square test) (see supplementary figure S14 for null distribution, Supplementary Material online). These results indicate the possibility of the evolution of HS HCNSs as the regulatory elements responsible for gene expression switches contributing to specific organ biology of Hominidae family.

Analysis of tissue-specificity of HS HCNSs also revealed that HS HCNSs have respectively intensified average chromatin immunoprecipitation signal and H3K4me3 epigenetic mark within fetal brain and placenta compared to flanking regions (see supplementary figure S15,

Supplementary Material online). H3K4me3 is associated with active promoter regions. These data are in line with overrepresentation of HS HCNSs in promoter regions and enrichment of developmental process in Gene ontology analysis of likely target genes of HS HCNSs. These results give evidence for the likely role of HS HCNSs as regulatory elements mainly involved in development which have been suggested to play key roles in phenotypic diversity across species (Carroll 2000).

Comparing properties of HS HCNSs with human genome regions under accelerated evolution (HARs) identified by Pollard et al. (2006) and conserved noncoding sequences under accelerated evolution in human (HACNs) identified by Prabhakar et al. (2006) revealed no significant overlap (see supplementary table S6, Supplementary Material online). These results were expected due to significant difference not only in the direction of evolutionary changes but also in the time intervals in which HARs, HACNs and HS HCNSs were under action of evolutionary forces, indicating age-dependent properties of conserved noncoding sequences, as also suggested by Babarinde and Saitou (2013).

Analysis of lincRNA from Ensembl, enhancer sequences from Fantom project (http://fantom.gsc.riken.jp/data/) and GWAS-tagged SNPs from NHGRI-EBI (http://www.ebi.ac.uk/gwas/) also showed neither significant overrepresentation of HS HCNSs in lincRNA or enhancer sequences nor enrichment of GWAS-tagged SNPs suggesting that the mode of action of the majority of these elements under strong purifying selection are yet to be fully understood

DISCUSSION

Unraveling the molecular mechanisms underlying unique cognitive specialization shared by humans and great apes such as language learning and problem solving ability has been of particular interest to researchers from a broad range of scientific fields and so far, several comparative genomic studies have been conducted to explore the genomic sequences underlying human-specific phenotypes (Sumiyama and Saitou 2011; Pollard et al. 2006; Prabhakar et al. 2006). However, due to unavailability of high throughput sequencing technology and whole genome data for apes until the first decade of new millennium, molecular evolutionary genetics has not progressed as much in deciphering underlying genomic components of Hominidae-specific unique phenotypes. Emergence of novel genes has been linked to appearance of novel developmental and behavioral phenotypes in several species. Examples include dry-nosed primate-specific insulin-like 4 (Arroyo et al. 2012), Arabidopsis-specific CYP84A4 (Weng et al. 2012) and Drosophila-specific Xcbp1 genes (Chen et al. 2012) which respectively affect fetal development, pollen development and foraging behavior. Although emergence of lineage-specific genes have been shown to be a major contributor to adaptive evolutionary innovation, there are still gaps in evolutionary genomics in explaining lineage specific characteristics and phenotypes which could not be answered by mere presence or absence of a particular set of genes. Within several kingdoms of species, lineage specific conserved noncoding sequences have been suggested to be involved in spatiotemporal regulation of gene expression (Babarinde and Saitou 2013; Hettiarachchi et al. 2014; Janes et al. 2011). Although the specific functions of these conserved elements are mainly unknown, functional analyses have shown CNSs to be under purifying selection and enriched in close proximity of genes involved in developmental process in mammals and amniotes (Babarinde and Saitou 2013; Janes et al. 2011; Takahashi and Saitou 2012). Since phenotypic evolution has been suggested to be primarily mediated by genes

involved in developmental process (Nei 2007), CNSs could be considered as a high-potential candidate for filling the knowledge gap in elucidating the molecular basis of phenotypic diversity across lineages.

In this study we identified one Hominidae-specific protein-coding gene and 1,658 CNSs originated in the common ancestor of Hominidae. Since comprehensive analysis of gene expression has not yet been uniformly accomplished for Hominoids and monkeys, projection of human's experimentally verified genes in great apes and monkeys were used as the sets of existing genes. We defined HS HCNSs as homologous regions with at least 100 bp length and conservation level of 100% within Hominidae members with no orthologous sequence with conservation level above neutral evolution threshold in non-Hominidae simians. Although it is possible that some putative genes with undetected expression or conserved noncoding sequences with less degree of conservation are functional, we assume that our conservative approach for HS novel gene and HS HCNS identification screens only genomic elements that are functionally important to Hominidae.

Down syndrome critical region (DSCR) has long been known to include genes involved in higher brain functions. This region has also been proposed to be responsible for the mental retardation phenotype observed in Down syndrome which is characterized by verbal short-term memory, spatial learning and deficits in speech and language (Olson et al. 2007). The critical importance of this region is consistent with our discovery that the only experimentally known Hominidae specific protein coding gene is placed in the DSCR region. Although the fact that this protein is mainly derived from transposable elements with no homology to any family of proteins raises doubts about the functionality of this protein, there are numerous evidences at RNA and protein level, indicating the functionality of this gene. These evidences are: i) higher absolute

expression values compared to flanking conserved genes (see supplementary figure 16, Supplementary Material online), ii) tissue-specific expression (Uhlén et al. 2015), iii) epigenetic marks for active regulatory region (see supplementary figure 17, Supplementary Material online), iv) being a binding site of several transcription factors (see supplementary figure 17, Supplementary Material online), v) the likely existence of secondary structures in DSCR4-coded protein (see supplementary figure S2, Supplementary Material online) and vi) acting as a fetal epigenetic marker for detection of down syndrome (Du et al. 2011). These evidences indicate active regulation and expression of DSCR4, which in turn suggests this gene to be a functional element in humans. Further functional analysis of DSCR4 might lead to better understanding of the genomic pathways involved in development of higher brain functions shared by Hominidae members and affected in Down syndrome.

Spatiotemporal regulation of gene expression has long been reported to be important in phenotypic diversity (Carroll 2000). The conservation level, coverage of polymorphism as well as DAF analysis supports that the potential Hominidae specific regulatory elements identified as HS HCNSs are under functional constraint and may be involved in regulatory functions restricted to members of Hominidae family. Nucleosome positioning analysis showed low nucleosome occupancy probability in HS HCNSs implying that these elements have lower probability to form nucleosomes. The finding by Bai and Morozov (2010), stating that regulatory sequences are more nucleosome-depleted, gives additional support to the hypothesis that HS HCNSs is functional and involved in transcriptional regulation of their target genes.

According to our finding, insertions and deletions along with accelerated substitution rate in the Hominidae common ancestor are the main driving force for the evolution of HS HCNSs. Lineage-specific accelerated evolution in noncoding sequences have been proposed to be

involved in evolution of species, potentially through lineage-specific changes in gene regulation (Bird et al. 2007). Evidence of prominent accelerated evolution on mappable HS HCNS ancestral sequences followed by strong purifying selection found in our study suggests that HS HCNSs have played key role in the emergence of Hominidae as a unique lineage among primates.

Gene ontology analysis carried out for HS HCNSs suggests HS HCNSs to be located close to genes mainly involved in developmental processes. Previous genome analyses of animals and plants have also demonstrated CNSs to be located near genes involved in developmental process. These findings agree with the idea that differences in the cis-regulatory elements involved in developmental process have a central role in intraspecific variation and phenotypic diversity across species (Carroll 2000) and gives further evidence for the contribution of HS HCNSs to the characteristics uniquely shared by Hominidae members. One interesting feature to note is the highest fold enrichment of likely target genes of HS HCNSs for the sensory perception of sound. Unlike the enrichment for developmental process which is shared between conserved elements within several lineages, sound sensory perception is uniquely overrepresented in HS HCNSs target genes. Sensory perception of sound is defined as the series of events required for an organism to receive an auditory stimulus, convert it to a molecular signal, and recognize and characterize the signal (Mi et al. 2013). Considering the unique sophisticated linguistic abilities observed within Hominidae (Miles and Miles 1990; Patterson and Linden 1981; Savage-Rumbaugh et al. 1985), one plausible reason to explain this observation is that HS HCNSs might be involved in development of unique sound sensory systems required for recognition and characterization of intricate communicative sounds used by humans and great apes.

Comparing genome wide analyses of primate specific genes (measured as transcriptional unit) and primate specific gene regulatory elements (measured as primate specific highly conserved noncoding sequences) shows that the ratio of lineage specific protein coding genes to lineage specific highly conserved regulatory elements is only 0.007 (59/8198) (Takahashi and Saitou 2012; Tay et al. 2009). The HS protein coding gene to HS HCNS ratio, 0.0006 (1/1658) found in this study, is more than 1/10 lower than the already low primate specific gene to HCNS ratio. These results are consistent with the notion that the morphological diversity is mainly accounted for by differences in regulatory elements (Carroll 2000), suggesting regulation alteration of existing protein-coding genes might have played a more significant role in Hominidae evolution than emergence of novel genes.

In this study, we identified one HS protein coding gene and 1,658 potential regulatory highly conserved noncoding sequences originated in the common ancestor of Hominidae clade members 15 to 19 million years ago. Although young, tissue-specific genes are of high medical relevance, functional characterizations of human genes have been biased against these genes (Hao et al. 2010). The Hominoide specific protein coding gene DSCR8 and Hominidae specific protein coding gene, DSCR4, are examples of such bias which despite being placed on medically important region, Down syndrome critical region of chromosome 21, their structure and function are not studied yet. In this study, HS HCNSs are shown to be under accelerated evolution in the Hominidae common ancestor, overrepresented in promoters, untranslated regions and in close proximity of genes involved in sensory perception of sound and developmental process. They also showed a significantly lower nucleosome occupancy probability. This study provides candidates of genes and regulatory elements which are expected to hold the key to the understanding of the phenotypic uniqueness shared by human and great apes, via mechanisms

majority of which are yet to be fully understood. Experimental verification of these elements is expected to shed light on the lineage specificity of Hominidae.

Acknowledgements

The authors thank Dr. Timothy A Jinam, for his 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 N.H. and by a grant-in-aid for scientific research from MEXT of Japan to N.S.

Literature cited:

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

Arnold K, et al. 2009. The Protein Model Portal. *J Struct Funct Genomics.* 10:1-8.

Babarinde IA, Saitou N. 2016. Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics.*Mol Biol Evol.* Doi: 10.1093/molbev/msw058

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

Bai L, Morozov AV. 2010. Gene regulation by nucleosome positioning. *Trends Genet.* 26:476-483.

Bird CP, et al. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol.* 8:R118.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* 4:e1000271.

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

Bryk J, et al. 2008. Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. *PLoS One.* 3:e2209.

Cai JJ, Petrov DA. 2010. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol Evol.* 2:393-409.

Cantalupo C, Hopkins WD. 2001. Asymmetric Broca's area in great apes. *Nature.* 414:505.

Carbone L, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature.* 513:195-201.

Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell.* 101:577-580.

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

Chan YF, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science.* 327:302-305.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res.* 20:393-402.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet.* 14:645-660.

Chen S, et al. 2012. Frequent recent origination of brain genes shaped the evolution of foraging behavior in Drosophila. *Cell Rep.* 1:118-132.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry.* 13:222-245.

Du Y, et al. 2011. Hypomethylated DSCR4 is a placenta-derived epigenetic marker for trisomy 21. *Prenat Diagn.* 31:207-214.

Faircloth BC, et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol.* 61:717-726.

Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet.* 41:331-368.

Finn RD, et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211-222.

Ghiasvand NM, et al. 2011. Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. *Nat Neurosci.* 14:578-586.

Hannenhalli S, Kaestner KH. 2009. The evolution of Fox genes and their role in development and disease. *Nat Rev Genet.* 10:233-240.

Hao L, et al. 2010. Human functional genetic studies are biased against the medically most relevant primate-specific genes. *BMC Evol Biol.* 10:316.

Harrow J, et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22:1760-1774.

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.

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.

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.

Hiller M, Schaar BT, Bejerano G. 2012. Hundreds of conserved non-coding genomic regions are independently lost in mammals. *Nucleic Acids Res.* 40:11463-11476.

Hopkins WD, Marino L, Rilling JK, MacGregor LA. 1998. Planum temporale asymmetries in great apes as revealed by magnetic resonance imaging (MRI). *Neuroreport.* 9:2913-2918.

Huerta-Cepas J, et al. 2011. PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* 39:D556-560.

Janes DE, et al. 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. *Genome Biol Evol.* 3:102-113.

Kaplan N, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* 458:362-366.

Kenigsberg E, Tanay A. 2013. Drosophila functional elements are embedded in structurally constrained sequences. *PLoS Genet.* 9:e1003512.

Kimura M. 1983. *The Neutral Theory of Molecular Evolution* (Cambridge University Press).

Klenke C, et al. 2012. Identification of novel cholesteatoma-related gene expression signatures using full-genome microarrays. *PLoS One.* 7:e52718.

Kosiol C, et al. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4:e1000144.

Kryukov K, Saitou N. 2010. MISHIMA--a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data. *BMC Bioinformatics.* 11:142.

Lamason RL, et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* 310:1782-1786.

Li CY, et al. 2010. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput Biol.* 6:e1000734.

Marques-Bonet T, et al. 2009. A burst of segmental duplications in the genome of the African great ape ancestor. *Nature* 457:877-881.

Matsuno M, et al. 2009. Evolution of a novel phenolic pathway for pollen development. *Science* 325:1688-1692.

McBride T, Arnold SE, Gur RC. 1999. A comparative volumetric analysis of the prefrontal cortex in human and baboon MRI. *Brain Behav Evol.* 54:159-166.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc.* 8:1551-1566.

Mi H, et al. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 38:D204-210.

Miele V, Vaillant C, d'Aubenton-Carafa Y, Thermes C, Grange T. 2008. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* 36:3746-3756.

Miles H, Miles H. 1990, The cognitive foundations for reference in a signing orangutan “Language” and intelligence in monkeys and apes (Cambridge University Press)

Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature* 447:167-177.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol.* 32:258-267.

Nakamura A, Hattori M, Sakaki Y. 1997. A novel gene isolated from human placenta located in Down syndrome critical region on chromosome 21. *DNA Res.* 4:321-324.

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

Nimchinsky EA, et al. 1999. A neuronal morphologic type unique to humans and great apes. *Proc Natl Acad Sci U S A.* 96:5268-5273.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302:205-217.

Olson LE, et al. 2007. Trisomy for the Down syndrome 'critical region' is necessary but not sufficient for brain phenotypes of trisomic mice. *Hum Mol Genet.* 16:774-782.

Ostlund G, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38:D196-203.

Patterson F, Linden E. 1981. The education of Koko (Holt Rinehart & Winston)

Prior HM , Walter MA. 1996. SOX genes: architects of development. Mol Med. 2(4):405-12.

Pennacchio LA, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444:499-502.

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

Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676-679.

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

Prado-Martinez J, et al. 2013. Great ape genetic diversity and population history. *Nature* 499:471-475.

Prufer K, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527-531.

Prufer K, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43-49.

Pybus M, et al. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* 42:D903-909.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genet. 10:e1004525.

Sabeti PC, et al. 2006. Positive natural selection in the human lineage. *Science* 312:1614-1620.

Saitou N. 2014. *Introduction to evolutionary genomics* (Springer).

Savage-Rumbaugh S, Rumbaugh DM, McDonald K. 1985. Language learning in two species of apes. Neurosci Biobehav Rev. 9:653-65.

Scally A, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:169-175.

Schones DE, et al. 2008. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132:887-898.

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. 2014. TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* 42:D922-925.

Semendeferi K, Lu A, Schenker N, Damasio H. 2002. Humans and great apes share a large frontal cortex. *Nat Neurosci.* 5:272-276.

Siepel A, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-1050.

Sotomayor M, Weihofen WA, Gaudet R, Corey DP. 2012. Structure of a force-conveying cadherin bond essential for inner-ear mechanotransduction. *Nature* 492:128-132.

Strick PL, Dum RP, Fiez JA. 2009. Cerebellum and nonmotor function. *Annu Rev Neurosci.* 32:413-434.

Sumiyama K, Saitou N. 2011. Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. *Mol Biol Evol.* 28:3005-3007.

Symmons O, Spitz F. 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. *Philos Trans R Soc Lond B Biol Sci.* 368:20120358.

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

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol.* 30:2725-2729.

Tay SK, Blythe J, Lipovich L. 2009. Global discovery of primate-specific genes in the human genome. *Proc Natl Acad Sci U S A.* 106:12019-12024.

Team RC. 2015. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2013. *Document freely available on the internet at: http://www. r-project. org.*

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673-4680.

Uhlen M, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* 347:1260419.

Van Rossum G. 2007. Python Programming Language, *USENIX Annual Technical Conference* (41).

Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* 461:199-205.

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.

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* 41:D358-365.

Wheeler TJ, et al. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* 41:D70-82.

Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A.* 106:7273-7280.

Xie C, et al. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* 8:e1002942.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics.* 9:40.

Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol.* 8:e1000494.

Figure legends

FIG. 1 — Hominidae specific gene identification pipeline. Human was used as focal species and its genes were searched against the rest of Hominidae members' genome to identify Hominidae shared genes, indicated by group I. Using the same strategy, pairwise orthologous genes were identified between human and outgroup species, indicated by group II. Intersection

of Group I and Group II were omitted from Hominidae shared genes which gives rise to Hominidae specific genes based on CCDS and Rfam databases. Group III genes were searched in orthology prediction databases (Inparanoid, Treefam, OrthoDB, PhylomeDB) along with DNA and protein databases (Genebank, EMBL, DDBJ, PDB, RefSeq, NCBI and Uniprot KB) and any of the gene queries with significant homology (coverage >70%, Identity >50%) in non Hominidae members were discarded.

FiG. 2 — Evolutionary origin of DSCR4 gene. (a) Multiple sequence alignment of DSCR4 homologous sequences in Hominidae family members along with gibbon and rhesus macaque. Multiple sequence alignment for the sequences was undertaken using combination of Mishima, ClustalW and T-coffee. Identical and variant sites are defined based on Human genome reference sequence. (b) Schematic representation of the evolution of DSCR4/8 genes and their shared promoter. Green arrows represent functional protein coding exons, yellow arrows represent exons coding only UTR, brown rectangles represent exons' nonfunctional ancestral sequences and cross marks represent absence of homologous sequences for corresponding exon. (c) Evolution of genomic region located between KCNJ6 and KCNJ15 genes and contribution of transposable elements in formation of DSCR4/8 genes along with their shared promoter.

FiG. 3 —The polymorphism coverage and DAF analysis of HS HCNSs. (a) The average number of polymorphisms (SNP and INDEL) in 114 bp (average length of HS HCNSs) of HS HCNS along with HS HCNS flanking regions. Complete polymorphism data of 1000 genome project along with polymorphisms with frequency less than one from great apes genome project were used. Polymorphisms are significantly underrepresented in HS HCNSs compared to random sequences (t-test p value<10^{-16} for all members). (b) DAF distribution for Yoruba from Nigeria. Error bars were estimated using binominal distribution as $\sigma=\sqrt{(pq)/N}$, where p

represents the fraction of polymorphisms in a particular bin, q represents (1-p), and N represented the total number of polymorphisms.(c) Conservation levels of HS HCNSs' flanking regions. Point 0 is the average percent identity of 100 bp at the center of the HCNSs, whereas other points are the average of 50-bp windows moved at 20-bp steps starting from 30 pb inside the HCNSs. The standard error of the mean for each window is represented as error bars.

FiG. 4 — HS HCNS substitution rate across catarrhini phylogenetic tree. (a) catarrhini phylogenetic tree color-coded based on the substitution rate per million year in HS HCSNs orthologous sequences. Nucleotide substitution rates in rhesus macaque, gibbon and Hominoidea common ancestor in HS HCNS orthologous sequences are significantly higher than that of neutral evolutionary rate (represented as green in color key). Strongest accelerated mutation rate was observed in Hominidae common ancestor. (b) Comparison of genomic divergence in 32 % of HS HCNS's ancestral sequences in Hominidae common ancestor along with (c) 60% of HS HCNS's orthologous and ancestral sequences in Hominidae common ancestor and gibbon with that of random sequences under pure neutral evolution reveals signature of accelerated evolution in HS HCNS orthologous and ancestral sequences.

FiG. 5 —Nucleosome occupancy probability for HS HCNSs including flanking regions. Zeroth nucleotide position represents the center of HCNSs and also the center of the random samples. Blue and red graphs show nucleosome occupancy probabilities of the HS HCNSs and random samples respectively. (b) HS HCNS average nucleosome occupancy score derived from genome occupancy profiling generated by ENCODE/Stanford/BYU. HS HCNSs do have significantly lower nucleosome occupancy compared to flanking regions. (c) HS HCNS overlap with H3K9me histone mark compared to random sequences. H3K9 methylation is the mark of

heterochromatin. HS HCNSs are significantly underrepresented in H3K9me marked regions defined by ENCODE project.

Figure 1

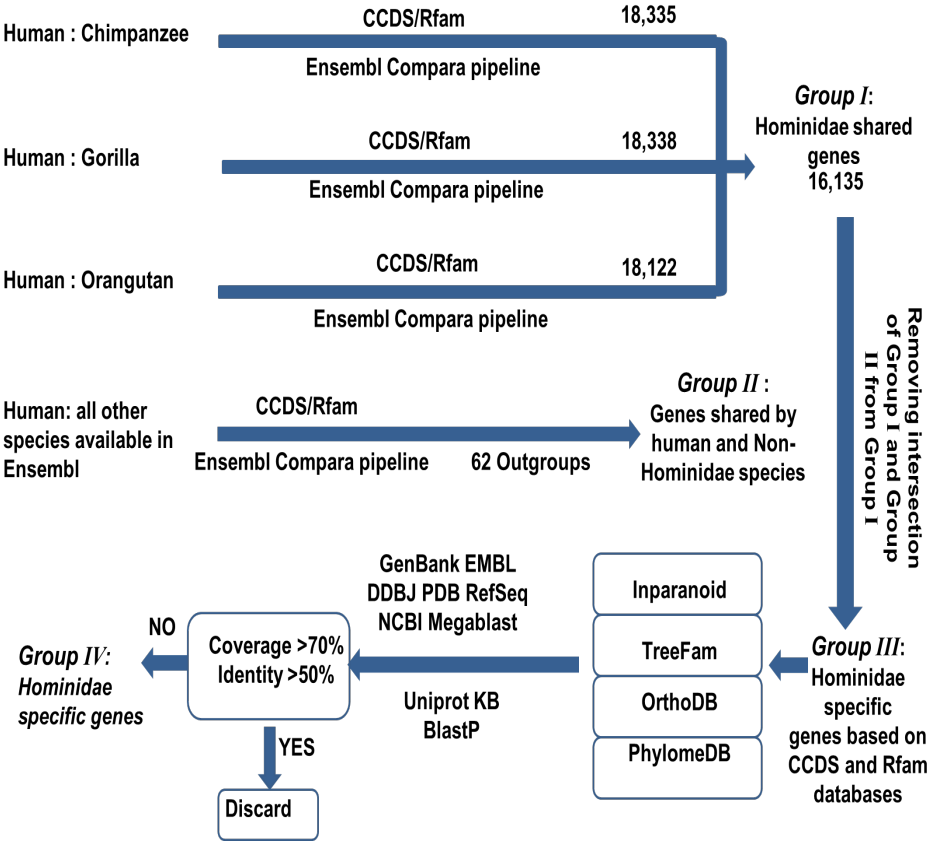


Figure 2

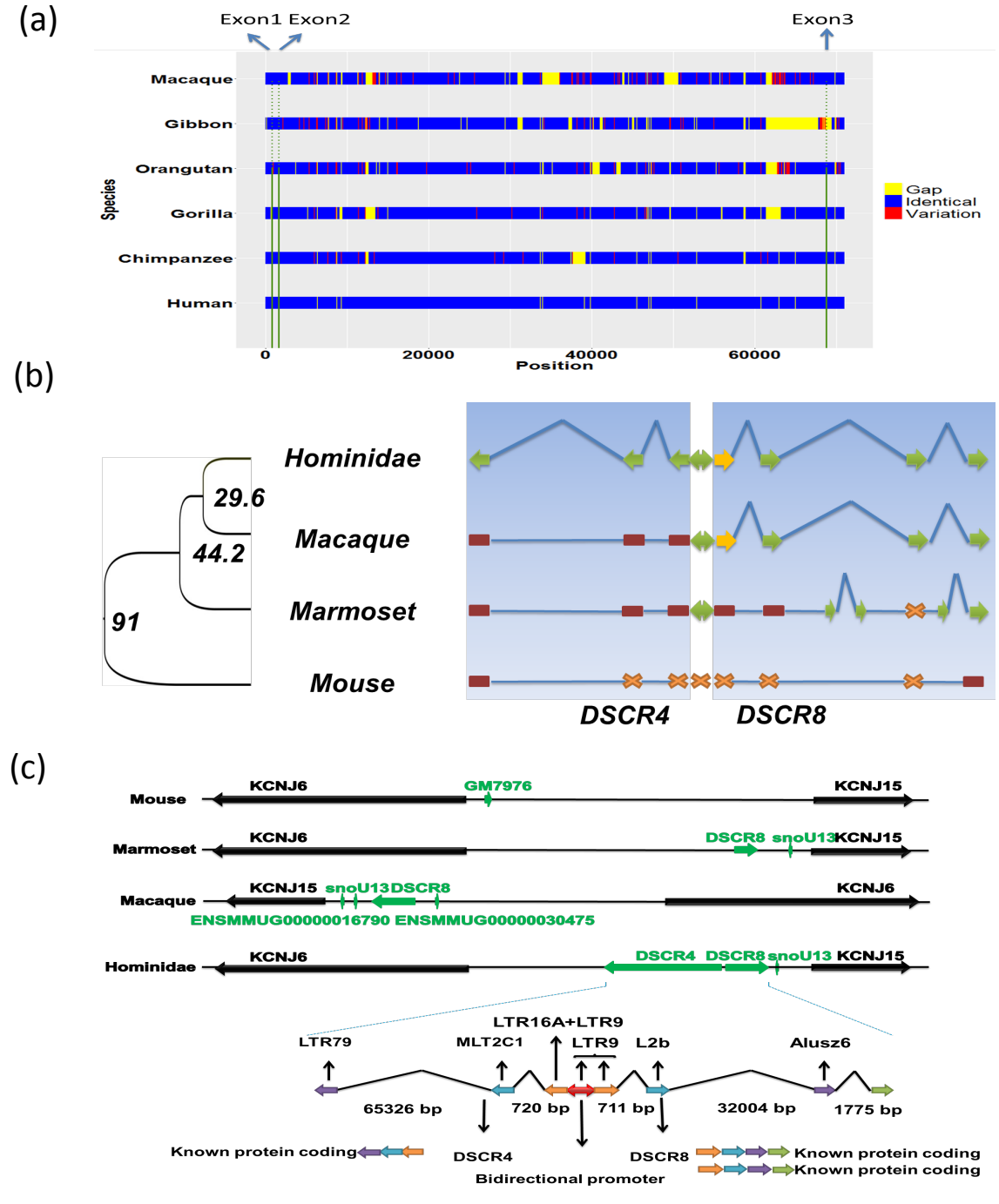
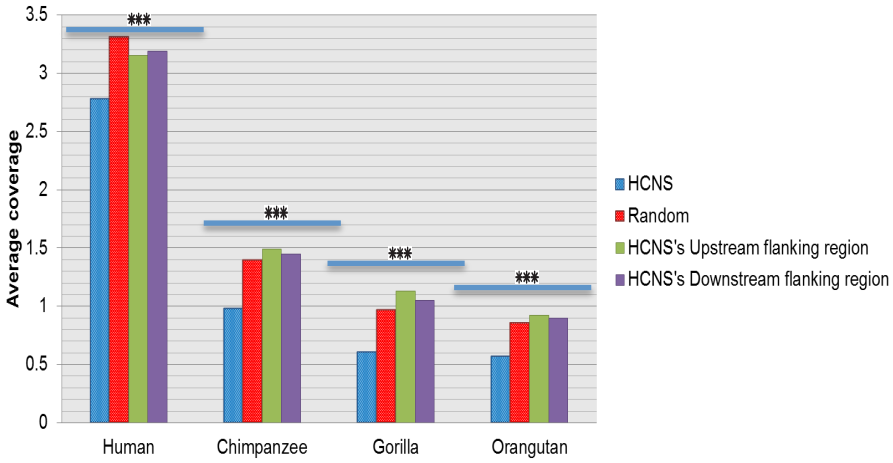
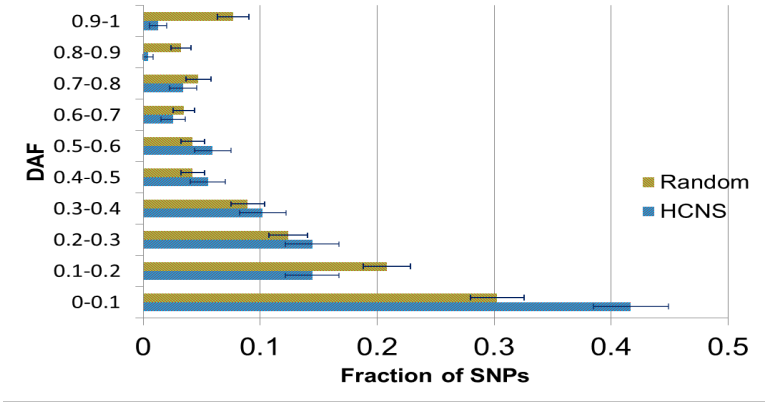


Figure 3

(a)



(b)



(c)

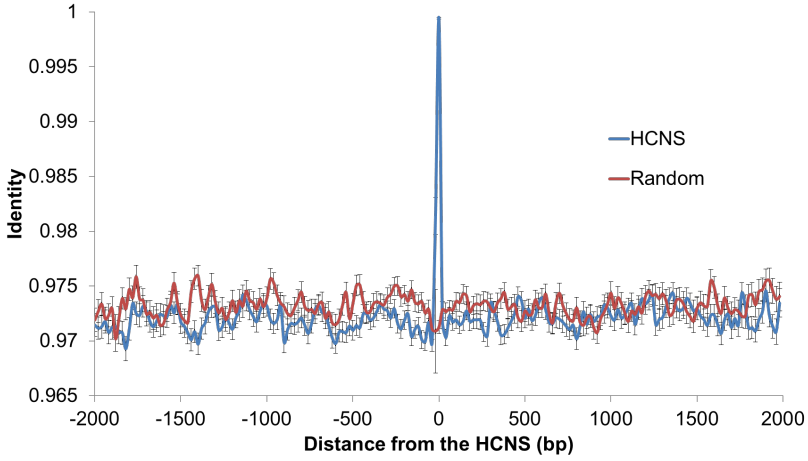


Figure 4

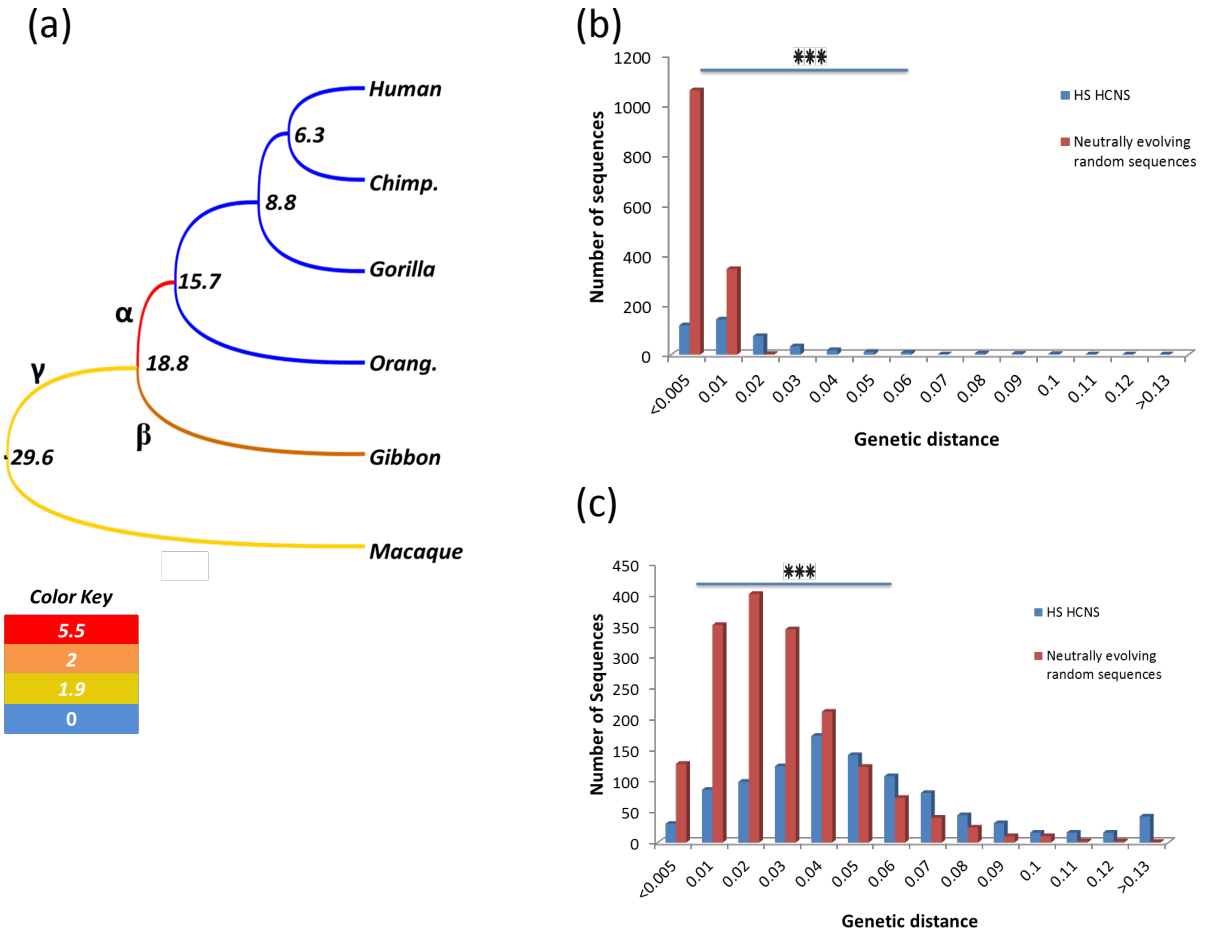


Figure 5

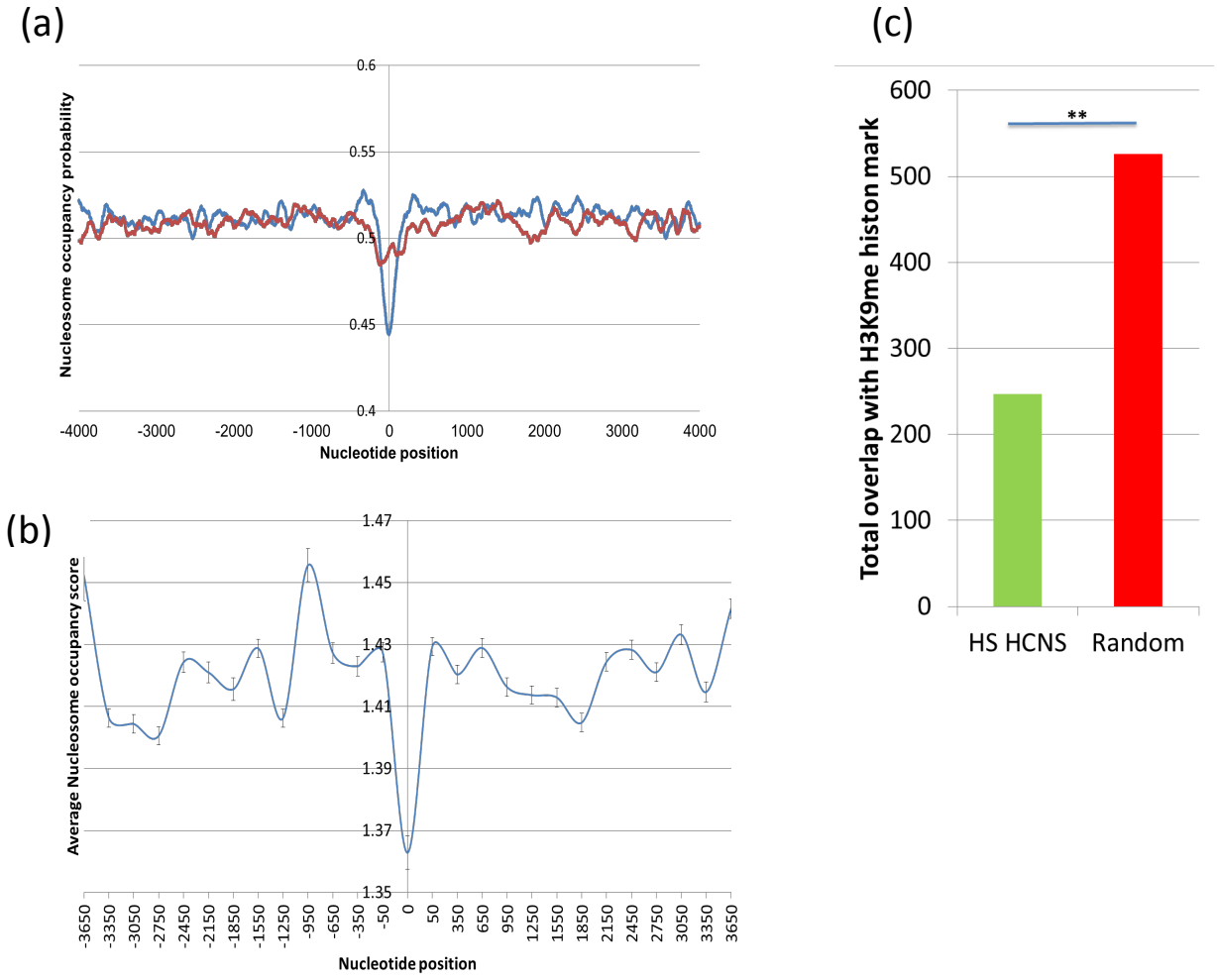


Table 1. Fractions (%) of genomic categories in HS HCNSs and the human genome

	HS HCNS*	Human genome
Intergenic	59.5 (1102)	74.4
Intronic	38.0 (703)	24.6
Promoter	1.1 (21)	0.6
UTR	1.4 (26)	0.4

* Absolute numbers are given in parentheses.

Table 2. Gene Ontology of HS HCNS-Associated Genes

Biological Process	Fold Enrichment
sensory perception of sound	3.40
cell-cell adhesion	1.76
mesoderm development	1.68
cell adhesion	1.68
biological adhesion	1.68
system process	1.58
neurological system process	1.57
system development	1.53
multicellular organismal process	1.48
single-multicellular organism process	1.48
developmental process	1.42
cell communication	1.28
cellular process	1.25