学位論文

# Computational and experimental analysis of evolutionary changes at Hominidae and Hominoidea specific coding and conserved noncoding genomic elements

（ヒト科とヒト上科特有のコード・非コードゲノム要素の進化的変化
に関するコンピュータ解析および実験的解析）

平成28年12月博士（理学）申請

東京大学大学院理学系研究科

生物科学専攻
マフモウディサベール　モルテザ

# Abstract

Superfamily Hominoidea, includes humans and apes, is one of the two living superfamilies of parvorder Catarrhini. Taxonomically this superfamily belongs to order primates and is comprised of two families; Hominidae (humans and great apes) and Hylobatidae (lesser apes). All members of this superfamily have large brains, well known for their complex social behavior and intellectual abilities. Despite the increasing genome data, the genetic factors underlying the phenotypic uniqueness of Hominoidea (apes) and Hominidae (great apes) have 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 which are comprehensively investigated in the superfamily Hominoidea and family Hominidae in this study. For identification of HCNSs, using the neutral evolution thresholds, 1,658 and 679 HCNSs were identified with 100 percent sequence similarity in Hominidae and Hominoidea, respectively.

HCNSs do have significantly lower overlap with genomic polymorphisms compared to random coordinates and their flanking regions indicating the functionality of these conserved elements. Derived allele frequency (DAF) analysis of HCNSs, further confirmed that the lower evolutionary rate within these elements is not due to them being located on mutation cold spots, but rather due to the evolutionary constraint that prevents spread of mutations occurring within these elements in human populations. To figure out the evolutionary forces leading to the formation of HCNSs, the properties of orthologs of HCNSs in outgroup species were investigated. With assumption of molecular clock, it has been shown that substitution rates in Hominidae and Hominoidea ancestral branches for HCNSs are respectively 5 and 2.3 times higher than that under neutral evolution, suggesting that these elements may have emerged through some kind of positive selection, and then purifying selection started to operate to keep the functions served by these elements.

HCNSs tend to cluster around genes involved in nervous system, transcription regulation and development. HCNS target genes were also shown to have significantly higher non-coding portion compared to the total human protein coding genes suggesting the action of evolutionary forces to prevent loss of conserved noncoding regions in these genes. HCNS target genes were also shown to be located in isolation with higher absolute distance to the next protein coding

genes compared to average human genome protein coding genes. Analysis of the production of enhancer RNA (eRNA) and the overlap with H3K4me1 and H3K27ac epigenomic marker revealed depletion of enhancer markers in young Hominoidea- and Hominidae-restricted HCNSs that have evolved less than 30 million years ago. Expression analysis of HCNS target genes within human tissues also suggest inhibitory effect of HCNSs on their target genes, especially in fetal brain tissue, the tissue in which HCNSs are supposed to be in their most active form considering Gene Ontology data.

For identification of coding genes restricted to superfamily Hominoidea and family Hominidae, major DNA, protein and gene orthology databases were analyzed. Setting strict thresholds in order to avoid false positives, only one protein coding gene named Down Syndrome Critical Region 4 (DSCR4) showed strong evidence of being bona fide Hominidae specific gene. Evolution of DSCR4 can be mainly classified in three evolutionary periods. Period (1): LTR79 retrotransposition, that took place in the common ancestor of mammals >100 Mya. This transposition formed DSCR4's exon 3 ancestral sequences. Period (2): during the evolution of common ancestor of primates at 29– 45 Mya, 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. Period (3): the final required mutation was a GC transversion that formed the stop codon and completed the formation of DSCR4 ORF.

There are numerous computational evidences at RNA and protein levels indicating the functionality of DSCR4 gene, however, due to the absence of known protein domains within the DSCR4-coded protein, the function of DSCR4 within the cell could not be determined without experimental investigations. For experimental verification of the functionality of DSCR4, I conducted DSCR4 gene perturbation analysis followed by transcriptome profiling in non-canceric human bone marrow cells which provided strong evidence for the involvement of this gene in regulation of cell migration, motility and locomotion. This hypothesis, is further supported by tissue-specific expression of DSCR4 in human cells in which migration is important for proper functioning.

This study provides candidates of gene and regulatory elements which are expected to hold the key to the understanding of the phenotypic uniqueness shared by Hominoidea and Hominidae, 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 apes

# Contents

# List of figures

# List of tables

# List of abbreviations

CNS: Conserved Noncoding Sequence
HCNS: Highly Conserved Noncoding Sequence
HS HCNS: Hominidae-specific Highly Conserve Noncoding Sequence
GO: Gene Ontology
GFP: Green Fluorescence Protein
DSCR: Down syndrome Critical Region

# Chapter 1. Introduction

## 1.1 A general introduction on superfamily Hominoidea and its unique phenotypes

The superfamily Hominoidea which include humans, chimpanzees, bonobos, gorillas and orangutans and gibbons are a group of Old World tailless primates, native to Southeast Asia and Africa. The sister superfamily of Hominoidea are Old World monkeys which include several species of monkeys such as rhesus macaque, proboscis monkeys and baboons. These two superfamilies, which have diverged at around 30 million years ago (Mya), form the Catarrhini parvorder (Figure 1-1 ).

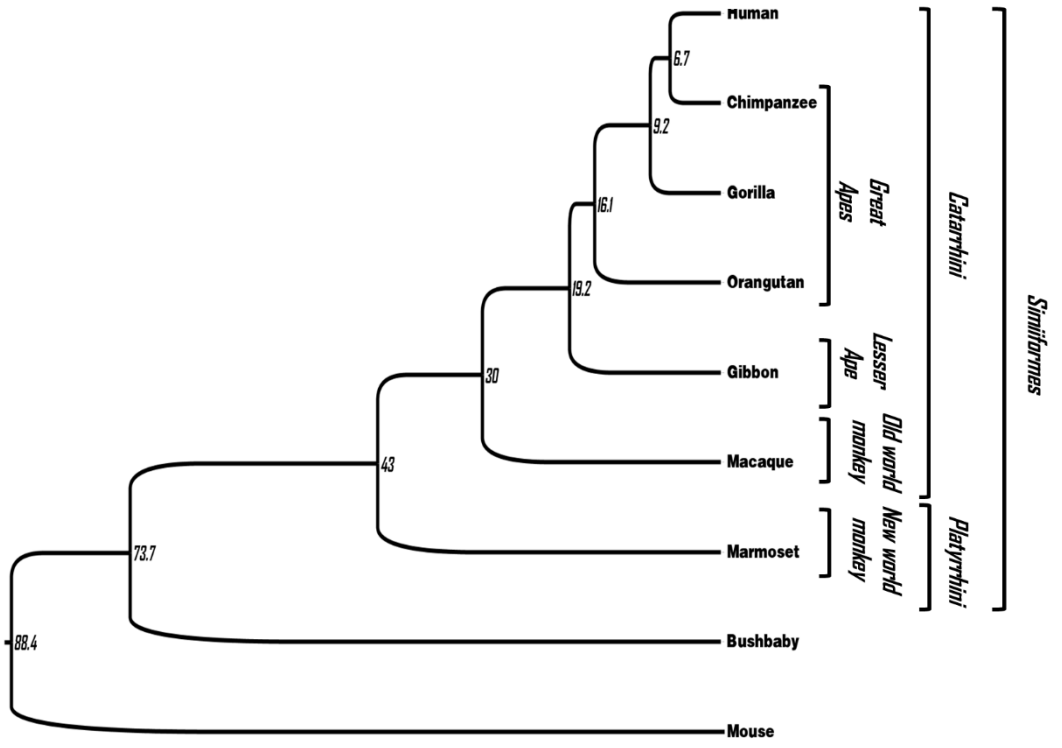



Figure 1-1. Phylogenetic tree and divergence times of lineages in order primate.

The phylogenetic tree is constructed based on data provided by Hedges et al. (2015). Divergence times are represented in the units of 'million years ago' in the tree.

The Hominoidea are the only tailless primates, the characteristic that phenotypically separates them from the closely related Old World monkeys and New World monkeys. This superfamily is composed of two living families: Hominidae including humans and Great Apes and Hylobatidae including lesser apes.

The family Hominidae includes chimpanzees, bonobos, gorillas, orangutans and humans. This family is also known as hominids. Currently, there are seven living species of Hominidae including two chimpanzees (*Pan troglodytes* or common chimpanzees and *Pan paniscus* or *bonobos*), two gorillas (*Gorilla gorilla* or *Western gorilla* and *Gorilla beringei* or *Eastern gorilla*), two orangutans (*Pongo pygmaeus* or *Bornean orangutan* and *Pongo abelii* or *Sumatran orangutan*) and a single extant species of humans (*Homo sapiens* or modern humans).

Members of superfamily Hominoidea are generally called as “apes” and this term has been broadly used in different senses within scientific settings. Traditionally, this term was used to name all members of superfamily Hominoidea except for human beings (Dixson 1981) but in other resources, it is used to name all members of Hominoidea including humans; for example in the book ‘The descent of man’ written by the pioneering evolutionary biologist, Charles Darwin, the humans were described as “big-brained apes” (Darwin 1872). So ‘ape’ is synonymous to ‘hominoid’ and includes all members of Hominoidea including humans (Benton 2009).

The superfamily Hominoidea, the name of which has originated from the Latin word meaning ‘Man-like’, are well-known for their complex social behaviors and intellectual abilities and as their name indicates, they share physiological, anatomical and behavioral similarities to that of human. To date, there have been numerous studies reporting the unique cognitive

abilities and anatomical characteristics of Hominoidea members. Some of these studies will be briefly introduced in here.

Understanding the relation between causal factors and their effect is an essential feature of advanced social and physical cognition. In a study conducted on orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes), bonobos (Pan paniscus) and gorillas (Gorilla gorilla) by Mulcahy and Call (2006) it was shown that some great apes do possess casual knowledge about the trap- tube task. In another study conducted by the same group, it was shown that gorillas (Gorilla gorilla) and orangutans (Pongo pygmaeus) possess the important elements of problem solving which are abilities to encode relevant task features and to combine multiple actions to achieve the goals. They also showed the ape subjects are proficient at using tools to achieve the reward (Mulcahy et al. 2005). Great apes are also shown to be able to learn how to discriminate objects based on achromatic color and weight (Schrauf and Call 2009).

Apes were shown to possess some problem solving skills(Martin-Ordas and Call 2009). In a study performed by Call (2007), apes were also demonstrated to know that hidden objects can affect the orientation of other objects. These findings showed that the apes could make some inferences about the reasons for inclined orientation of the boards they confronted, not simply associated with the presence of the award.

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-Hominoidea 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 apes family members and a distinctive feature compared to the rest of the species.

Phenotype is the result of a collective network of genes along with other regulatory elements. The unique cognitive and anatomical phenotypes observed in superfamily Hominoidea and the family Hominidae indicate the existence of unique shared coding and non-coding functional regulatory genomic elements which underlies such characteristics. There are two categories of high-potential evolutionary candidates underlying clade-specific phenotypes including clade-specific genes and clade-specific highly conserved noncoding sequences (HCNSs). These categories of genomic elements will be introduced within the next sections.

## 1.2 Origin of novel phenotypes

### 1.2.1 Novel highly conserved noncoding sequences (HCNSs)

Conserved sequences in noncoding sections of genomes have been examined for nearly two decades for their functional importance (Babarinde and Saitou 2016; Bejerano et al. 2004; de la Calle-Mustienes et al. 2005; Jareborg et al. 1999; Lindblad-Toh et al. 2005; Saber et al. 2016; Siepel et al. 2005). Stretches of highly conserved noncoding elements in non-repetitive untranslated regions of protein coding genes, intergenic and intronic regions have captured the attention of researchers to their enigmatic properties. In some cases, the levels of homology present in these conserved elements were even higher than protein coding genes (Babarinde and Saitou 2013; Saber, et al. 2016) so I would like to refer to such elements as 'highly conserved noncoding sequences' or HCNSs in this dissertation. These highly conserved elements have been identified spanning across various evolutionary periods. For instance, Babarinde and Saitou (2016) documented that the conserved elements they found, shared by primates, rodents, carnivores and cetartiodactyls, have emerged more than 300 million years ago while in another study by the author (Saber, et al. 2016), highly conserved elements restricted to humans and great apes which have emerged less than 20 million years ago were identified.

The conserved noncoding elements have called by different terminologies such as UCEs -ultraconserved elements (Bejerano, et al. 2004), CNEs - conserved noncoding elements, HCNEs highly conserved noncoding element (Lindblad-Toh, et al. 2005), HCNRs - highly conserved noncoding regions (de la Calle-Mustienes, et al. 2005), however, the functional importance of these conserved element are consistently shown by nearly all of the studies. In various studies of HCNSs in vertebrates (Babarinde and Saitou 2013; Babarinde and Saitou 2016; Bejerano, et al. 2004; Saber, et al. 2016; Takahashi and Saitou 2012), Invertebrates (Siepel, et al. 2005) and plants (Hettiarachchi et al. 2014; Kaplinsky et al. 2002; Kritsas et al. 2012), HCNSs were reported

to have significant potential to have regulatory functions related to lineage- and clade-specific characteristics. HCNSs have also been shown to be under strong purifying selection (Babarinde and Saitou 2016; Drake et al. 2006; Saber, et al. 2016; Takahashi and Saitou 2012). More recently, the functional importance of some of HCNSs have been functionally verified and confirmed. For instance, in an investigation by Lee et al. (2011), a HCNS have been shown to be able to drive expression of a reporter gene in mouse and zebra fish embryos. HCNSs were also shown to have unique enrichment pattern regarding epigneomic markers. Babarinde and Satiou (2016) reported that ancestral HCNSs shared by amniotes are associated with enriched H3K4me1 histone modification signal, especially in the tissue of fetal brain. This histone modification is known to be strongly related with active enhancer elements.

There are dozens of examples for phenotypic changes generated by mutations in cis-regulatory elements such as breadth and length of Darwin's finches' beak (Abzhanov et al. 2004) and stickleback fish's pelvic fins (Wray 2007). Therefore, cis-regulatory elements seem to have played crucial roles in the evolution of phenotypic characters. This hypothesis is further supported considering the small degree of amino acid differences between closely related species with dramatic phenotypic differences (King and Wilson 1975). Although there is still considerable uncertainty regarding the function of majority of HCNSs, 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 these elements to be under purifying selection probably due to their functional importance as regulatory elements (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 embryonic stage (Benko et al. 2009; Kritsas, et al. 2012). Such shared properties suggests that HCNSs are potent candidates to be involved in emergence of order-specific phenotypes.

### 1.2.2 Novel protein coding genes

Various studies investigating the nature of mutations underlying adaptive evolutionary innovations have revealed that along with subtle genomic alterations of existing protein coding genes which lead to differences in encoded RNA or protein activities, emergence of novel genes with novel functions might have significantly contributed to the evolution of clade- or lineage-specific phenotypic characteristics (Kaessmann 2010). As a result, the phenomenon of the birth and evolution of novel genes has received much attention from evolutionary biologists in the past several decades. In fact, investigation of the origin and fate of novel genes dates back to the time even before identification of the structure of DNA double helix, where refashioned copies of old genes were introduced as a method for emergence of new genes (Haldane 1933) .

The recent detailed analysis of genomes by comparative genomics after completion of whole genome sequencing projects for humans and dozens of other species, have accelerated the discovery of mechanisms underlying the emergence of novel genes. One of the mechanisms of evolution of novel genes which has drawn lots of attention is origination of novel genes “from scratch” that is from formerly non-functional genomic sequences (Wu and Zhang 2013). The de novo origination of novel protein-coding genes from scratch was considered as extremely unlikely for a long time. For example, it was once stated by Francxois Jacob in an essay that the ‘‘probability that a functional protein would appear de novo by random association of amino acids is practically zero’’ and thus the ‘‘creation of entirely new nucleotide sequence could not be of any importance in the production of new information’’ (Jacob 1977).

In spite of this long-held belief, recent studies have identified several novel protein-coding genes arose from formerly noncoding genomic elements. One of the cases of this phenomenon, reported in literature, is the *Morpheus* gene family which emerged in Old World primate ancestor (Johnson et al. 2001). Although the exact evolutionary mechanism underlying the origination of coding sequence of this gene family has remained unclear, the absence of any orthologous gene in species other than Old World primates indicates the de novo origination of this gene family. It was also shown that ancestral protein coding gene of this gene family have enormously expanded by segmental duplication in Old World primates, and the multiple copies of Morpheus gene family revealed to have experienced strong positive selection within the coding sequence, suggesting accelerated function of adaptive evolution on this gene family. Although the functional role of Morpheus genes in Old World primates is yet unclear, the strong purifying selection functioning on this gene family indicate the critical and rapidly evolving role of the encoded protein within Old World primates.

In yet other studies, fourteen de novo originated genes have been identified in Drosophila genome which are mainly expressed only in testes (Levine et al. 2006; Zhou et al. 2008). These findings suggest that a large proportion of novel genes in this genus may have emerged through de novo gene formation. De novo originated genes have also been identified in primates such as human and single cell eukaryotic cells such as yeast (Cai et al. 2008; Toll-Riera et al. 2009). In a study of human de novo originated protein coding genes, three genes were identified to have arisen from scratch in human genome, the functionality of which are supported by evidence of translation of their coding sequences (Knowles and McLysaght 2009). They further proceeded to estimate that 0.075% of human genes may have originated de novo from scratch leading to a total expectation of 18 such cases in human genome containing 24,000 protein-coding genes.

Lineage-specific genes, which are often called orphan genes, have been described in a multiple organisms, including primates, rodents, and plants. These studies have revealed that orphan genes tend to have a simple structure, a short protein size, and are preferentially expressed in tissue-specific manner. As orphans lack homologues in other species, many of these genes are likely to have arisen de novo while the rest are the result of gene duplication, frame-shift fixation, creation of overlapping genes, horizontal gene transfer, and exaptation of transposable elements.

In conclusion, these studies indicates that de novo emergence of novel genes is more frequent than previously expected, although the functional relevance and phenotypic implications of the majority of de novo originated genes are yet unclear. These studies also indicate that two key steps must precede the origination and fixation of novel protein coding genes from an ancestrally noncoding DNA sequence, first, the DNA must turn into transcriptionally active elements and second, it must evolve a complete and coherent open reading frame with potential to encode a beneficial protein (Kaessmann 2010). In fact, there are several mechanisms through which both of these steps could occur within the genome that increase the likelihood of emergence of novel de novo originated genes. In summary, based on recent findings, it is well-confirmed that emergence of novel genes have contributed to the functional genomic evolution and emergence of novel phenotypes which in turn, demonstrates the critical importance of birth and function of novel genes in the evolution of organisms.

## 1.3 Scope and objective of this study

Studies of lineage-specific characteristics and phenotypes, have so far emphasized on the importance of emergence of novel regulatory elements and novel protein coding genes that are unique to particular taxa or clade. HCNSs, in multiple studies, have been suggested to be potential novel regulatory elements contributing to the emergence of clade- and lineage specific phenotypes. However, previous studies of HCNSs in primates do not have the proper resolution to identify ape-specific HCNSs (Babarinde and Saitou 2013; Keightley et al. 2005; Takahashi and Saitou 2012) and focused mainly on potential regulatory functions of HCNSs that must have driven lineage specific characteristics unique to primates.

There have also been multiple findings regarding the functional importance of novel de novo originated protein coding genes in several kingdoms of species. In a study of Drosophila, it was shown that a novel gene *Xcbp1,* has gained expression in specific brain structure through adaptive evolution which in turn has led to the modifications in neural circuits, contributing to the evolution of foraging behavior in this species (Chen et al. 2012). A novel human-specific protein coding gene, named FLJ33706, was also shown to be involved in brain functions and pathogenesis of Alzheimer disease (Li et al. 2010).

These evidences indicate the clade- and species- specific regulatory and coding architecture which govern lineage specific characteristics. Comparative genomic analysis is the primary approach for identification of such functional group-specific coding and regulatory genomic elements. The idea of identification of Hominoidea- or Hominidae-specific coding and conserved regulatory elements underlying unique ape-restricted phenotypes has not comprehensively tested so far. This idea formed the foundation of this study to uncover lineage

specific functional genomic sequences in family Hominidae and superfamily Hominoidea and to characterize their genomic properties.

In chapters 2 and 4, I analyzed whole genome sequences along with gene orthology data retrieved from major DNA databases to find Hominidae- and Hominoidea-specific genes and HCNSs, respectively. The initial objectives for these parts of the study were,

- Identification of HCNSs originated during the evolution of common ancestor of Hominidae and Hominoidea
- Determination of the mechanisms of emergence and evolution HCNSs and novel genes
- Functional analysis of Hominoidea- and Hominidae-restricted HCNSs and genes
- Characterization of HCNS associated genes, their distribution and characteristics
- Identification of unique epigenomic enrichment pattern of HCNSs across human tissues
- Determining the role of HCNSs within the genome.

As identified in Chapter 2 (Already reported as Saber et al. 2016), *Down syndrome critical region 4 (DSCR4),* is a functionally unknown authentic Hominidae-specific protein coding gene that is located on medically important region name Down Syndrome critical region of chromosome 21. To elucidate the functional role of DSCR4, I conducted overexpression experiments using non-canceric human bone marrow cells and deduced the regulatory networks in which this gene is involved using differential gene expression analysis. The experimental analysis is the subject of chapter 3 of this thesis.

All the analyses were conducted via custom made python scripts which were coded by myself and genome and proteome data were retrieved from public genome databases.

## 1.4 Thesis structure

This dissertation entails five chapters with two chapters that address multiple aspects of HCNSs in family Hominidae and superfamily Hominoidea and a chapter that discuss the likely functional roles of DSCR4 gene in human cells.

Chapter 1 provides a general overview over highly conserved noncoding sequences (HCNSs) and de novo originated protein coding genes and their importance in emergence of novel clade-specific characteristics and phenotypes.

Chapter 2 mainly deals with analysis of Hominidae-specific HCNSs and de novo originated protein coding genes. This chapter is further subdivided into multiple sections to handle methodology, results and discussion on the importance and functionality of HCNSs and de novo originated protein-coding genes restricted to Hominidae. Chapter 2 has been published as Saber et al. (2016) in the journal of Genome Biology and Evolution in 2016 June (doi:10.1093/gbe/evw132).

Chapter 3 deals with the experimental functional analysis of DSCR4 gene. It discusses the experimental methodology, results and discussion regarding the likely biological pathways affected by DSCR4 gene perturbation.

Chapter 4 covers the analysis of unique characteristics of HNCSs restricted to superfamily Hominoidea. This chapter tries to elaborate the functional role of HCNSs within genome and compares their characteristics with those of ancestral CNSs.

Chapter 5 was written to summarize the results on chapter 2, 3 and 4 and provides description on future directions to further corroborate and extends the results generated in this study.

# Chapter 2. Hominidae-specific coding and highly conserved noncoding genomic sequences

## 2.1 Introduction

Family Hominidae which includes humans and Great apes, is one of the two living families of ape superfamily Hominoidea (See Appendix A1 for phylogenetic representation). Taxonomically, this family belongs to the order Primates. All members of this family have large brains, well-known for their unique complex intellectual abilities. Apart from humans, other species of this family have also shown signs of problem solving (Volter and Call 2012), the phenotype which have not been observed in other closely related monkeys.

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 family Hominidae. 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.

Here I explored the unique genomic elements underlying phenotypes restricted to the family Hominidae by identifying Hominidae-specific de novo originated genes and Hominidae-specific (HS) highly conserved noncoding sequences. I analyzed whole genome sequences along with gene expression and orthology data retrieved from multiple databases to identify Hominidae specific genes. I also analyzed Hominidae 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, I used stringent thresholds for identifying HS novel genes and HCNSs to minimize type I error. I found that there is a low proportion of HS protein-coding gene to HS highly conserved potential regulatory HCNSs, suggesting a likely stronger contribution of regulatory elements than novel genes in defining Hominidae-clade specific phenotypes.

## 2.2 Materials and methods

### 2.2.1 Retrieving genome sequences 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 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.

### 2.2.2 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 Appendix A2-a). 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 family Hominidae.

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 Appendix A3 for list of outgroup species and the build of their genomes). 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 2-1.

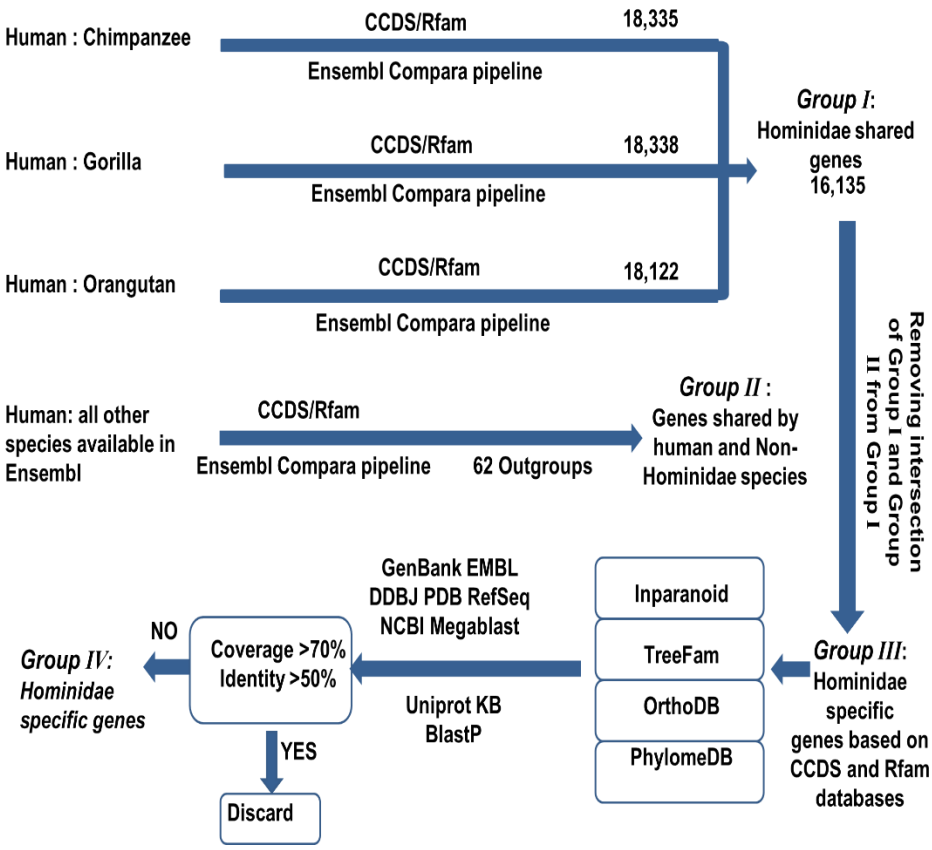



Figure 2-1. Hominidae specific gene identification pipeline.

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.

### 2.2.3 Identifying the evolutionary origins of Hominidae-specific genomic elements

To identify evolutionary processes leading to the emergence of Hominidae-specific 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.

### 2.2.4 Analysis of selection

Codon-wise and nucleotide-wise analysis of selection using the method described by Haygood et al. (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 Hominidae-specific 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 Hominidae-specific protein-coding gene.

### 2.2.5 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 Appendix A1), 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. To minimize the probability of false positives due to short divergence time, I set the threshold as 100% identity in conservation and 100bp in length for the identification of sequences under purifying selection.

For accurately determining the threshold for neutral evolution, I compared protein coding sequences' synonymous site variation rate with that of noncoding genomic divergence rate between species; I 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, I constructed the distribution plot of sequence divergence values. Mode of the plot was considered as the threshold of neutrally evolving sequences within noncoding sequences.

### 2.2.6 Homology search for identification of highly conserved noncoding sequences

After masking coding sequences in each genome, I searched for sequences that are under selective constraint in the Hominidae family. To this end, I 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 $3 \times 10^{9}$. 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 Hominidae-specific 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. Hominidae-specific HCNSs with conserved orthologous in interspersed repeats were also discarded.

### 2.2.7 Evolutionary origin of Hominidae-specific 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).

### 2.2.8 Single nucleotide polymorphism and derived allele frequency analyses

I 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 I retrieved and combined Single Nucleotide Polymorphism (SNP) and insertion/deletion variation data and since the variations were mapped to human genome, I filtered out all variations with allele frequency of 1.0. For each of the three Great Apes, I generated random sequences with the same number and size as Hominidae-specific HCNSs in each species and investigated the coverage of variation in HCNSs and random sequences. For derived allele frequency analysis, I retrieved SNP frequency data of the Yoruba population of Nigeria, from the International HapMap project. The ancestral alleles of SNPs overlapping the Hominidae-specific HCNSs or random sequences were determined using pyliftover and chimpanzee sequence.

I also extracted Hominidae-specific HCNSs along with 2000bp upstream and downstream flanking sequences and aligned the sequences using ClustalW. For each alignment, I made sliding windows of 50bp and step size of 20bp starting from 30bp inside the CNSs and

calculated the percent identity in each window. Afterwards, I calculated the average of the percent identity for each window.

### 2.2.9 Nucleosome occupancy probability analysis

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 Hominidae-specific 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, I 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 Hominidae-specific 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, I retrieved H3K9me mapping

data from ENCODE project and analyzed Hominidae-specific HCNS overlap with H3K9me histone mark compared to random sequences.

### 2.2.10 Genome distribution and Gene ontology analysis

I 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. I then found HCNSs that are located on UTR, promoter, intronic and intergenic regions. I 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.

I retrieved the coordinates of protein-coding genes from Gencode project. For Hominidae-specific HCNSs, I 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. I checked the functional analysis of Hominidae-specific HCNSs using Panther 9.0 (Mi et al. 2010). P-value corrected for multiple testing using Bonferroni correction was calculated.

### 2.2.11 Tissue specificity of Hominidae-specific HCNSs

To investigate whether Hominidae-specific HCNSs do have unique properties in tissue-specific manner, I 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 Hominidae-specific 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/).

## 2.3 Results

### 2.3.1 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. (1997) via EST mapping, was found to be the only annotated Hominidae-specific 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 Appendix A4a); but it is 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 Appendix A4a);. 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 structure with >90% sequence identity to DSCR4 in protein 3D databases. However, the secondary structure of DSCR4 based on Chou and Fasman algorithm (1974) suggests the existence of potential α-Helices and β-Sheets (See Appendix A5a). 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 Appendix A5b).

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 my shift to macaque genome as template for evolutionary analysis of the origin of DSCR4 gene (Figure 2-2a).

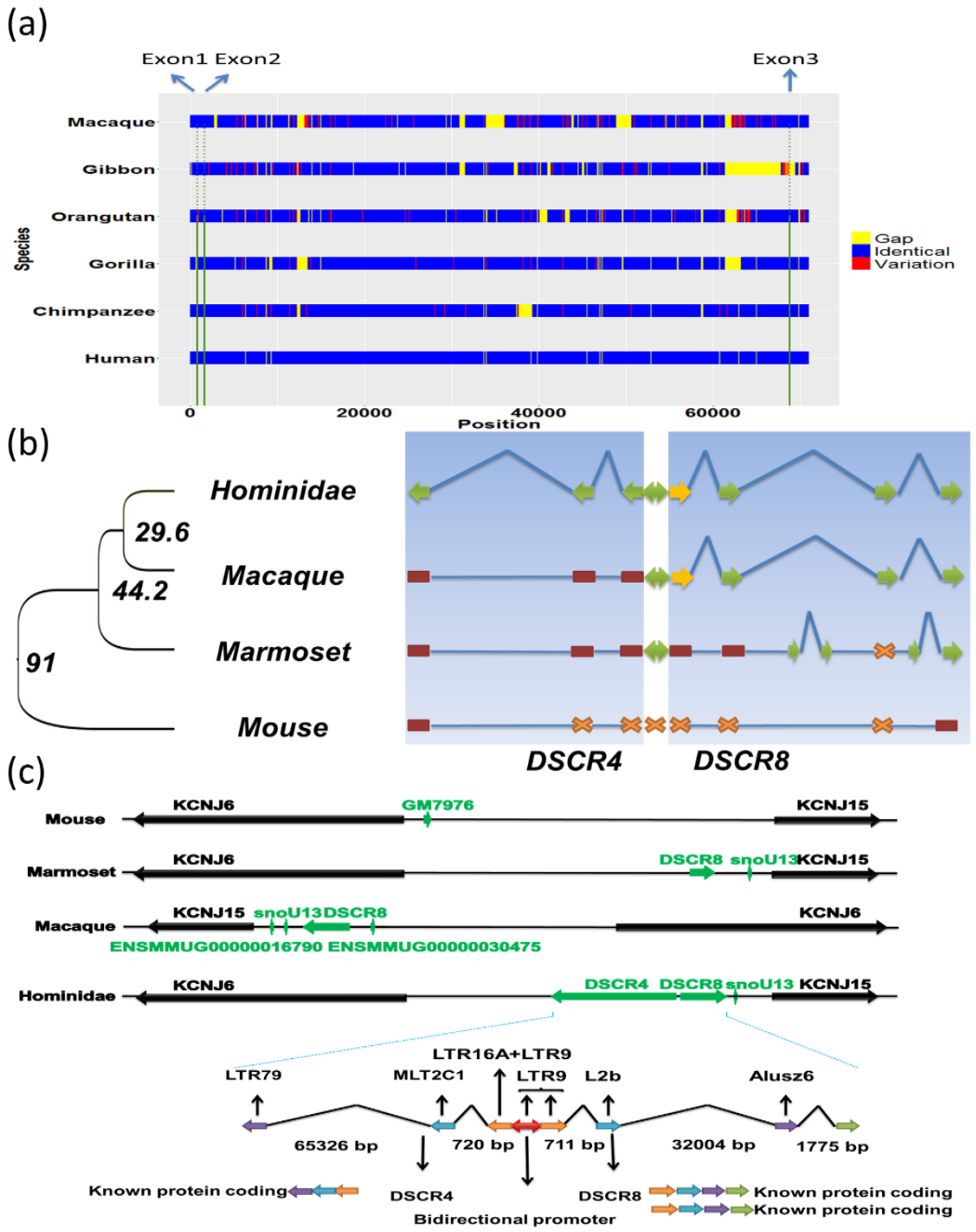



Figure 2-2. Evolutionary origin of DSCR4 gene.

(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. Divergence times represented in phylogenetic tree are in units of 'million years ago' (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.

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 family Hominidae, DSCR8 is present in Hominidae, Old World and New World monkeys. Multiple sequence alignment of DNA sequences corresponding to DSCR4/DSCR8 gene and their shared promoter in family Hominidae 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 *Appendix A7*b). 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 2-2c).

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 *Appendix A6*a). 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 *Appendix A8*). Period 3) The final ORF-enabling mutation was a GC transversion at DSCR4 exon 3 that formed the stop codon TGA (Appendix

A4). This transversion, which took place in the common ancestor of Hominidae 15-19 million years ago, completed the formation of the DSCR4 gene.

### 2.3.2 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 did not 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 *Appendix A9*a). Although analysis of the nonsynonymous and synonymous substitution rates indicate the action of purifying selection on DSCR4 in human and chimpanzees, the dn/ds ratios do not show selection constraint on this gene in gorilla or orangutans (See *Appendix A9*b). These results are consistent with previous findings stating that young genes are subject to weaker purifying selection (Cai and Petrov 2010).

### 2.3.3 Highly conserved noncoding sequences

Gorilla diverged from the common ancestor of *Homo* and *Pan* Genera 9.2 mya and Homo and Pan diverged about 6.7 mya. The common ancestor of Hominidae diverged from Hylobatidae 19.2 mya and 3.1 million years later at 16.1 mya, orangutan, the most distant member of Hominidae family emerged (Fig. 1-1). Such short divergence times within family members and between family Hominidae and phylogenetically close species have made discerning Hominidae-specific functional noncoding sequences under purifying selection from neutrally evolving sequences a challenging objective.

The crucial parameter for identifying linage-specific HCNSs in closely related species is the nucleotide identity thresholds 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), I considered the mode of the noncoding sequence alignment plot as the neutral evolution threshold. To verify the authenticity of this threshold, I also analyzed the neutral substitution rate in protein coding sequences. I constructed the synonymous substitution rate plot between human and three closest outgroup species, namely, gibbon, rhesus macaque and marmoset. 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 *Appendix A10*a-b) which suggests that the determined thresholds for neutrally evolving sequences are accurate. I filtered out all Hominidae-specific HCNS with orthologous sequences in any outgroups with divergence levels lower than neutral evolution threshold. Using this strategy I identified 1,658 Hominidae-specific HCNSs (The sequences and alignments of the discovered HS HCNSs are available online at: http://gbe.oxfordjournals.org/content/early/2016/06/10/gbe.evw132/suppl/DC1). Length distributions of Hominidae specific HCNSs are shown in *Appendix A11*. Probability analysis using whole genome blast hits frequency data (See *Appendix A12*) showed that the frequency of sequences meeting all these conditions by chance is $3.88 \times 10^{-8}$. Since the total number of

pairwise blast hits in each of reference genomes pairs are much less than 3.88 x $10^8$, it is extremely unlikely for Hominidae-specific HCNSs to be only cases of the outliers of neutral evolution or in other words, false positives.

### 2.3.4 Functional analysis of Hominidae-specific HCNSs

Genetic variation is a suitable indicator of selective constraint on a sequence. I investigated the frequency of SNPs, deletions and insertions overlaid on the Hominidae-specific 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 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 2-3a).

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. I found a higher proportion of Hominidae-specific HCNSs having lower derived alleles than random expectation. This suggests that Hominidae-specific HCNSs are under purifying selection (Figure 2-3b). At the level of DAF<0.1, HCNS showed a significant excess of rare-derived polymorphisms compared to random expectations (fisher test p-value: 0.004) and by comparing all categories I noticed a significant shift in Hominidae-specific HCNS polymorphisms' allele frequency toward rare allele frequencies (chi squared p=0.001).

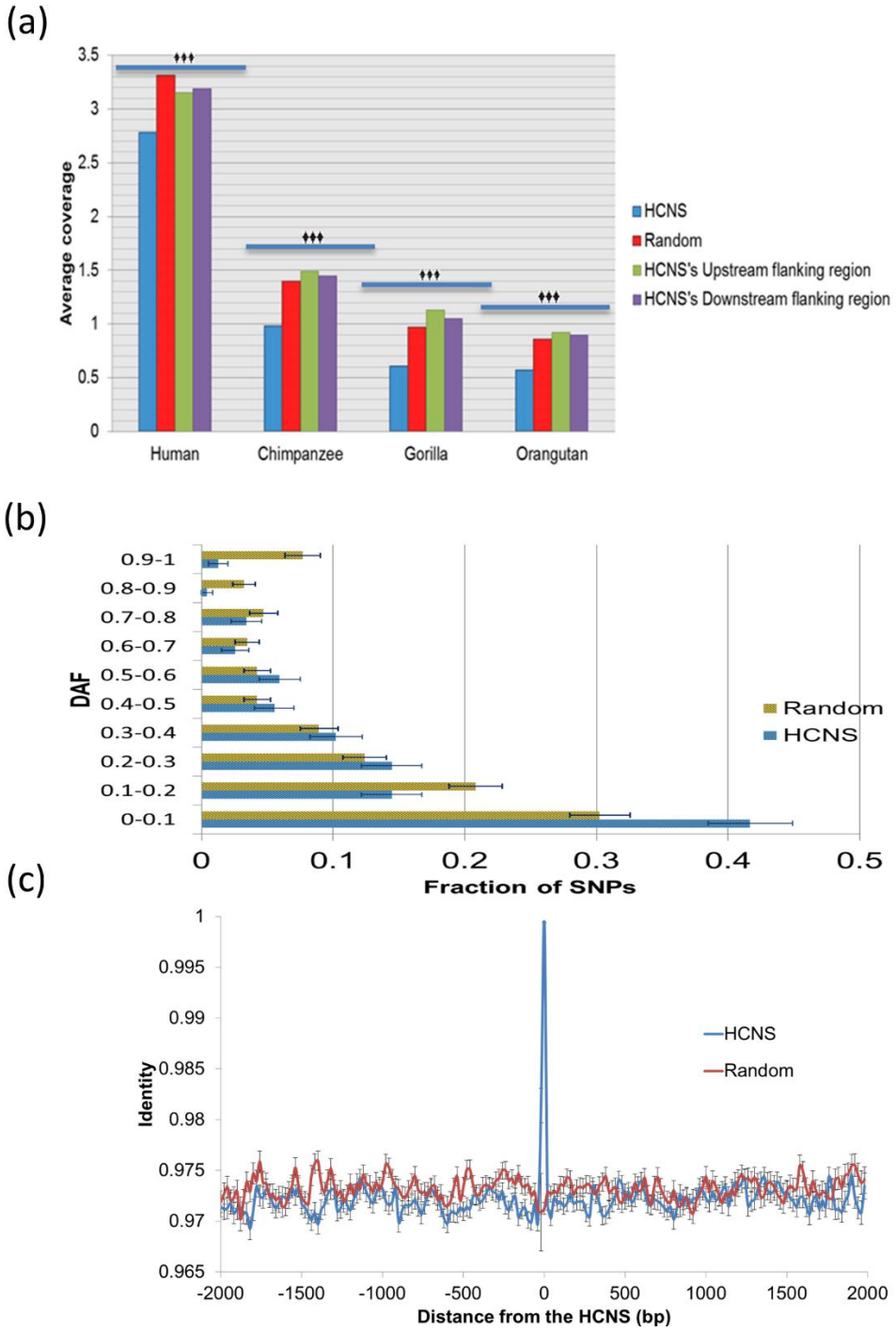



Figure 2-3. The polymorphism coverage and DAF analysis of HS HCNSs.

(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.

Are HS HCNSs located on local mutation cold spot regions in all the Hominidae family? To address this question, I checked the conservation level of the HCNS flanking regions. Figure 2-3c shows the pattern of conservation within 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 HCNSs are highly conserved, indicating that they are under the strong constraints, relative to their flanking regions. I also investigated genetic variation frequency at upstream and downstream regions within the same length of each HCNS that do not overlap with known coding sequences. The genetic variations at 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 2-3a). These results indicate that Hominidae-specific HCNSs are not located on mutation cold spots.

### 2.3.5 Evolutionary origins of Hominidae-specific HCNSs

How did Homindae-specific HCNSs emerge? It was needed to compare outgroup species sequences of HS HCNSs to answer this question. Using whole genome mapping data, 32% (527) of Hominidae-specific HCSNs were mapped to gibbon and rhesus macaque genomes while the rest could not be mapped to both of these outgroup species genomes. I thus examined 527 multiple alignments of three sequences. Length size distribution analysis revealed that average length difference of the mapped sequences in gibbon and rhesus macaque genomes from HCNSs is significantly higher than that of random sequences (See *Appendix A13*).

I also estimated substitution rates (/site/year) at three branches (α, β, γ) of Figure 2-4A for Hominidae-specific HCNSs orthologous and ancestral sequences using mapped gap-removed alignments. Divergence time estimates shown in Appendix A1 are used for rate estimations. I particularly focused on branch on branch α of the phylogenetic tree shown in Figure 2-4A,

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 2-4a), which is five times higher than that (1.1 x$10^{-9}$ ) 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 2-4a).

A very high mean substitution rate for branch α of Figure 2-4a suggests the existence of positive selection at this branch followed by purifying selection in the later Hominidae lineages. I therefore examined the distribution of substitution numbers at branch α for those 527 Hominidae-speicifc HCNSs, as shown in blue bars of Figure 2-4B. Red bars of Figure 2-4b 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) 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 Hominidae-specific HCNSs experienced positive selection which enhanced their substitution rates.

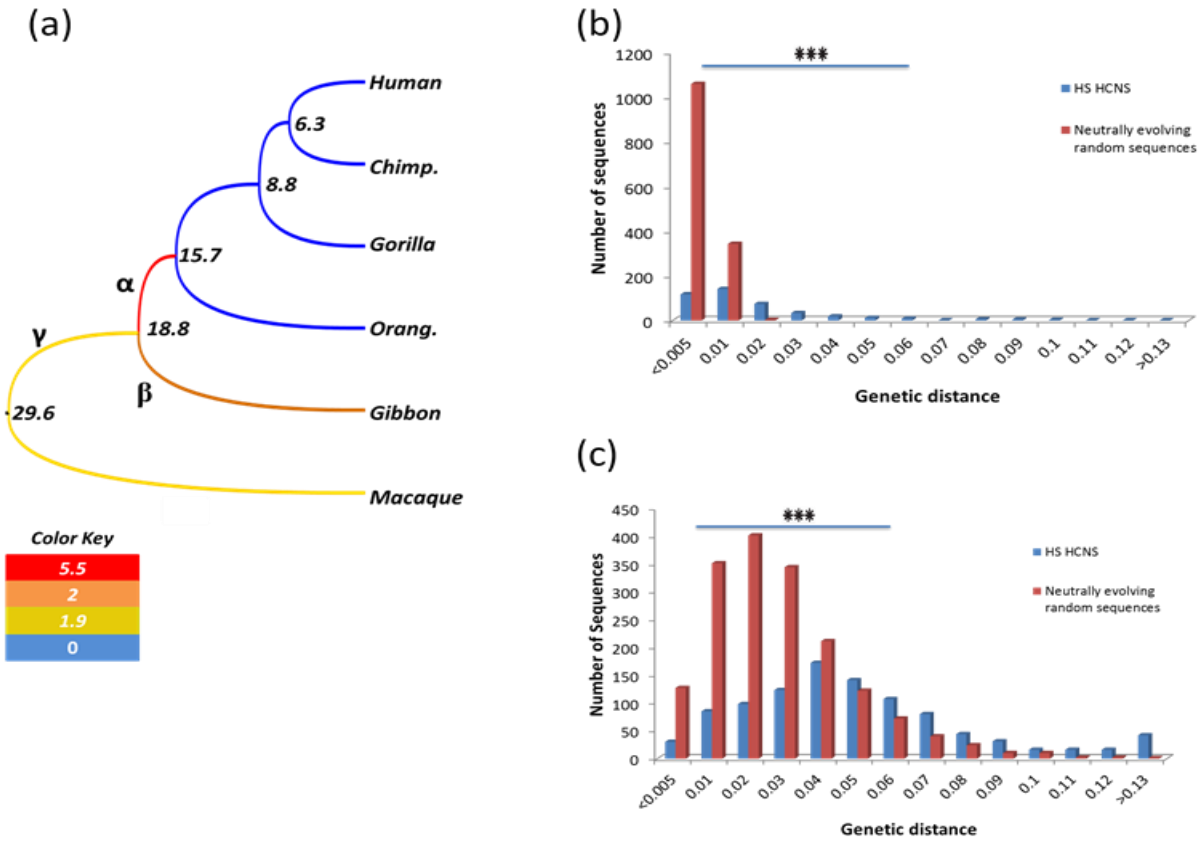



Figure 2-4. Hominidae-specific (HS) HCNS substitution rate across catarrhini phylogenetic tree. (a) catarrhini phylogenetic tree color-coded based on the substitution rate per million year in Hominidae-specific HCSNs orthologous sequences. Nucleotide substitution rates in rhesus macaque, gibbon and Hominoidea common ancestor in HCNS orthologous sequences are significantly higher than that of neutral evolutionary rate. 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 (represented as α) along with (c) 60% of HS HCNS's orthologous and ancestral sequences in Hominidae common ancestor and gibbon (represented as β) with that of random sequences under pure neutral evolution reveals signature of accelerated evolution in Hominidae-specific HCNS orthologous and ancestral sequences.

I found that 527 Hominidae-specific 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, the finding based on only 527 HCNSs can be strengthened. In fact, the mean substitution rate for branches α and β combined for 1,001 Hominidae-speicifc HCNSs was 2.3 x$10^{-9}$/site/year (Figure 2-4C). If we subtract the contribution of branch β from this rate, we obtain the new substitution rate estimate (3.9 x$10^{-9}$/site/year ) for branch α. This value is slightly lower than that 5.5 x $10^{-9}$ for 527 Hominidae-specific 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 HCNSs is shown as blue bars in Figure 2-4C with random regions shown in red bar. In Figure 2-4C, the mode of genetic distance for neutrally evolving sequences is 0.02 while the mode of HCNSs' genetic distance is equal to 0.04 which clearly indicate branch lengths for Hominidae-specific HCNSs to be shifted toward larger ones compared to neutrally evolving sequences.

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 Hominidae-specific HCNSs. I examined neighboring genes (protein coding genes with minimum distance to HCNSs) of these HCNSs with very high substitution rate for branch α of Figure 2-4A. Appendix A13b 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. 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 Hominidae-specific HCNSs. The genomic coordinates and sequences of HCNSs under strong accelerated evolution in Hominidae common ancestor are presented in Appendix A13b.

### 2.3.6 Genomic distribution of Hominidae-specific HCNSs

I 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 HCNS within these categories is shown in Table 2-1. Their distribution significantly differs between Hominidae-specific HCNSs and rest of the genome (P value = 2.2E-16, Chi Square test). The fraction of 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 HCNSs within UTR and promoter regions are consistent with previous findings of the genomic distribution of CNSs in primates (Babarinde and Saitou 2013; Takahashi and Saitou 2012) who reported the notably increased fraction of HCNSs in UTR and promoter regions.

Table 2-1. Fractions (%) of genomic categories in Hominidae-specific 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.

### 2.3.7 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). I computed the nucleosome position probability of Hominidae-specific HCNSs and their flanking regions using the nucleosome prediction probability algorithm developed by Kaplan et al. (2009), 4000bp region from the center of each 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 Hominidae-specific HCNS regions (Figure 2-5a). 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 2-5b). Analysis of H3K9me heterochromatin mark also revealed significant underrepresentation of HCNSs within H3K9me-marked heterochromatin regions (Figure 2-5c). Babarinde and Saitou (2013) discussed the possibility of the association of their low-GC mammalian HCNSs, as also found for Hominidae-specific HCNSs (See Appendix 13C) to nucleosome occupancy, and Kenigsberg and Tanay (2013) found a similar nucleosome positioning pattern in Drosophila HCNSs.

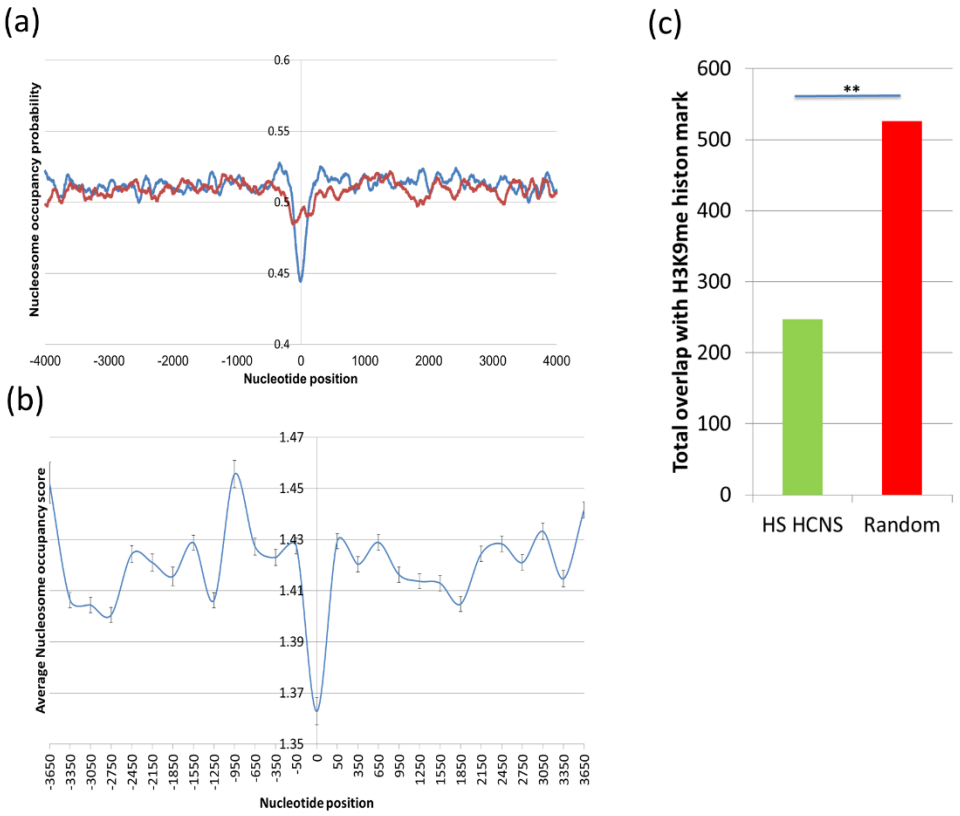



Figure 2-5. Nucleosome occupancy probability for Hominidae-specific HCNSs including flanking regions.

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) HCNS overlap with H3K9me histone mark compared to random sequences. H3K9 methylation is the mark of heterochromatin. Hominidae-specific HCNSs are significantly underrepresented in H3K9me marked regions defined by ENCODE project.

### 2.3.8 Gene ontology analysis

I considered the closest genes to Hominidae-specific 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 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 *Appendix A14*A). This observation is significantly different from the random expectation (p value: 1e-05, empirical chi-square test). 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 Hominidae-specific HCNSs being regulatory elements of their nearby genes could not be ruled out.

Table 2-2 shows the top categories in which Hominidae-specific HCNSs are enriched. The gene enrichment analysis for the 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 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 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.

Table 2-2. 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 |

Consistent with previous analysis of conserved noncoding sequences, genes involved in developmental process are also mainly enriched as likely target genes of Hominidae-specific HCNSs (Table 2-2). *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 Hominidae-specific HCNSs. Among the other likely target genes of 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.

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 Hominidae-specific HCNSs (See *Appendix A14*B) 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). These results indicate the possibility of the evolution of Hominidae-specific HCNSs as the regulatory elements responsible for gene expression switches contributing to specific organ biology of Hominidae family.

Analysis of tissue-specificity of Hominidae-specific HCNSs also revealed that HCNSs have respectively intensified average chromatin immunoprecipitation signal and H3K4me3 epigenetic mark within fetal brain and placenta compared to flanking regions (See *Appendix A14*C). H3K4me3 is associated with active promoter regions. These data are in line with overrepresentation of Hominidae-specific HCNSs in promoter regions and enrichment of developmental process in Gene ontology analysis of likely target genes of HCNSs. These results

give evidence for the likely role of Hominidae-specific 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 Hominidae-specific 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. (Prabhakar et al. 2006) revealed no significant overlap (See *Appendix A14*C). 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 Hominidae-specific 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 Hominidae-specific 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.

## 2.4 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 (Pollard, et al. 2006; Prabhakar, et al. 2006; Sumiyama and Saitou 2011). 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 HCNSs 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; Nei 2013), HCNSs 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 I identified one Hominidae-specific protein-coding gene and 1,658 HCNSs 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. I 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, I assume that my conservative approach for Hominidae-specific novel gene and Hominidae-specific 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 my 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 Appendix A15B), ii)

tissue-specific expression (Uhlen et al. 2015), iii) epigenetic marks for active regulatory region (See Appendix A15C), iv) being a binding site of several transcription factors (See Appendix A15C), v) the likely existence of secondary structures in *DSCR4*-coded protein (See Appendix A5) 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 (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 my 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 my 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 HCNSs 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 (Patterson and Linden 1981), 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 Hominidae-specific protein coding gene to Hominidae-specific 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.

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.

# Chapter 3. Functional analysis of Down syndrome critical region of 4 gene

## 3.1 Introduction

In the past few decades, gene perturbation by overexpression, knockout or knock-down through RNA interference technology have been extensively used to determine function of novel genes, the results of which have significantly impacted multiple areas of medical and biological research (Milhavet et al. 2003). So far, gene perturbation screening has been conducted for dozens of genes in humans and multiple model organisms. Generally, these researches have focused on identification of genes linked with specific biological and physiological phenotypes such as growth rate, viability and cell morphology (Mohr et al. 2014). The development of high-throughput screening methods such as CRISPR/Cas9 targeted genome editing, have further facilitated the comprehensive and accurate identification of critical genes involved in specified phenotypes. The gene perturbation could contribute to the modification of phenotype through interrupting with specific biological pathways or interaction with other key molecule involved in important biological pathways or processes. However, accurate characterization of molecular mechanisms underlying the effects of perturbed gene and the mechanism by which perturbed genes contribute to phenotypic changes have still remained challenging.

Considerable number of studies have conducted whole genome expression profiling by transcriptome analysis after gene perturbation using microarray experiments in order for functional characterization of their genes of interest and also for the identification of biological pathways in which the perturbed genes are involved. For instance, a gene network regulated by SOX2 was uncovered through expression profile analysis of SOX2-knocked out squamous-cell carcionam cells (Boumahdi et al. 2014). In another study, through transcriptome profile analysis

of 147 lincRNA knocked-down samples, it was found that lincRNAs mainly regulate the expression of genes involved in maintaining the pluripotency and repression of the differentiation of embryonic stem cells (Guttman et al. 2011). These studies indicate that the global gene expression modifications occurring due to gene perturbation could be employed to infer the context-dependent function, regulatory networks and cellular cascades in which the perturbed-gene is involved (Xiao et al. 2015).

As it was shown and discussed in detail in Chapter 2, Down Syndrome Critical Region of 4 (DSCR4) gene is the only one bona fide experimentally verified protein coding orphan gene restricted to Hominidae family which is also located on medically important region called Down Syndrome Critical (DSCR) Region. DSCR4's location on Down syndrome critical region and its unique existence in Hominidae lineage which is well-known for unique and novel phenotypes along with its regulated tissue-specific expression, suggest that DSCR4 is a functional gene within the genome of human and Great apes.

Even though the existence of DSCR4 protein is documented at RNA and protein level (Nakamura et al. 1997; Uhlen et al. 2015), there is no experimental evidence reporting the functionality of this gene. Overexpressing wild-type genes in wild-type background has long been used for identification of function of unknown genes in multiple species (Halder et al. 1995; Palatnik et al. 2003; Sokol et al. 1991; Wright et al. 1988). Considering the fact that phenotypic effect of misregulation of Down syndrome critical region (DSCR4) in Down syndrome patient is caused by trisomy and hence overexpression of the genes located on DSCR region; the gene perturbation analysis through overexpression is likely the most suitable approach for functional analysis of unknown genes located on this medically important region, hence, here I used this approach for the identification of biological pathways in which DSCR4 is involved which will in turn pave the way for functional characterization of this gene.

## 3.2 Materials and methods

### 3.2.1 Identification of proper host for DSCR4 functional analysis

Overexpressing wild type genes in wild type backgrounds have been used in multiple model organisms and have been proven to be an efficient approach for functional characterization of novel genes. However, since the gene of interest in this study, DSCR4, do not exist in any model organisms, the overexpression analysis of this gene at organism level is not feasible. Human cells provide a suitable alternative host for gene perturbation studies in cases where organism level overexpression is not feasible or investigation of tissue-restricted effect of gene perturbation is desired. On this premise, human cells were used as the host for functional analysis of DSCR4.

Examining the expression profile of all human cells, for which the transcriptome data are available at GeneInvestigator (Hruz et al. 2008), it was found that human bone marrow cells, HS27a, is the top non-canceric cells that naturally and consistently express DSCR4 at medium to high levels (Figure 3-1). This result indicates the cell-specific expression of DSCR4 and also suggests the functional role of this gene in metabolism of human bone marrow cells represented by HS27a cells.

HS27a, is papillomavirus 16 (HPV-16) E6/E7 transformed long-term human bone marrow cells, and while it is a non-canceric cell, the regulation of cell cycle in this cell is disrupted through amphotropic retrovirus vector, LXSN16E6E7, transformation in the presence of polybrene. It possess epithelial morphology and it is an adherent cell type. HS27a cells were purchased from ATCC® in frozen format and were cultured in RPMI-1640 media according to the protocols provided by the ATCC®.

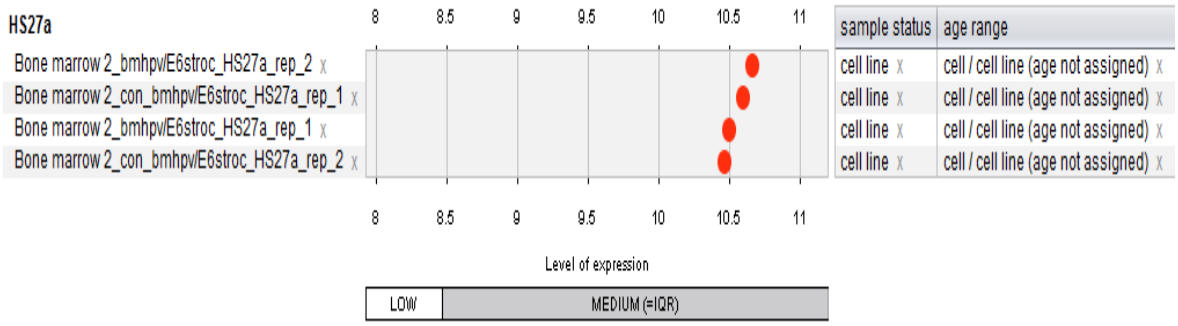



Figure 3-1. Expression profile of DSCR4 gene in HS27a cell.

The gene, DSCR4, is consistently and uniformly expressed naturally in significant levels in all samples of human papillomavirus 16 (HPV-16) E6/E7 transformed HS27a cells whose transcriptome profile is available through microarray analysis. Data retrieved from GeneInvestigator (Hruz et al. 2008).

### 3.2.2 Determining the optimal selection antibiotic concentration

In order for generation of HS27a cells overexpressing DSCR4 gene, the first critical step is optimization of the concentration of antibiotic for selecting stable cell colonies. The optimal concentration of antibiotic is cell-type dependent and varies across different cells, based on their origin, rate of growth and resistance. G418 (also known as G418 sulfate and Geneticin) is an aminoglycoside antibiotic, similar in structure to gentamicin, which blocks peptide synthesis in both prokaryotic and eukaryotic cells. This antibiotic, has so far been widely used as selection reagent in generating stably transfected eukaryotic cells and has been proven efficient.

I conducted a kill curve assay for optimization of the concentration of G418 selection reagent for treating HS27a cells. Kill curve assay is dose-response experiment in which the cells are treated with increasing concentrations of selection reagent to determine the minimum concentration of selection reagent required to kill all the cells over the course of 7 or 14 days. I treated the cell with G418 antibiotic selection marker with concentration gradient ranging from 0 µg/ml (negative control) to 1500 µg/ml. In concentration of 1400, after two weeks, cells were uniformly and completely killed, visible by checking under microscope (Figure 3-2a-b). To further confirm this result, I also conducted cell cytotoxicity assay using cell counting kit-8 (Figure 3-2c-d). Cell Counting Kit-8 (CCK-8) allows convenient kill curve assay using WST-8 (2-(2-methoxy-4-nitrophenyl)-3-(4-nitrophenyl)-5-(2,4-disulfophenyl)-2H-tetrazolium, monosodium salt), which produces a water-soluble formazan dye upon bio-reduction in the presence of an electron carrier, 1-Methoxy PMS, that is present only in living cells. This property of WST-8 allows quantitative measurement of the relative number of living cells in different samples under investigation.

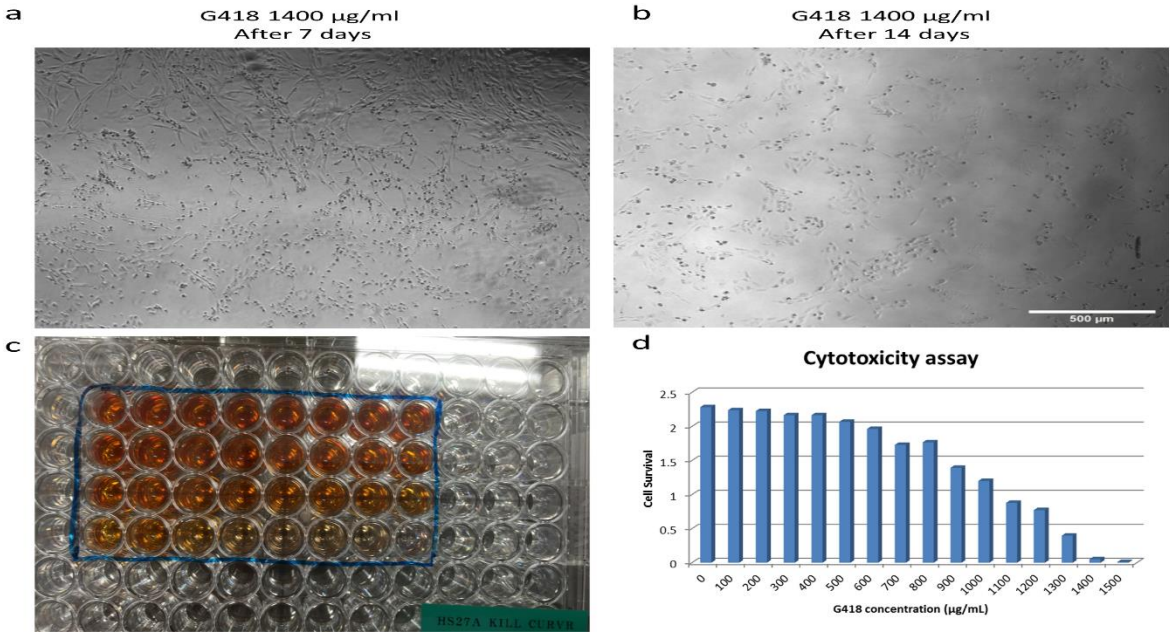



Figure 3-2. Hs27a kill curve assay using G418.

(a-b) Cell cytotoxicity analysis of HS27a cells revealed that the concentration of 1400 µg/ml of G418, is the minimum concentration of this antibiotic that kills all the HS27a cells over the course of 14 days visible under the optical microscope, (c) cell cytotoxicity assay using cell-counting kit-8, revealed color-gradient, from strong red in low concentrations of G418, to color-less in high concentrations of G418 as expected due to abundance of living cells in low concentrations and absence of living cells in high concentrations. (d) Quantitative analysis of cell-counting kit-8 results confirmed the concentration of 1400 µg/ml of G418 to be the minimum concentration required to kill all the HS27a cells over the course of 14 days.

### 3.2.3 Constructing DSCR4-containing carrier

Plasmids are self-replicating extrachromosomal circular DNAs that could contain a wide variety of elements including the genes controlling their own replication to antibiotic resistance genes and genes encoding various enzymes or even toxins. Plasmids are commonly found in bacteria as small circular, double-stranded DNA molecules; however, sometimes they could also be present in archaea and eukaryotic cells.

In the past few decades, artificially constructed plasmids have been extensively used as efficient carries of genetic materials in genetic engineering. There are a dozen types of vectors specifically designed for expression compatibility with different prokaryotic and eukaryotic host cells. PTCN vectors are a class of mammalian cell vectors containing a neomycin resistance marker, designed for expressing a cDNA from the CMV promoter and CMV enhancer in eukaryotic cells. The enhancer and promoter of PTCN vector, derived from cytomegalovirus (CMV), are commonly used in vectors designed for expression in mammalian host cells due to their potent effect in driving gene expression. For construction of DSCR4-cotaining plasmid vector, I incorporated DSCR4's coding sequence (CDS) along with its 3′ and 5′ untranslated regions (UTRs) into the PTCN plasmid (Figure 3-3a). The UTR elements were incorporated into the vector since these elements might be involved in proper expression and localization of DSCR4-coded protein inside the compartmentalized eukaryotic cells. An additional Polyadenylation signal, was inserted at the end of 3′ UTR for guaranteed generation of poly-adenylated DSCR4 mRNA from the plasmid that is a requisite for mRNA translation in eukaryotic cells. Finally, along with CMV promoter and CMV enhancers, NeoR/KanR (G418) resistance and Ampicillin resistance genes were also incorporated into the PTCN vector which act as selection markers in eukaryotic and prokaryotic cells, respectively (Figure 3-3a). A plasmid vector with the

same structure and genomic contents of PTCN-DSCR4 plasmid but without DSCR4 CDS and UTR elements were also constructed as control to minimize the confounding factors (Figure 3-3b).

The constructed PTCN plasmid were transformed into DH5α competent *E. coli* cells. The transformed cells were cultured overnight in ampicillin-containing medium and then the amplified plasmids were harvested. To confirm that constructed plasmid do possess DNA elements with specified sequences, its nucleotide content was analyzed by sequencing which demonstrated that both PTCN-DSCR4 and PTCN-control plasmids possess desired structure and sequence.

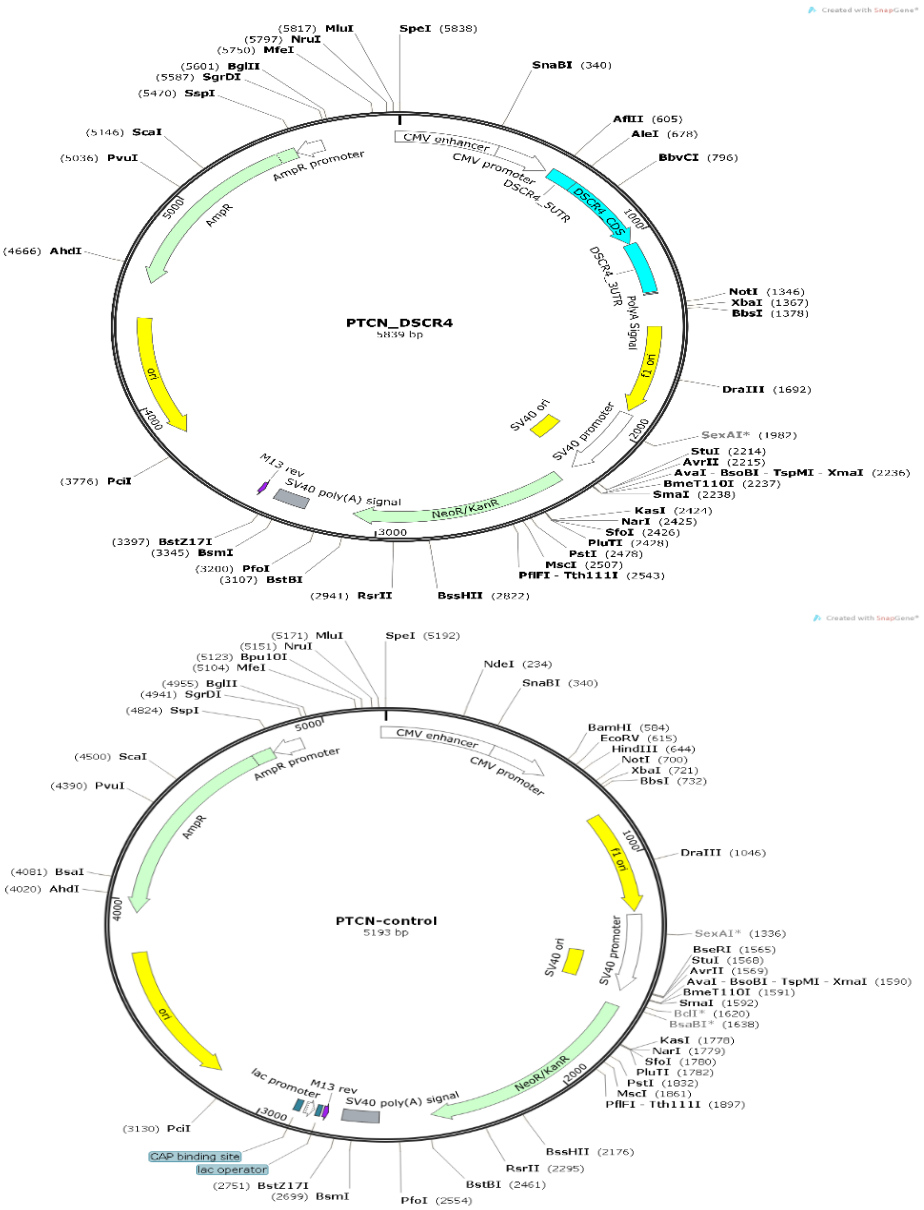



Figure 3-3. PTCN-DSCR4 and PTCN-control plasmid vector construction.

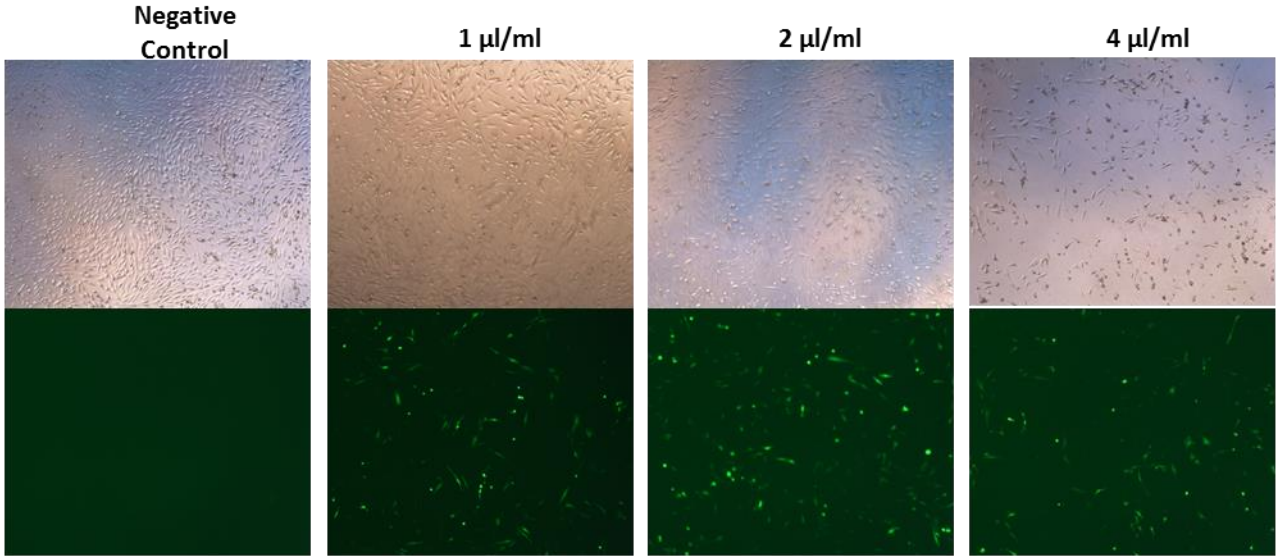
(a) PTCN-DSCR4 vector containing DSCR4 CDS and UTR elements and (b) PTCN-control vector with the structure and genomic content identical to PTCN-DSCR4 but without CDS and UTR elements.

### 3.2.4 Determining the optimal concentration of transfection reagent

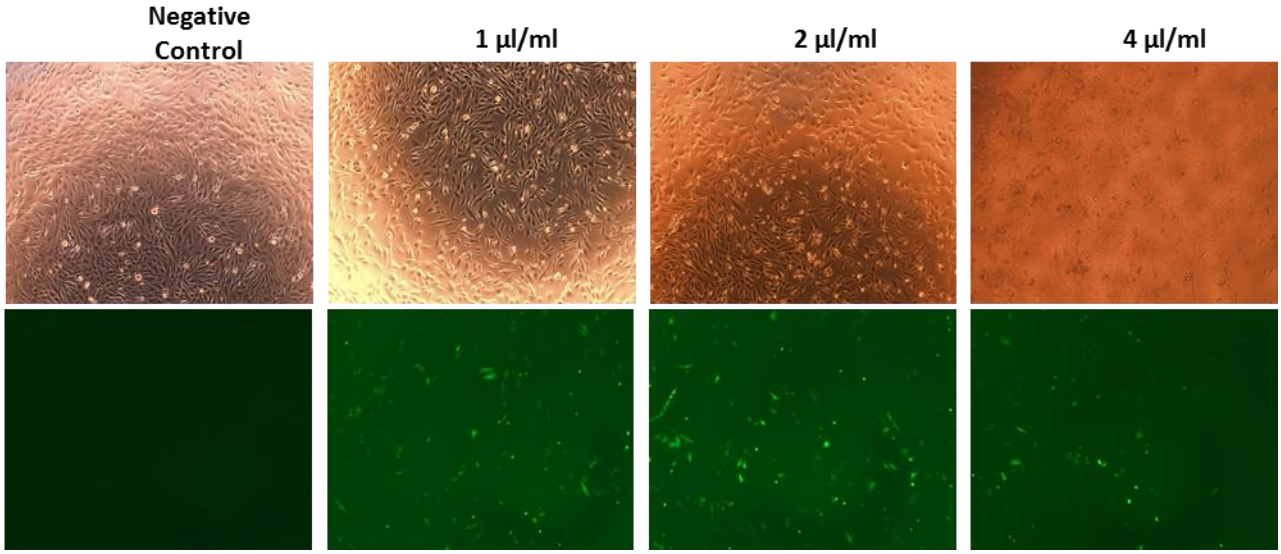
After harvesting and purification of DSCR4-cotaining PTCN plasmid, it should be transfected into HS27a cell's nucleus so that it could merge into the genome. Transferring DNA vectors into eukaryotic cells is a more challenging objective compare to prokaryotic cells due to complex structure and compartmentalized nucleus of eukaryotic cells. There are several types of transfection reagents designed for eukaryotic cell transfection such as lipofectamine, FuGene and omnifect. These reagents differ from each other regarding their toxicity to the host cells, the speed and rate of transfection. According to the nature of transfection reagent and type of host cells, various concentrations of transfection reagent are required for optimized transfection rate. Therefore, to obtain the minimum cell cytotoxicity effect and maximum transfection efficiency, the optimized concentration of transfection reagents should be determined for each types of host cells.

I chose omnifect reagent for transfection of HS27a cells due to its lower toxicity which is required for treatment of eukaryotic non-canceric cells. To measure the optimized concentration of omnifect reagent for transfection of HS27a cells, I used a green fluorescence protein (GFP)-containing plasmid with nearly the same size as PTCN-DSCR4 plasmid and treated HS27a cells with this reagent with concentrations ranging between 1µl/ml to 4µl/ml. The rate of transfection and cell cytotoxicity were determined by measuring the GFP expression using fluorescence microscopy within 3 days. Measuring the GFP signal after 24 hours, 48 hours and 72 hours, it was revealed that Omnifect reagent has the optimized transfection efficiency on HS27a cells at the concentration of 2µl/ml (Figure 3-4).

## 24 hours




## 48 hours



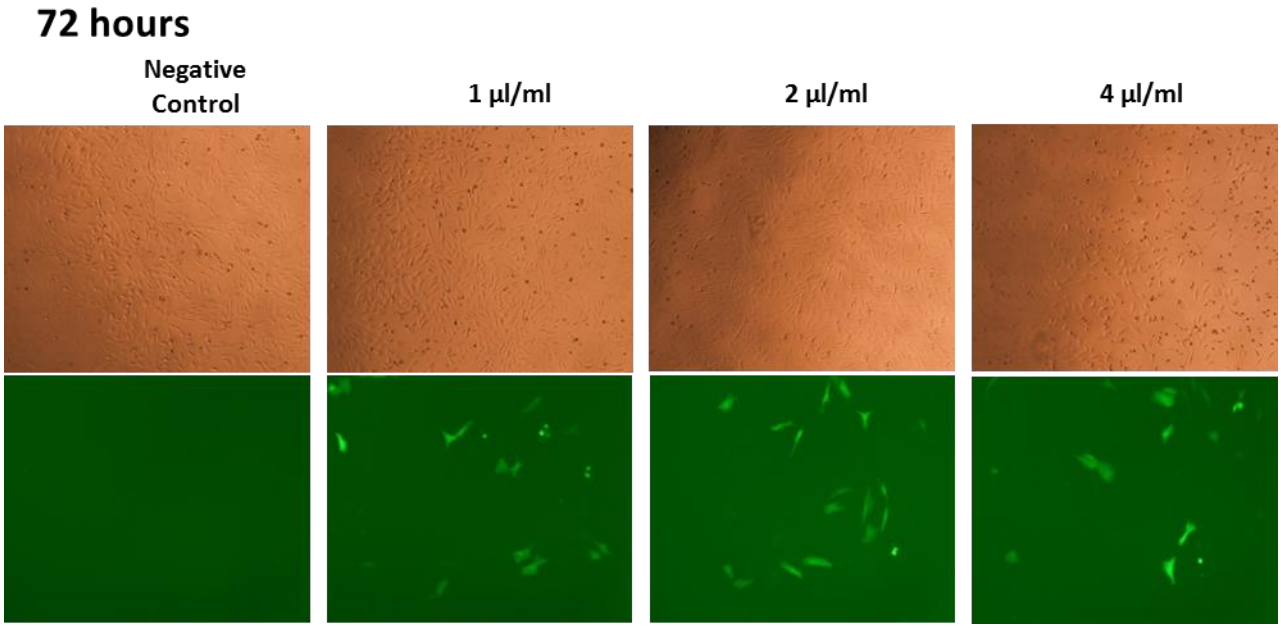



Figure 3-4. Omnifect transfection efficiency on HS27a cells.

Efficiency of transfection of HS27a cells by omnifect transfection reagent was measured using a GFP containing plasmid over the course of 3 days. The concentration of 2µl/ml of omnifect was revealed to have the best balance regarding the cell toxicity effect and transfection rate across all time intervals.

### 3.2.5 Generating HS27a cells stably transfected with DSCR4 and verification of DSCR4 overexpression

After construction of DSCR4-containing and control plasmid vectors and optimizing the concentration of omnifect transfection reagent, in order for the generation of HS27a cells stably transfected with the two prepared plasmids, the HS27a cells were transfected with DSCR4-containing (Figure 3-3a) and control plasmid vectors (Figure 3-3b) using omnifect at concentration of 2µl/ml (Figure 3-4) and treated with G418 at the concentration of 1400 µg/ml (Figure 3-2) over the course of one month.

After treating the PTCN-DSCR4 transfected HS27a cells along with PTCN-control transfected and non-transfected normal HS27a cells with G418 for a month, they were reanalyzed by optical microscope. As expected, PTCN-DSCR4 and PTCN-control transfected cells survived the long-term treatment of G418, indicating that they have successfully incorporated the PTCN plasmid into their genome which encode G418 deactivating enzyme (Figure 3-5). On the other hand, non-transfected HS27a cells were completely killed by G418 due to lack of resistance gene in only a few days (Figure 3-5).

To ensure that PTCN-DSCR4 transfected cells are actually overexpressing the incorporated DSCR4 carrier, the expression values of DSCR4 was measured in PTCN-DSCR4, PTCN-control transfected and normal non-transfected HS27a cells using SYBR-green real time PCR assay. As expected, PTCN-DSCR4 transfected cells were revealed to be overexpressing DSCR4 gene, 3 to 6 folds more than either PTCN-control transfected or normal HS27a cells (Figure 3-6). These results confirm that I could successfully transfect DSCR4 gene stably into the HS27a cells and significantly overexpress it.

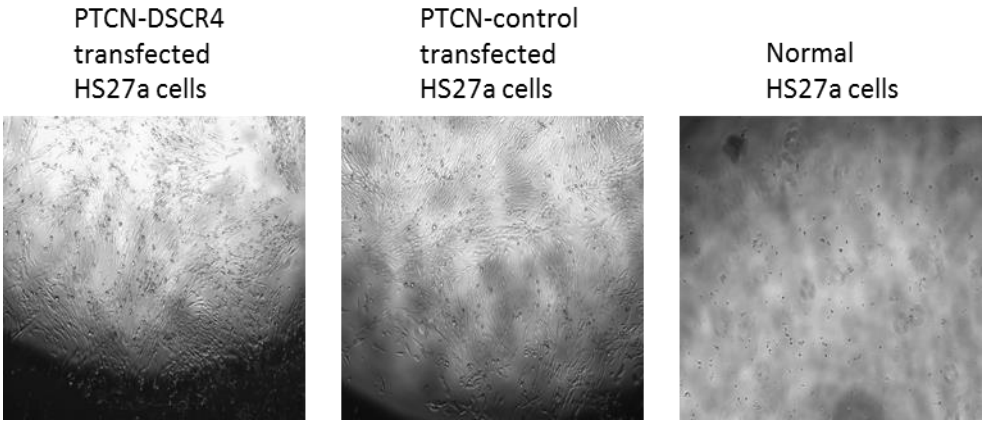



Figure 3-5. Stably transfecting HS27a cells with PTCN-DSCR4 and PTCN-control plasmids.

HS27a cells were transfected with PTCN-DSCR4 and PTCN-control plasmids and treated with G418 over a month. At the end, as expected, the transfected cells could survive the G418 treatment due to gain of G418 resistance gene from PTCN plasmids, while the normal non-transfected cells could not.

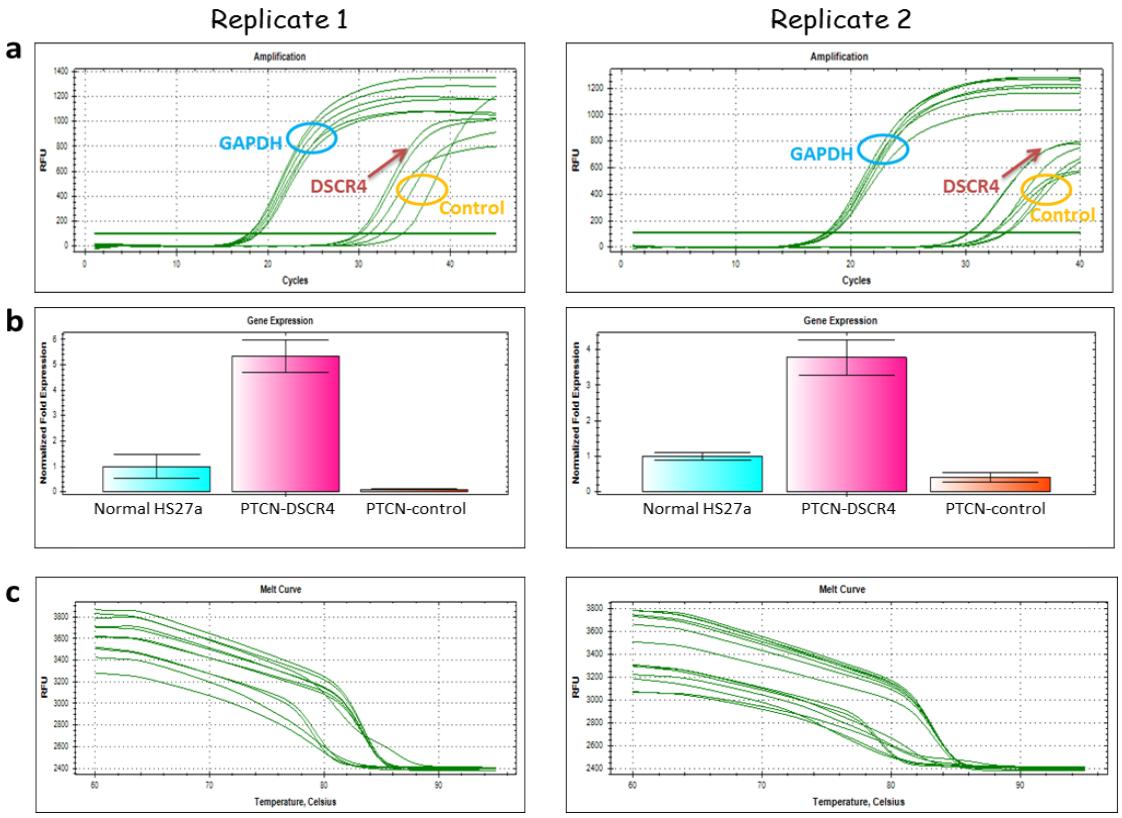



Figure 3-6. Confirmation of the overexpression of DSCR4 in PTCN-DSCR4 transfected cells.

(a) Expression analysis of DSCR4 gene in PTCN-DSCR4 and PTCN-control transfected HS27a cells and normal HS27a cells consistently reveal lower $C_t$ (Threshold Cycle) for DSCR4 gene in PTCN-DSCR4 (specified as 'DSCR4') compared to PTCN-control and normal HS27a cells (specified as 'Control') in two independent replicates. (b) After normalization of the data using the expression values of GAPDH housekeeping gene, DSCR4 gene was shown to be significantly overexpressing in PTCN-DSCR4 compared to both PTCN-control and normal HS27a cells. (c) Melt curve analysis confirmed that there are no significant non-specific band nor primer dimers leading to the erroneous measurement of the expression values of either DSCR4 or GAPDH genes.

## 3.3 Results and Discussion

After confirmation of the successful overexpression of DSCR4 in PTCN-DSCR4 transfected HS27a cells, I conducted whole genome transcriptome analysis to identify what are the molecular effects of overexpression of this gene on human cells. For this end, I purified whole genome RNA from PTCN-DSCR4 transfected, PTCN-control transfected and normal HS27a cells using PureLink RNA minikit. The sizing, quantitation and quality control of extracted RNAs were measured using Agilent 2100 Bioanalyzer and also Nanodrop, which confirmed that extracted RNAs from all three samples are of proper size and quality without any significant degradation (See Appendix A16).

The whole genome transcriptome analysis of PTCN-DSCR4 transfected, PTCN-control transfected and Normal HS27a cells were conducted using Agilent SurePrint G3 Human Gene Expression v3 8x60K microarray chips which provide comprehensive coverage of genes and transcripts using the latest annotation databases. This chip is sourced from RefSeq, Ensembl, UniGene, GenBank, and LNCipedia databases and present full coverage of the whole human mRNAs and also long non-coding RNAs (lncRNAs) using over 58,000 probes.

The results of microarray data were quality checked, background noise subtracted, log-transformed and normalized, then the differentially expressed genes (DEGs) were determined using GeneSpring GX software. DEGs were defined as saturated features (at least 50% of the pixels in a feature are above the saturation threshold) with differences in expression values of more than two-fold ($\log_2$ transformed fold change values of > 1 for over-expressed genes and < -1 for under-expressed ones). Finally, Gene ontology (GO) enrichment analysis of DEGs were conducted and biological processes enriched with DEGs were determined. The enriched biological processes with significant modification in gene expression due to DSCR4 gene

perturbation, are the regulatory networks where DSCR4 is likely to be actively involved, which in turn, provide clues to the functional role of DSCR4 inside human cells.

### 3.3.1 Quality control analysis of microarray results

Microarrays represent a strong technology that provides the capability to measure the expression quantities of thousands of genes simultaneously. However, it enclose several steps that create multiple potential source of variation which if left uncontrolled, can interfere with data analysis and interpretation. Thus, quality control protocols which examine the reproducibility of the results via identification of abnormal or deviating trends, are mandatory in data quality and assay performance of microarrays.

To investigate whether the microarray analysis have been conducted properly and accurately, I conducted two types of microarray quality control analysis, first, scatter plot analysis for verification of quantity of variation among the arrays and second, box-and-whisker plot analysis for investigation of similarity of expression distribution among arrays.

In microarray data analysis, under most experimental conditions less than 10 percent of all genes are expected to change in a biologically relevant way (Alizadeh et al. 2000; Bilban et al. 2002), so deviation of gene expression profile across investigated samples above this threshold after $\log_2$-scaling, is likely to be an indicator of variation caused by artifact which would increase the false-positive results in identification of differentially expressed genes across samples. A scatter plot of gene expression variation across all samples was constructed which revealed that the total significant expression variation across samples are far less than 10 percent (Figure 3-7).

PTCN-DSCR4 transfected HS27a cells and PTCN-control transfected cells show significantly less variation to each other compared to normal HS27a cells (Figure 3-7).These results were expected since PTCN-DSCR4 transfected HS27a cells and PTCN-control transfected

cells have been under the same transfection process and G418 treatment over a month while the normal HS27a was not.

Box-and-whisker plots of the un-normalized expression quantities of each chip give a global overview of the signal intensity distributions. Ideally, all the expression distributions across chips, in an experiment, should have comparable similar distribution even before normalization. Constructing box-and-whisker plot for visualization of expression profile distribution of PTCN-DSCR4 transfected, PTCN-control transfected and normal HS27a cells, revealed highly similar distributions (Figure 3-8), indicating again the proper conduction of microarray experiments for all samples.

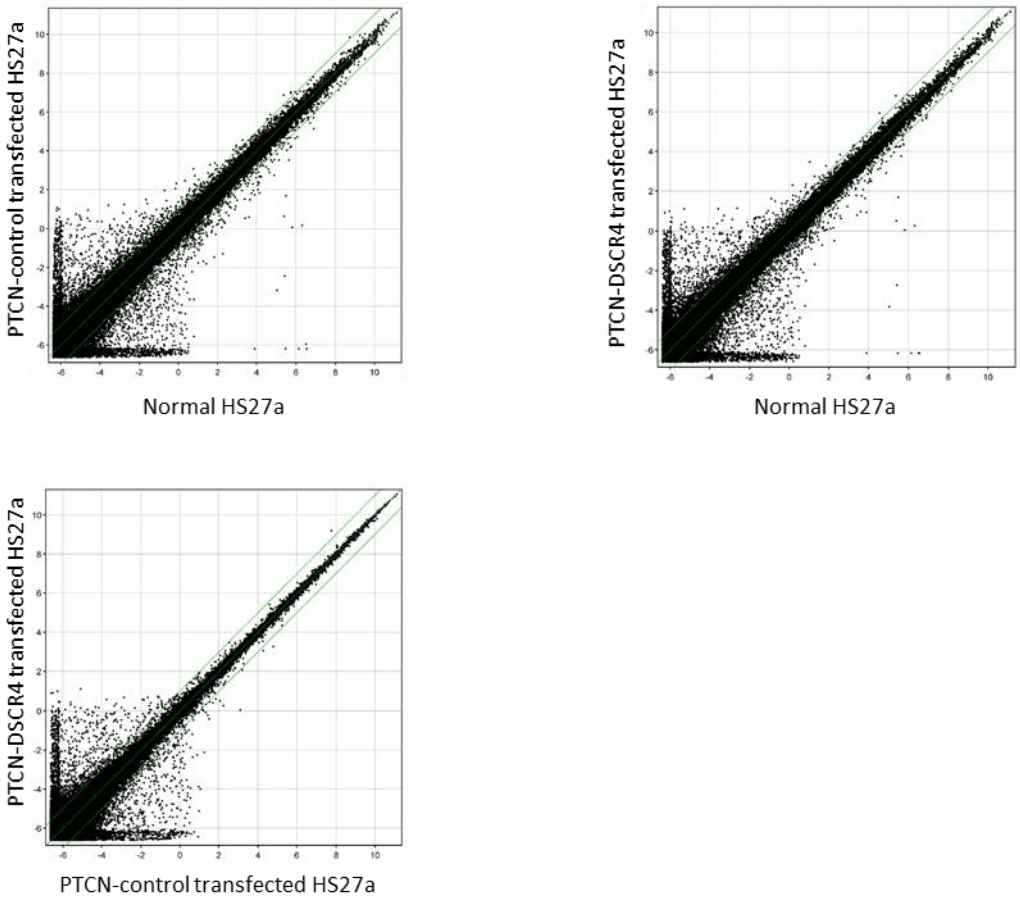



Figure 3-7. Scatter plot of gene expression variation across arrays.

Significant variations in gene expression (two fold expression change as threshold) were observed in less than 5 percent of total genes in all pairwise comparisons indicating the absence of artifacts in the obtained transcriptome data.

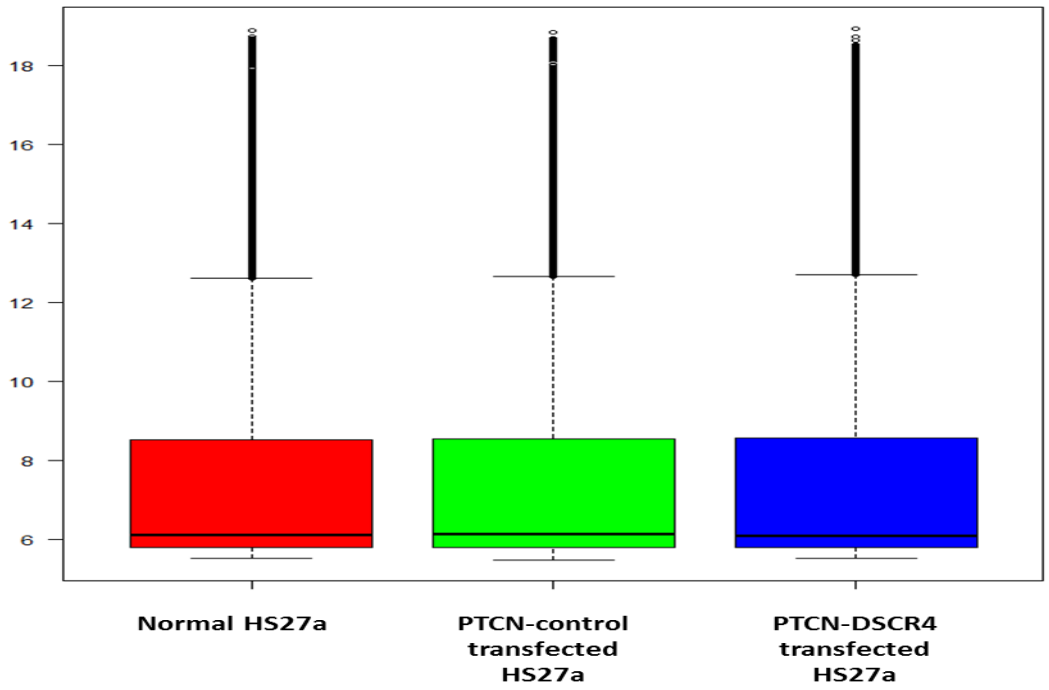



Figure 3-8. Box-and-whisker plot of the un-normalized expression distribution.

The similar distribution of expression quantities across un-normalized microarray results indicate proper conduction of microarray experiments.

### 3.3.2 Identification of DEGs and Gene ontology analysis

After confirmation of the proper conduction of microarray experiments through quality check analysis, the differentially expressed genes (DEGs) were determined using Genespring GX software which is specifically designed for analysis of Agilent array data, provided and recommended by the Agilent Company. A gene is called differentially expressed by GeneSpring software when it is saturated in all the samples under investigation with up-regulated and down-regulated genes having a ratio above a pre-set threshold for significantly higher (≥2 fold change) and lower expression (≤0.5 fold change) intensities.

As it is deducible from scatter plot analysis of array data (Figure 3-7), the treatment processes applied for transfection and permanent incorporation of PTCN plasmids into the HS27a genome, have partially affected the expression of HS27a cell genes. Consequently, comparison of expression profiles of PTCN transfected HS27a cells to normal HS27a cells may yield to identification of DEGs which are due to the processes of cell transfection and treatment and not the overexpression of DSCR4 gene. As a result, to refrain from obtaining false positive results, the PTCN-control transfected HS27a cell's expression profile was only used as control for identification of DEGS in PTCN-DSCR4 transfected HS27a cells. By this approach, of the total human protein coding gene and lncRNA genes present in the array, 310 genes met these prerequisites. Among them, 166 were down regulated and 144 were up regulated.

The DEGs were then further analyzed according to their attributions to known biological functions. Functional annotation tools were used to arrange genes in associated categories based on associated gene ontology (GO) terms and participation in biological pathways. The significantly enriched biological pathways are represented in Table 3-1.

Table 3-1. Gene Ontology (GO) analysis of DEGs obtained by comparing PTCN-DSCR4- and PTCN-control transfected HS27a cells' expression profiles. Significantly enriched categories and enrichment values are shown.

| GO ACCESSION | GO Term | p-value | Fold-enrichment |
|---|---|---|---|
| GO:0035413 | Positive regulation of catenin import into nucleus | 1.37E-05 | 60.58 |
| GO:0043508 | Negative regulation of JUN kinase activity | 3.24E-05 | 46.60 |
| GO:0030334 | Regulation of cell migration | 6.12E-05 | 4.18 |
| GO:2000145 | Regulation of cell motility | 9.22E-05 | 3.99 |
| GO:0040012 | Regulation of locomotion | 1.51E-04 | 3.77 |
| GO:0051270 | Regulation of cellular component movement | 1.60E-04 | 3.75 |

Interestingly, the six GO terms which are significantly enriched in DEGs, represent interconnected pathways in gene regulatory network of human cells (Figure 3-9). Out of the six pathways, four of them (namely, regulation of cell migration, regulation of cell motility, regulation of locomotion and regulation of cellular component movement) are directly involved in the regulation of movement, migration and motility of cell or cells compartments. Regulation of cell migration is defined as any process that modulates the frequency, rate or extent of cell migration, regulation of cell motility is described as any process that modulates the frequency, rate or extent of cell motility, regulation of locomotion is outlined as any process that modulates the frequency, rate or extent of locomotion of a cell or organism and regulation of cellular component movement is defined as any process that modulates the frequency, rate or extent of the movement of a cellular component (Ashburner et al. 2000). The significant fold-enrichment of DEGs related to these interconnected pathways (Table 3-1, Table 3-2) suggest the likely role of DSCR4 gene as a regulatory factor modulating the migration and locomotion of cell or cell compartments.

This prediction is further supported by the fact that the other two enriched pathways in which DEGs are enriched are also indirectly involved in cell migration. The pathway, Positive regulation of catenin import into nucleus, is defined as any process that increases the rate, frequency or extent of the directed movement of a catenin protein from the cytoplasm into the nucleus (Ashburner, et al. 2000). The import of β-catenin from cytoplasm to nucleus with assistance of Wnt protein, leads to activation a signaling pathway named Wnt/β-catenin signaling (Jang et al. 2015). It has been shown that Wnt/β-catenin signaling pathway is involved in cell migration of breast cancer cells and metastasis (Cai et al. 2013; Jang, et al. 2015). The other pathway enriched with DEGs by DSCR4 gene perturbation is regulation of JUN kinase activity. Negative regulation of JUN kinase is described as any process that stops, prevents, or

reduces the frequency, rate or extent of JUN kinase activity (Ashburner, et al. 2000). JUN N-terminal Kinases are a group of kinase enzymes that bind and phosphorylates the c-JUN proteins. C-jun is a proto-oncogene and is the homolog of the viral oncoprotein v-jun (Wisdom et al. 1999). In breast cancer cells, c-jun is known to play key role in migration and invasion of mammary epithelial cells (Jiao et al. 2008; Jiao et al. 2010). In summary, all the enriched pathways and processes with DEGs consistently indicate the functionality of DSCR4 as a regulator of cell migration.

If DSCR4 is involved in cell migration and motility, we would expect this gene to be expressed mainly in cells in which migration is important for functionality. To investigate this hypothesis, I analyzed the expression of DSCR4 in all human tissues for which the transcriptome data is available using Geneinvestigator (Hruz, et al. 2008). As expected, it was revealed that DSCR4 is mainly expressed in human immune system cells (Figure 3-10) where cell migration is critical for proper functioning (Madri and Graesser 2000). Moreover, the placenta, the tissue in which DSCR4 was first identified (Nakamura, et al. 1997), is known to mediate cell migration between mother and fetus (Dawe et al. 2007; Morgan and Wooding 1983).

In conclusion, transcriptome data of DSCR4 gene perturbation analysis along with tissue-specific expression of this genes across human cells, imply the functional importance of DSCR4 in regulation of cell migration. Here, for the first time, I provided evidence and clues for functional importance of DSCR4 gene in human cells, which later, could be subject of direct investigations for further confirmation of the functional role of this mysterious gene.

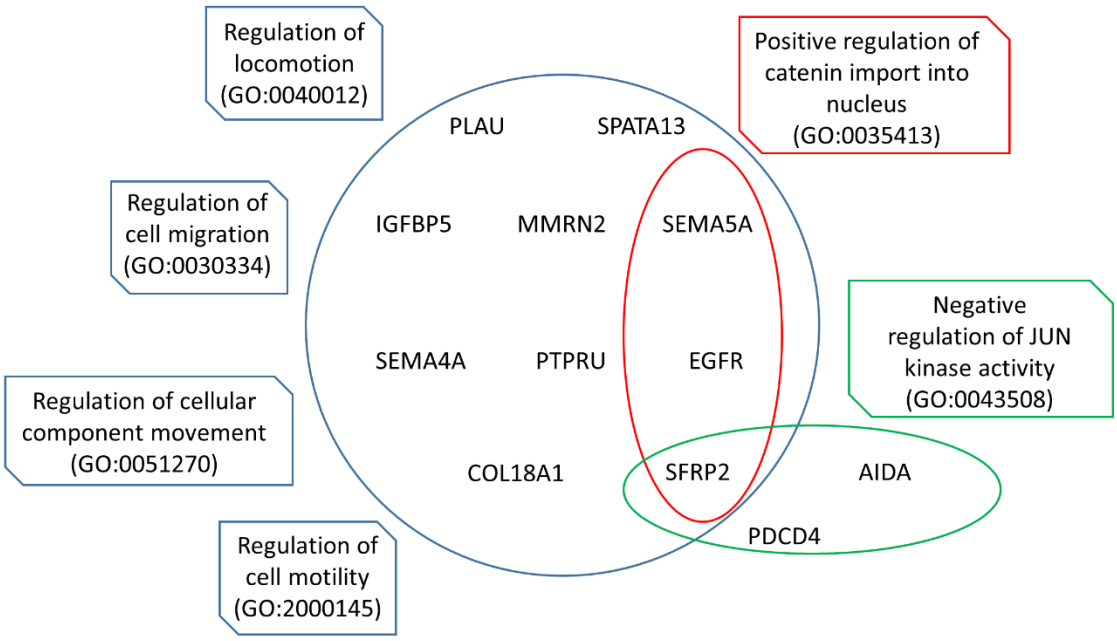



Figure 3-9. Biological processes significantly affected by over-expression of DSCR4 in HS27a cells.

Six biological processes are significantly affected by overexpression of DSCR4 in HS27a cells. All these processes are interconnected and are mainly involved, directly or indirectly, to cell migration.

Table 3-2. Expression data of DEGs involved in biological processes significantly affected by DSCR4 gene perturbation.

| Gene | Control cell raw expression | DSCR4-overexpressing cell raw expression | Control cell normalized expression | DSCR4-overexpressing cell normalized expression | Fold Change ($log_2$) |
|---|---|---|---|---|---|
| EGFR | 228.93 | 40.47 | 0 | -2.53 | -2.53 |
| PLAU | 105.32 | 35.46 | 0 | -1.60 | -1.60 |
| IGFBP5 | 78.64 | 28.34 | 0 | -1.50 | -1.50 |
| SEMA4A | 86.61 | 35.30 | 0 | -1.32 | -1.32 |
| SEMA5A | 4347.68 | 1896.85 | 0 | -1.22 | -1.22 |
| COL18A1 | 498.30 | 237.77 | 0 | -1.09 | -1.09 |
| SPATA13 | 49.29 | 23.92 | 0 | -1.07 | -1.07 |
| SFRP2 | 37.00 | 17.97 | 0 | -1.07 | -1.07 |
| PTPRU | 438.21 | 215.17 | 0 | -1.05 | -1.05 |
| MMRN2 | 41.90 | 20.86 | 0 | -1.03 | -1.03 |
| PDCD4 | 263.84 | 114.82 | 0 | -1.23 | -1.23 |
| AIDA | 46.00 | 23.39 | 0 | -1.01 | -1.01 |

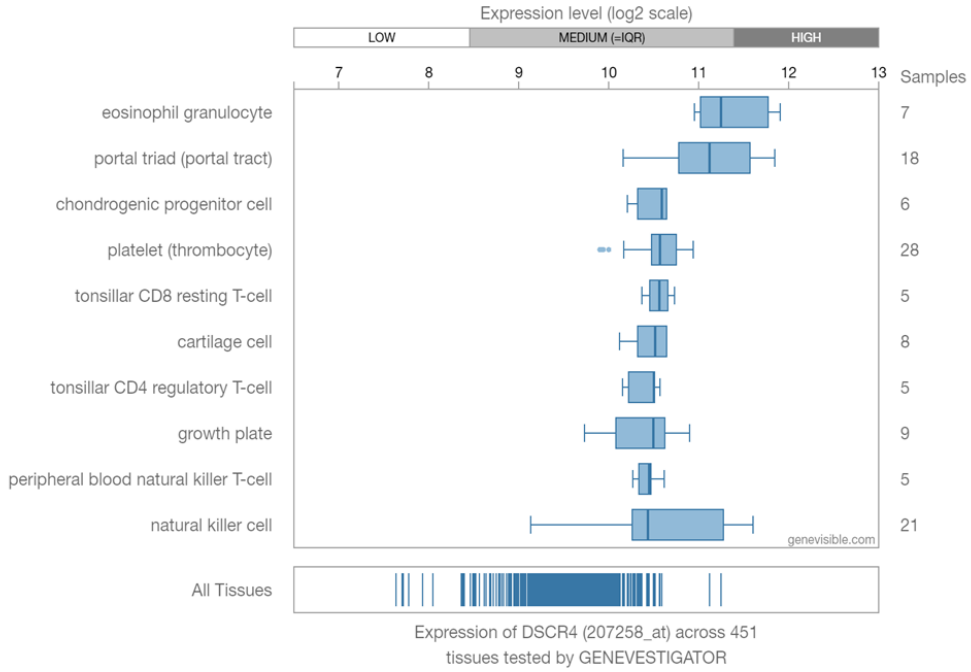



Figure 3-10. DSCR4 expression in human tissues.

DSCR4 is mainly expressed in human immune system cells where migration is critical for proper functioning.

### 3.3.3 Evolutionary importance of DSCR4 in emergence of Hominidae-unique phenotypes

The embryonic evolution and development of complex organisms involves morphogenetic cell movement and migration, during which, tremendous number of cells migrate in a coordinated fashion to form organs and tissues. In adults, cell migration also occurs during the processes such as tissue renewal, wound healing, angiogenesis and is involved in cancer invasion and metastasis. Therefore, the role of DSCR4 in regulation of cell migration could indicate the potential importance of this gene in several aspects of the evolution of humans and great apes.

Phenotypic evolution has been suggested to be primarily mediated by genes involved in embryonic developmental process (Nei 2013) featured by major cell migration (Weijer CJ, 2009). One of the embryonic cell populations highly related to emergence of unique facial characteristics of Hominidae are neural crest cells. Neural crest cells emerge during the first five weeks of gestation from dorsal part of neutral tube ectoderm and migrate into the branchial arches and the region which later turn into the face, consequently, pinpointing the central plan of face morphology. Our analysis of expression of DCSR4 in the data provided by Prescott et al. (2015), revealed significant expression of DSCR4 in human and chimpanzee in neural crest cells. The significant expression of DSCR4 in migrating neutral crest cells in human and chimpanzees are interesting if we consider the non-significant expression of DSCR4 across majority of tissues in these species (Brawand et al. 2011). These results are in line with regulatory roles of DCSR4 in migrating cells and indicate possible roles of DSCR4 as one of regulatory components of the facial morphology of human and great apes.

The children born with Down syndrome typically have some distinctive facial features including almond shaped eyes (due to epicanthal folds); a small, somewhat flat nose, a small mouth with a protruding tongue; and small ears. They may also have round faces and somewhat flatter profiles (Stray-Gundersen K, 1995). Overexpression of DSCR4 is one of genomic factors leading to Down syndrome which indicate the possibility of the contribution of DSCR4 mysregulation in emergence of facial morphology defects observed in Down syndrome patients.

T cell migration and motility is indispensable for T cell-dependent immune response, which allows for the identification of cognate antigens at the surface of foreign antigen-containing cells and also for interactions with other immune system cells. Although previously it was thought that T cell migration is random, a growing number of evidence are emerging to support the notion that T cell migration patterns are strategic and ruled by processes which are optimized for the activation of T cells and also for environment-specific cues (Krummel et al. 2016). Consistent and significant expression of DSCR4 in T cells (Figure 3-10) further support the hypothesis of the functional role of DSCR4 in the regulation of cell migration and suggests the likely role of DSCR4 in evolution of immune system in family Hominidae.

Defective immune system is one of the medical symptoms of Down syndrome patients. Some of the abnormalities of the immune system associated with Down syndrome include: impaired mitogen-induced T cell proliferation, reduced specific antibody responses to immunizations and defects of neutrophil chemotaxis (Ram and Chinen, 2011). Significant expression of DSCR4 in immune system cells along with association of overexpression of DSCR4 with impaired immune system in Down syndrome patients suggest the likely role of DCSR4 in evolution and proper functioning

of immune system in humans and great apes through regulation of the movement of these actively migrating cells.

The evolutionary importance of the emergence of DSCR4 in determining the face morphology and function of immune system in humans and great apes, are in line with gene perturbation and tissue-specific expression data; however, they are yet speculations and need further experimental investigations for confirmation.

# Chapter 4. Unique features of highly conserved noncoding genomic sequences in Hominoidea

## 4.1 Introduction

Identification of the molecular basis of phenotypic evolution has been an active area of 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 play crucial roles in the evolution of species (Nei 2007). Complex interaction of genes associated with development forms gene regulatory networks which are involved in signaling pathways producing phenotypes (Davidson 2006) and the number of genes involved in the network generally increases as the phenotypic character involved becomes more complex.

The evolution of the complex systems of gene regulatory networks has occurred, at least partly, by mutational changes of the regulatory regions which control the spatial and temporal expression of surrounding protein coding genes. In fact, there are dozens of examples for phenotypic changes generated by mutations in cis-regulatory elements such as breadth and length of Darwin's finches' beak (Abzhanov et al. 2004) and stickleback fish's pelvic fins (Wray 2007). Therefore, cis-regulatory elements seem to have played crucial roles in the evolution of phenotypic characters. This hypothesis is further supported considering the small degree of amino acid differences between closely related species with dramatic phenotypic differences (King and Wilson 1975). Moreover, it is believed by many developmental biologists that mutations in cis-regulatory elements are of more importance compared to changes in protein

coding regions, since 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 new species in each phylum are emerged by mutational changes of gene regulatory networks in late stages of development (Nei 2007, 2013).

So far, several mechanisms have been proposed to explain the immense phenotypic diversity observed among organisms. At nucleotide level, mutations including all kinds of genetic changes are the only driving force of phenotypic evolution (Nei 2007). Theoretically, the majority of mutations are evolving under neutral or nearly neutral evolution for which natural selection do not affect the spread of the mutation in the population (Kimura 1983). It is possible for a phenotype-associated mutation to be fixed in a population only through neutral evolution (Kimura 1984). However, the genetic changes contributing to the adaptationally 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 conserve the beneficial mutation by eliminating disrupting deleterious mutations reversing the beneficial ones.

To infer directional or purifying selection on a sequence, first the null model of neutral evolution must be rejected (Kimura 1983). In protein coding sequences, the synonymous mutations serve as a convenient proxy for estimation of the neutral evolution rate, then by comparing accumulation rate of nonsynonymous mutations (abbreviated as Ka) to the synonymous rate (Ks), selection can be inferred in protein coding regions. On average, 85% of loci in protein coding sequences of human genome are evolving under purifying selection (Nei 2007). 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 negative selection (Bejerano et al. 2004; Babarinde and Saitou 2016; Hettiarachchi et al. 2014; Matsunami and Saitou 2013; Takahashi and Saitou 2012, Suzuki and Saitou, 2011).

Conserved noncoding sequences (abbreviated as CNSs) are the sequences under selection constraint that been repeatedly shown to be potential cis-regulatory modules, regulating gene expression (Dermitzakis et al. 2004; Nobrega et al. 2003). 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 these elements to be 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 embryonic stage (Benko et al. 2009; Kritsas et al. 2012). Such shared properties suggests that CNSs are potent candidates to be involved in emergence of order-specific phenotypes.

The superfamily Hominoidea which includes humans and apes, is one of the two living superfamilies of catarrhini parvorder, diverged from old world monkeys around 30 million years ago (Mya) (Hedges et al. 2015) (See Appendix 17a). Members of Hominoidea superfamily 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 mainly unclear. Setting neutral evolution thresholds using coding and noncoding genomic sequences, I identified conserved noncoding genomic sequences under accelerated evolution in common ancestor of Hominoidea and under strong purifying selection within all members of this superfamily including humans, chimpanzees, gorillas, orangutans and gibbons and showed their potential role in expression regulation of genes mainly during embryonic brain developmental stage. It has been reported that ancestral CNSs shared by mammals and aves which have evolved more than 300 million years ago are likely to be functioning as enhancer elements up-regulating the close-by protein coding genes during developmental stage (Babarinde and Saitou 2016). Investigating the differences in characteristics of ancestral CNSs and young Hominoidea-restricted HCNSs which have evolved less than 30 Mya, using a combination of evolutionary and statistical approaches, I found that, in contrary to ancestral CNSs, recently evolved HCNSs in Hominoidea tend to have silencing effects on their target protein coding genes.

## 4.2 Materials and Methods

### 4.2.1 Setting thresholds for negative and positive selection

The thresholds of neutral evolution were determined using the same approach used in Chapter 2 (Saber et al. 2016). In order for the identification of Hominoidea-shared negatively evolving sequences, by comparing the human reference genome and three outgroup species, namely rhesus macaque, marmoset and bushbaby, the nucleotide substitution rates were determined in protein coding synonymous sites and whole genome non-repetitive noncoding sequences. The substitution rates in protein coding 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 protein coding synonymous sites and non-coding DNA sequences were respectively considered as neutral evolutionary rate in protein coding and non-coding regions of the genome (See Appendix A17b). The rate of neutral evolution in protein coding 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 action 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 no non-eliminated mutation after divergence of common ancestor of Hominoidea with 100 percent identity in all members, were considered as Hominoidea-shared sequences under negative selection, as in Saber et al. (2016).

### 4.2.2 Dataset resources

The repeat-masked genome sequences of human, chimpanzee, gorilla, orangutan, rhesus macaque, marmoset and bushbaby were retrieved from Ensembl database version 75. The protein-coding sequences (CDS) and lincRNA coordinates were downloaded from Ensembl biomart. Gene orthology data were also retrieved from Ensembl biomart. The vista enhancer elements with verified enhancer activity were retrieved from VISTA enhancer browser (Visel et al. 2007). Tissue expression data were retrieved from Necsulea et al. (Necsulea et al. 2014) and Epigenome roadmap project (Kundaje et al. 2015). The chip-Seq and enhancer RNA data for 47 human primary tissues were also retrieved from Epigenome roadmap project. The genome polymorphism data of phase 3 of 1000 Genome project were used for polymorphism coverage and DAF analysis. For mapping HCNS coordinates in outgroup species' genomes, the liftover chain files were obtained from UCSC Genome browser database.

Repressor and activator elements in human were retrieved from the UniProt database and transcription factor binding sites were retrieved from Encode project. Silencer elements were defined as the intergenic binding site of monofunctional repressors which do not have any overlap with any of the activator elements binding sites.

For each Hominoidae-specific HCNS, at least 10 random sequences with the same length randomly distributed throughout the genome were picked using Mersenne Twister approach implemented in Python random module. The random coordinates were selected so that they do not have overlap with each other and HCNSs. The random coordinates which have overlap with protein coding or repetitive DNA sequences were also discarded.

For extracting Chip-Seq and eRNA signals, The bigwig file of each histone modification signal and enhancer RNA signal were downloaded from Roadmap Epigenome Project. Using

UCSC bigWigToWig tool, the bigwig data were first transformed into wig files. The average score of each nucleotide within HCNS and random coordinates and standard deviations were then calculated using Python SciPy module. These series of methods essentially followed those used in Saber et al. (2016).

### 4.2.3 Hominoidea-specific HCNS Retrieval

Following a similar approached used in Saber et al. (2016) 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 garnetti*i genomes were first masked. Pairwise homology searches using human genome as query against other four Hominoidea and three outgroup species were performed using BLASTn with E-value threshold of $10^{-5}$. The sequences with at least 100-bp length under negative selection in all members of Hominoidea which do not have any orthologs in outgroup species with conservation level above 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. A schematic depiction of the pipeline used for identification of Hominoidea-specific HCNSs is represented in Appendix 17c.

### 4.2.4 Derived allele frequency (DAF) spectrum

The frequency of genetic polymorphisms overlapping my datasets along with the state and frequency of derived alleles were extracted from the VCF files of total human population generated by 1000 Genomes project (Abecasis et al. 2012). The distribution of derived allele frequencies was calculated for HCNSs and random coordinates.

### 4.2.5 HCNS–gene association

For each protein-coding gene in human genome a proximal gene regulatory domain and a distal gene regulatory domain were defined based on the methodology used by McLean et al. (McLean et al. 2010). Proximal gene regulatory domain was defined as the region 5 kb upstream of transcription start site (TSS) into promoter region and 1 kb downstream of TSS into untranslated region (UTR). Proximal regulatory domain was determined regardless of other nearby protein coding genes. For each protein coding gene also the distal gene regulatory domain was defined as 1000 kb region extended at both upstream and downstream of TSS up to the nearest protein coding gene's basal domain. Potential target of each HCNS were determined upon its overlap with the calculated gene regulatory domains.

### 4.2.6 Selection analysis and nucleotide substitution rate estimation

Each HCNS along with randomly picked coordinates in human genome were mapped to rhesus macaque and marmoset genome using UCSC whole genome alignment chain files. Sequences were then aligned using Muscle software (Edgar 2004), alignment gaps caused by insertion and deletions were discarded, phylogenetic tree was constructed using Neighbor joining method and genetic p-distances were calculated for each branch within phylogenetic tree for each sequence using MEGA-CC software (Kumar et al. 2012). Linear molecular clock has been assumed and applied in calculation of nucleotide substitution rates per site per year.

Insertions and deletion rates within HCNS and random coordinates' ancestral sequences were calculated upon measuring the length difference between coordinates in human genome and their mapped sites in rhesus macaque and marmoset.

### 4.2.7 Gene enrichment test

Gene ontology analysis of HCNS target genes were conducted using a similar approach used by Babarinde and Satiou (2016). First, a list of all genes with GO terms (Atotal) were retrieved from Ensembl biomart build 75. Then a list of HCNS potential target genes with GO terms (AHCNS) were 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.

## 4.3 Results

### 4.3.1 Identification of Hominoidea HCNSs

Pairwise whole-genome homology searches were conducted using NCBI BLASTN on coding sequence masked and repeat-masked genomes of human, chimpanzee, gorilla, orangutan, gibbon, rhesus macaque, marmoset and bushbaby to identify HCNSs shared only by Hominoidea. Human genome was used as query. HCNSs in study of superfamily Hominoidea were defined as noncoding sequences in human genome at least 100bp long with 100% identity in chimpanzee, gorilla, orangutan and gibbon that do not have orthologous sequences in rhesus macaque, marmoset and bushbaby with conservation level above neutral evolution. 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 outgroups species that occurs due to blastN software errors, each HCNS in human genome was also individually mapped to rhesus macaque and marmoset using whole genome alignment data. The HCNSs which had conserved orthologs in rhesus macaque or marmoset regardless of being repetitive, were discarded. This approach (Appendix 17c) identified 679 HCNSs uniquely shared by five members of Hominoidea. (DNA sequences of the discovered HCNSs are available upon request to the author)

### 4.3.2 Functional analysis of Hominoidea HCNSs

I focused on human and great apes genomes due to the availability of data for functional analysis. Sequences under functional constraint do have lower mutations not eliminated in the population, which otherwise would disrupt the functional genomic element. Using the genomic polymorphisms available in phase 3 of 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 compare to random expectations (Figure 4-1A) indicating the existence of functional constraint on these elements.

To confirm that the lower evolutionary rate in HCNSs are not due to them being located on mutation cold spots, I conducted derived allele frequency analysis. For regions under purifying selection, derived alleles of mutations would not be able to fix in the population and tend to remain at low frequencies which leads to the excess of low-frequency derived alleles. Analysis of DAF spectra for HCNSs (Figure 4-1B), as expected, revealed HCNSs to have significantly higher proportion of low-frequency derived alleles compare to random coordinates. I also checked the conservation level of HCNS and their upstream and downstream flanking regions in humans and great apes which show that conservation of HCNSs are not extended to either upstream or downstream regions (Figure 4-1C). These results clearly demonstrate that HCNSs are like neither their up/downstream flanking regions nor random coordinates regarding the action of purifying selection and prove the existence of functional constraint on these elements.

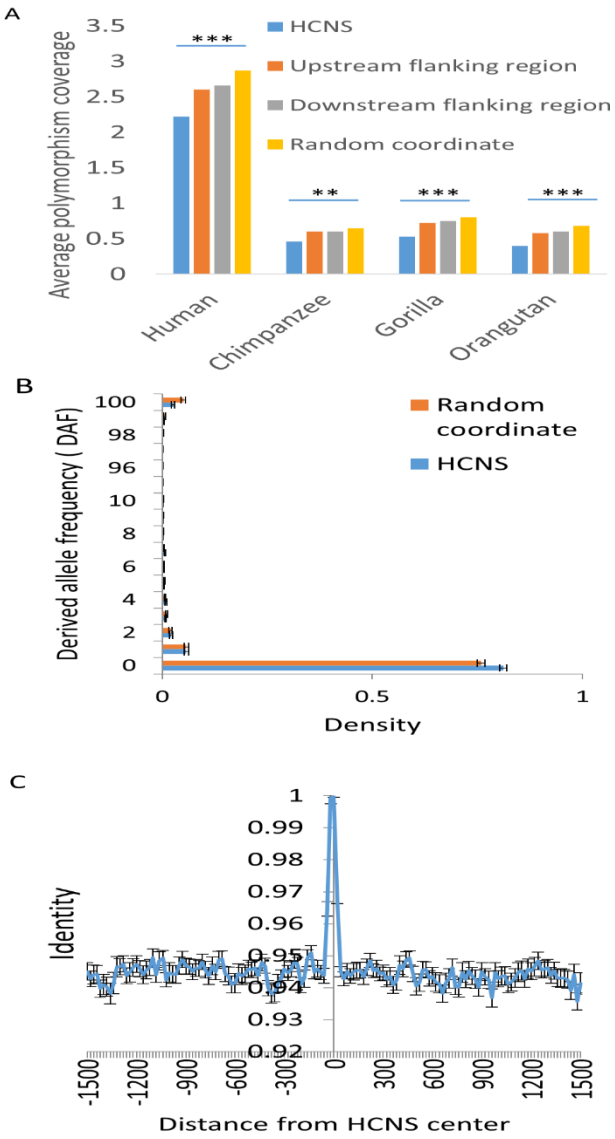



Figure 4-1. HCNSs are under functional constraint in Hominoidea genomes.

(A) HCNSs have lower non-eliminated mutation compare to their upstream and downstream flanking regions and random coordinates. (B) HCNSs have higher frequency of low-frequency derived allele polymorphisms indicating the action of purifying selection (Chi square P value <0.001). (C) Divergence in HCNS flanking regions is equal to the whole genome average (Error bars are 95% CI).

### 4.3.3 Evolution of Hominoidea-specific HCNSs

Having confirmed that HCNSs are under purifying selection, I then asked how these elements have evolved in the common ancestor of Hominoidea. This important question is mostly unanswered in studies of conserved noncoding sequence. Setting and using neutral evolution threshold for identification of HCNSs provide the opportunity for identification of HCNS orthologs in closely related species, it also makes it feasible to analyze and characterize the evolutionary changes occurred at HCNS ancestral sequences during the evolution of common ancestor of Hominoideas. For this investigation, I first mapped each of Hominoidea HCNSs in the two closest species for which whole genome sequencing data are available, namely, rhesus macaque and marmoset. Out of 679 HCNSs, 364 (53.6%) could be mapped in rhesus macaque genome and 352 (51.8%) could be mapped to marmoset genome. Out of the mapped sequences in rhesus macaque and marmoset, 203 (30%) were shared. I then aligned the sequences and calculated the genetic distances between sequences for each mapped HCNS. By constructing phylogenetic tree using the mapped HCNSs in rhesus macaque and marmoset the genetic distance in Hominoidea ancestral sequences were calculated. Same analysis was also performed for random coordinates with the same size but ten times higher in number compare to HCNSs. Deriving nucleotide substitution rate from genetic distances at Hominoidea HCNS ancestral sequences revealed a bimodal graph with one mode at 9E-10 that is nearly identical to the single mode of the nucleotide substitution graph for random coordinates and a second mode at 2.8E-9 that indicates accelerated nucleotide substitution rate at HCNS ancestral sequences for a portion of HCNSs (Figure 4-2B). The total average nucleotide substitution rate at

Hominoidea ancestral sequences for Hominoidea-restricted HCNSs is also 2.38 times higher than that of random coordinates (Figure 4-2A).

Accelerated nucleotide substitution rate observed at Hominoidea HCNSs ancestral sequences was computed using 30% of total HCNSs successfully mapped in rhesus macaque and marmoset genome. To confirm this result with higher confidence using higher proportion of HCNSs, pairwise nucleotide substitution rate between HCNSs in human and their orthologous sequences in rhesus macaque was computed using 53.6% of total HCNSs. Similar to nucleotide substitution rate 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 neutrally evolving random sequences (Figure 4-2C) and second mode at 1.7E-09. These results, in total, give 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, I also probed the rates of insertions and deletions at HCNS ancestral sequences. To this end, the average length difference of HCNSs and their orthologs in rhesus macaque and marmoset were calculated. The same analysis were 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 genome have experienced no insertions or deletions during the evolution of common ancestor of Hominoidea, however, only 17% of HCNSs showed the same characteristic (See *Appendix A18*A). On the other hand, the proportion of HCNSs with orthologous sequences in rhesus macaque and marmoset with length difference above ten nucleotides is significantly higher than that of random coordinates (See *Appendix A18*A). In summary, these results indicate that HCNS ancestral sequences have been under accelerated evolution in aspects of 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.

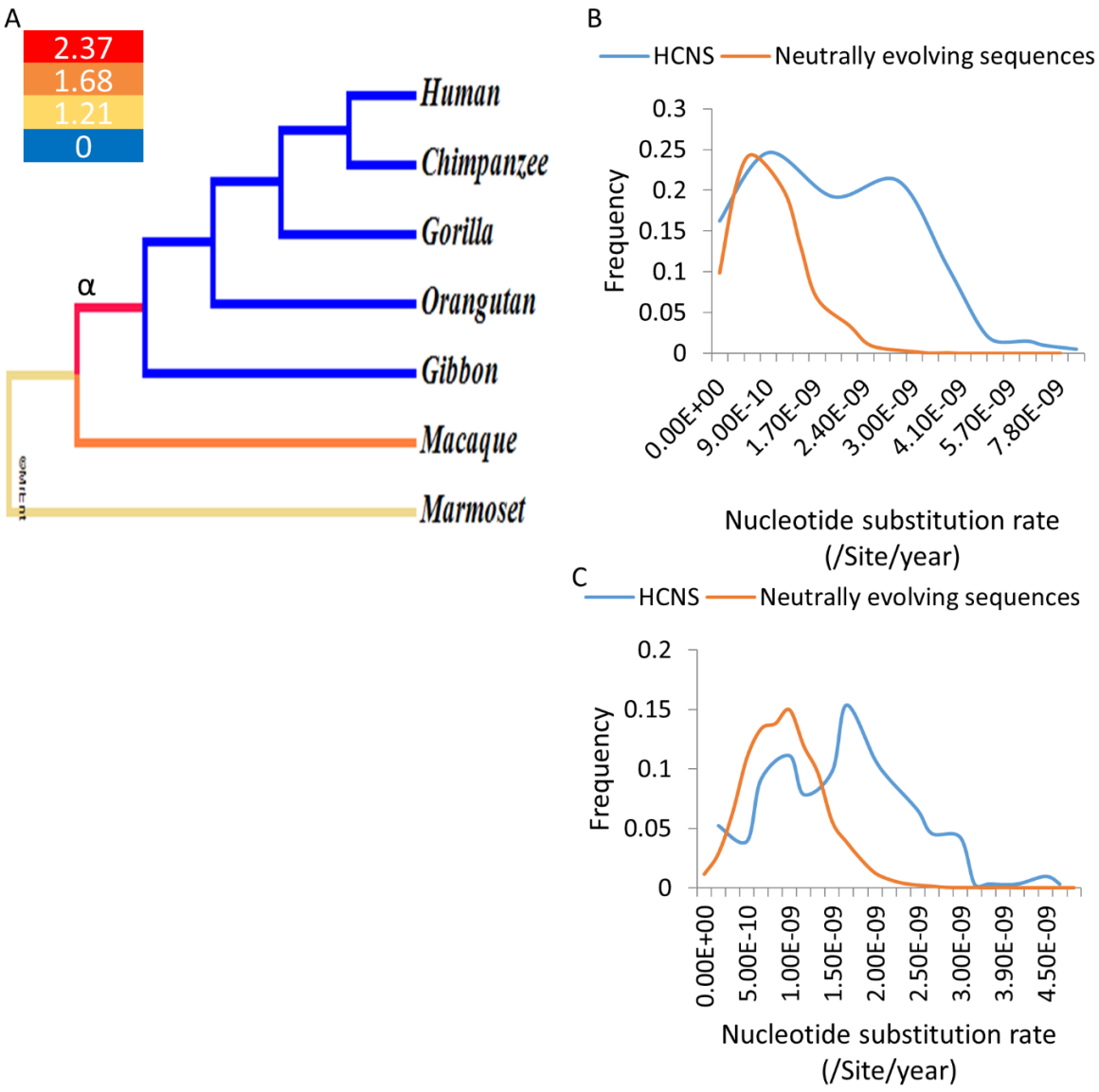



Figure 4-2. Nucleotide substitution rate at Hominoidea-restricted HCNSs' ancestral sequences.

(A) Color-coded phylogenetic tree of simians, representing nucleotide substitution rate ratio to neutrally evolving sequences in Hominoidea-restricted HCNSs' ancestral sequences. Nucleotide substitution rates during the evolution of common ancestor of Hominoidea (α) for HCNSs' ancestral sequences is 2.37 times higher than neutral evolution. (B) Distribution of nucleotide

substitution rates for HCNS ancestral sequences during the evolution of common ancestor of Hominoidea and (C) distribution of nucleotide substitution rates during the evolution of common ancestor of Hominoidea along with rhesus macaque compare to neutrally evolving random coordinates provide strong evidence for accelerated evolution in Hominoidea HCNS ancestral sequences.

### 4.3.4 Examination of Hominoidea HCNSs distribution

Having established that HCNSs emerged through adaptive evolution and conserved by purifying selection, I then analyzed their genomic distribution. I asked whether HCNSs are preferentially located close to protein coding genes. To answer this, I defined proximal regulatory domain (1kb downstream and 5kb upstream of TSS) and 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. Figure 4-3 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 compare to HCNSs, lincRNAs and random coordinates. Conducting genomic distribution analysis using GREAT genomic regions enrichment annotation tool (McLean et al. 2010) also revealed enrichment of Hominidae-restricted HCNSs at distance range of 5-50 kb at upstream and downstream regions compared to random coordinates and vista enhancers (*Appendix A18*B). On the other hand, silencer elements are enriched in close proximity of TSSs at distance ranges of <50kb and underrepresented at farther distances similar to distribution pattern of HCNSs (See Figure 4-3 and *Appendix A18*B). These results, confirm nonrandom distribution pattern of HCNSs within the genome and , demonstrate similarities in genomic locations of HCNSs to silencer elements.

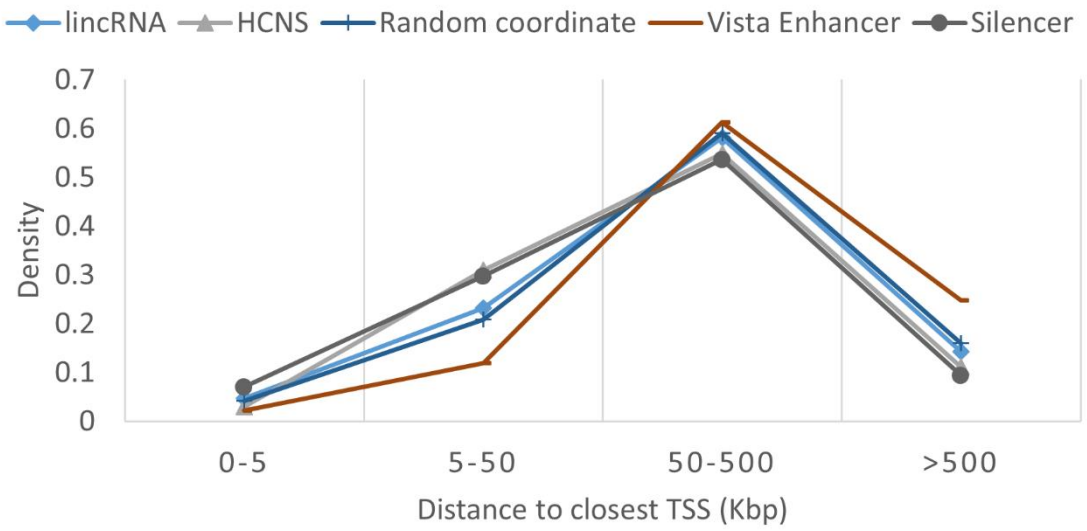



Figure 4-3. Nonrandom distribution of Hominoidea-restricted HCNSs in the human genome.

Hominoidae-restricted 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.

### 4.3.5 Features of Hominoidea-restricted HCNS target genes

Proving nonrandom distribution of HCNSs, I then investigated the properties of HCNS target genes. Conserved noncoding elements have been previously reported to be enriched in close proximity of genes involved in development, transcription and nervous system (McEwen et al. 2009; Saber et al. 2016). Gene ontology analysis confirmed that same enrichment pattern holds for Hominoidea-restricted HCNSs. Hominoidea-restricted HCNSs also tend to be underrepresented in proximity of genes involved in defense and immunity (Figure 4-4A). Unique distribution and pattern of enrichment in gene functional categories of HCNS-associate genes suggest that these conserved elements are likely to be involved in evolution of gene expression especially in the tissue of fetal brain, since at this stage, genes involved in transcription regulation, development and nervous system are mainly expressed.

To investigate the hypothesis that Hominoidea-restricted HCNSs associated genes have unique expression pattern in human tissues following the GO enrichment prediction, RNA-Seq data of human tissues from Roadmap Epigenome project (Kundaje et al. 2015) were retrieved and analyzed. Average RPKM score for all HCNSs target genes along with target genes of random coordinates and vista enhancer elements were calculated across human tissues. As expected, Hominoidea-restricted HCNS target genes have unique expression pattern in embryonic brain, however, surprisingly the expression of HCNS target genes in fetal brain is significantly lower than not only compare to experimentally verified vista enhancer element but also compare to random expectations (Figure 4-4B). For further confirmation, RNA-seq data of human tissues provided by Necsulea et al. (2014) were also retrieved and analyzed. The results of this analysis consistently revealed the same expression pattern across all the investigated tissues (See *Appendix A18*C). These results give clear evidence for association of HCNSs with lower gene expression in their proximal genes during the development of embryonic brain.

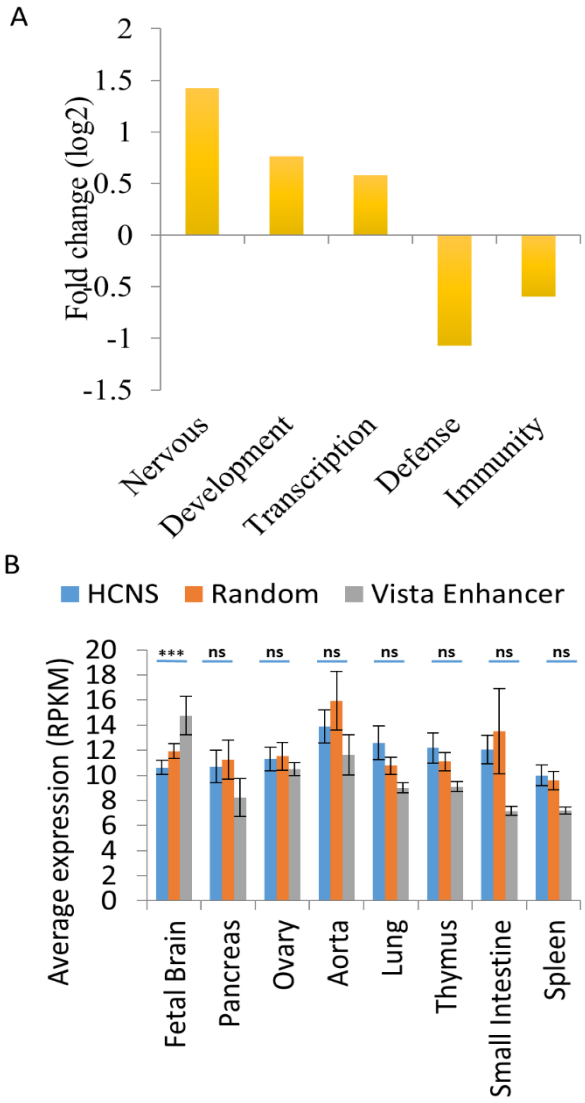



Figure 4-4. Enrichment of HCNS-target genes.

(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).

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, I would expect HCNS target genes to have larger noncoding proportions compared to the genes not targeted by HCNSs. Analysis of HCNS target genes' structure indeed confirmed this hypothesis (Figure 4-5A). HCNS target genes have considerably higher proportion of noncoding sequences (93.79%) compared to the whole genome average of genes not targeted by HCNSs (86.56%).

To figure out whether HCNS target genes have unique distribution pattern in the genome, I also analyzed the distance of HCNS associated genes from their proximal protein coding genes. Analysis of distance to proximal genes demonstrates whether HCNS target genes are located in clusters or in isolation. Calculating the median distance to upstream and downstream flanking genes revealed that HCNS target genes are located dramatically farther away from the nearest protein coding genes compared to the whole genome average of genes not associated with HCNSs (Figure 4-5B). These results indicate HCNS target genes to be unique not only in their structure but also in their location throughout the genome.

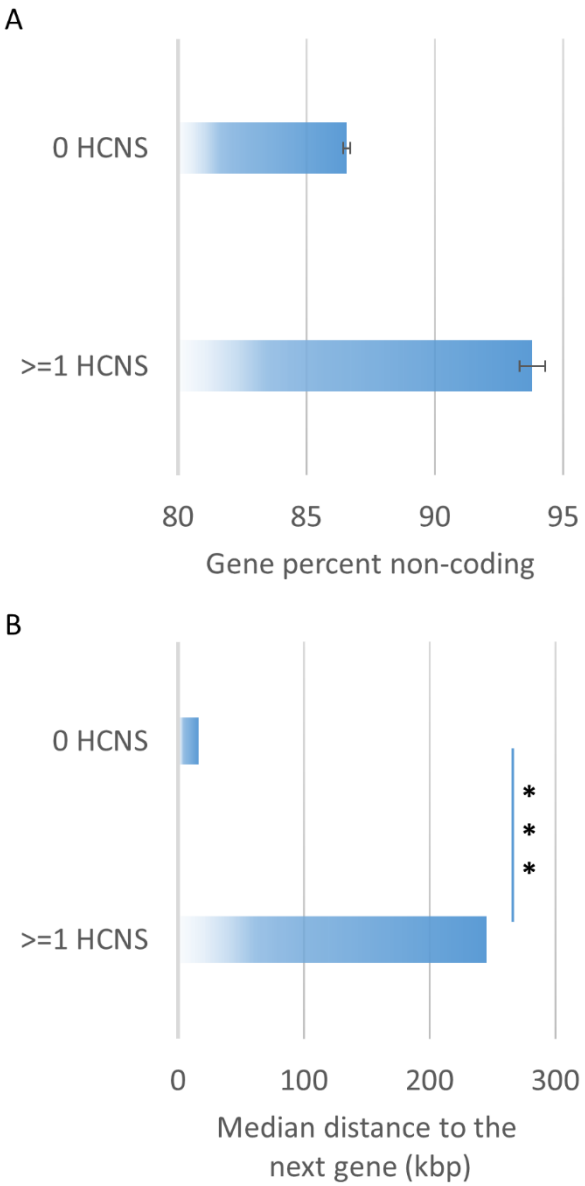



Figure 4-5. Unique features of Hominoidea-restricted HCNS target genes.

(A) Hominoidea-restricted HCNS target genes have significantly higher proportion of noncoding sequences (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 (Mann–Whitney U test P values <0.00001).

### 4.3.6 Epigenomic characterization of Hominoidea-specific HCNSs

Analyzing features of Hominoidea-specific HCNS-associated genes, I 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, I would also expect epigenomic markers for active enhancer elements such as H3k4me1 (Akhtar-Zaidi et al. 2012; Creyghton et al. 2010) to be depleted in HCNSs, especially in fetal brain. To investigate this hypothesis, I analyzed the chip-seq data from roadmap epigenome project. Human tissues for which chip-seq data are available, were classified into four categories, namely, fetal brain, other fetal tissues, adult brain and other adult tissues. As shown in Figure 4-6a, the lowest signal for H3k4me1 in fetal brain was found for HCNSs while the highest signal was found to be for vista enhancer elements. The signal intensity difference between HCNSs, vista enhancer elements lincRNA and random coordinates is the least in adult non-brain tissues. LincRNAs were also shown to have no significant difference from random coordinates in any of the tissue categories (Figure 4-6a).

The pattern of signals for H3K4me3 that is the epigenomic mark for active promoter elements (Cain et al. 2011) is similar to that of H3k4me1 in that HCNSs have the lowest signal in fetal brain and other fetal tissues, however, for H3K4me3 the difference is also visible at adult brain and other adult tissues (Figure 4-6b). The other main difference in signal pattern of H3K4me3 is the signal intensity of lincRNAs which is the highest across all four categories except fetal brain where lincRNAs and vista enhancer elements have equal intensities, both significantly higher than random coordinates and HCNSs. The high signal intensities of lincRNAs for H3K4me3 are as expected probably due to their transcription and proximity to protein-coding genes (Babarinde and Saitou, 2016).

Enhancer RNAs or eRNAs represent a class of bidirectionally transcribed non-coding RNAs 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) which suggest tissue-specific expression of enhancer RNAs. Based on this hypothesis I would expect that 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 several human tissues 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 enhanceric 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 compare vista enhancer elements most significantly at embryonic brain tissue (See Appendix A19a). This result gives further evidence for likely role of Hominoidea-restricted HCNSs as silencer elements in tissue-specific manner.

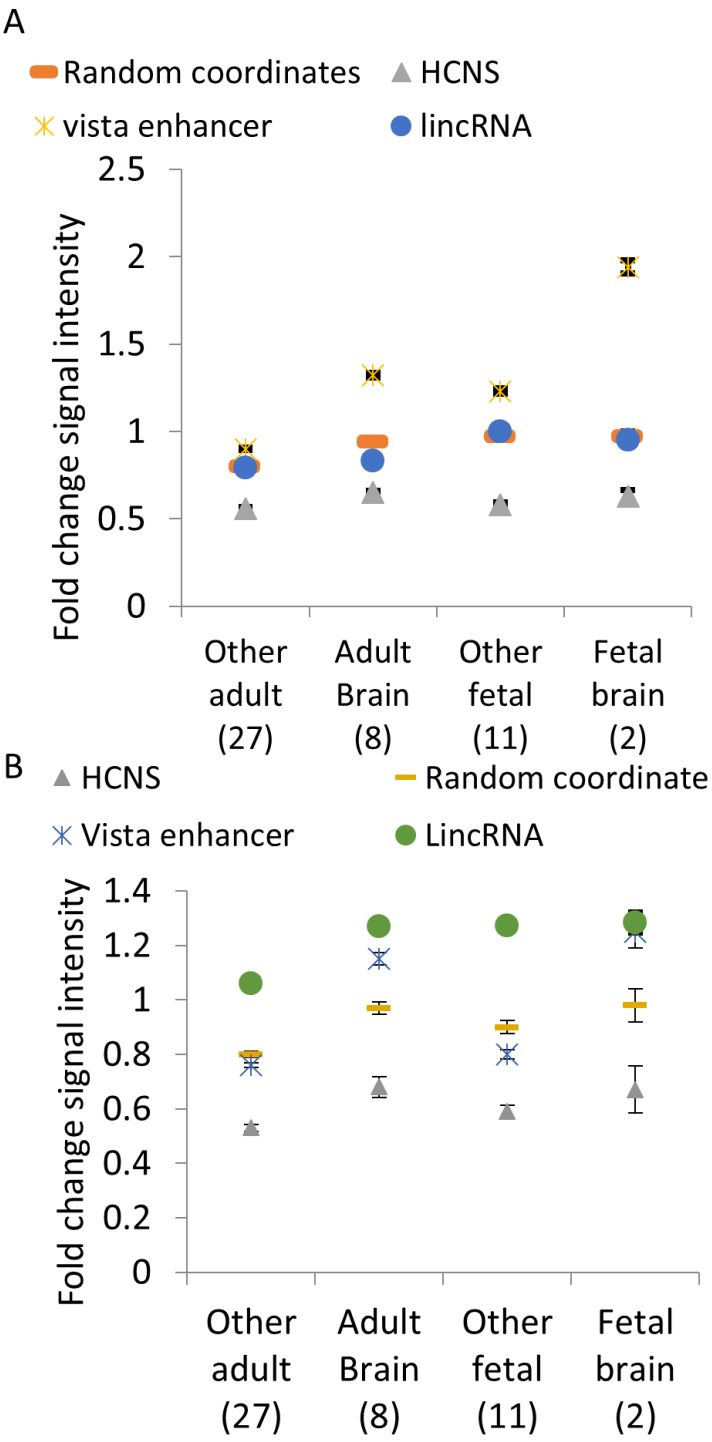



Figure 4-6. Hominoidea-restricted HCNS are depleted in enhancer and promoter epigenomic markers.

(A) Hominoidea-specific HCNSs have remarkably weaker signal for H3K4me1 (enhancer) compare to random coordinates and lincRNAs. The difference is most significant in the tissue of fetal brain. Experimentally verified vista enhancer elements were used as positive control. (B) Hominoidea-specific 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.

## 4.4 Discussion

The superfamily Hominoidea, literary meaning 'Human-like', includes humans and apes which are well-known for their unique man-like anatomy, physiology and cognitive characteristics. The majority of such similarities have been suggested to be result of synapomorphies and a small fraction attributed to homoplasy (Pilbeam 1996). If synapomorphy holds, Hominoidea-restricted functional genomic elements which have evolved in the common ancestor of Hominoidea through adaptive evolution, would account for the majority of Hominoidea-specific characteristics and phenotypes.

Using a computational approach, I have identified 679 highly conserved noncoding genomic elements shared by all members of Hominoidea including humans, chimpanzees, gorillas, orangutans and gibbons. These conserved elements are 100 percent identical in all the investigated Hominoidea members that is significantly higher than the conservation level of protein coding genes, without any orthologs in outgroup species namely, rhesus macaque, marmoset and bushbaby with conservation level above neutral evolution threshold. The stronger purifying selection acting on HCNSs further indicates the functionality of these conserved elements as it proves the constant action of natural selection to eliminate mutations occurring within these sequences. The potent purifying selection acting on HCNSs also demonstrate the critical functional importance of these conserved elements in the evolution and adaptions of Hominoidea members.

Mammalian conserved noncoding elements have been proposed to be classified into two groups with different modes of evolution, first group, consists of HCNSs where a single parameter, models the nucleotide substitution rate throughout the phylogeny (Kim and Pritchard 2007) and second group which departures from the basic model with speed-ups and

slow-downs on particular branches (Doan et al. 2016; Kim and Pritchard 2007). Hominoidea-specific HCNSs mainly follow the evolution pattern of second group for which accelerated nucleotide substitution rate (Figure 4-2) along with accelerated rate of insertions and deletions (*Appendix A18*a) in the common ancestor of Hominoidea is followed by strong selection constraint which has led to absolute conservation of these elements in the superfamily Hominoidea. It has been argued that many of the reported human accelerated regions (HARs) are likely to be simply a result of biased gene conversion (Galtier and Duret 2007). One of the main characteristics of biased gene conversion in mammalian genome 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 elements are not GC-rich but in contrary they are GC-poor sequences (See *Appendix A*19b). Another phenomenon which might interfere with proper calculation of evolutionary rate in Hominoidea-restricted HCNSs ancestral sequences is that HCNSs might occasionally align not with orthologs but with paralogs in outgroup species; however, I 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 formation of Hominoidea-restricted HCNSs is the result of adaptive evolution in common ancestor of Hominoidea superfamily.

Hominoidea-restricted HCNSs are overrepresented in close proximity of protein coding genes, in distance range of 5 to 50 kb from the transcription start sites. In distances less than 5 kb there is no significant differences between Hominoidea-restricted HCNSs and random expectations, lincRNAs or vista enhancer elements and at distances farther than 50 kb and over 500 kb, HCNSs are underrepresented. It is interesting to note that lincRNA distribution is identical to that of random coordinates and at distances farther 5 kb from transcription start

sites, the two of them have distribution frequency standing between HCNSs and vista enhancer elements (Figure 4-3). The significant non-random and contradictory genomic distribution of Hominoidea-restricted HCNSs regarding transcription start sites compare to verified enhancers suggests that HCNSs are not enhancer elements. On the other hand, the distribution pattern of intergenic silencer elements in human genome clearly indicates that in contrary to enhancer elements that tend to be located far away from protein coding genes TSSs, 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 conserved noncoding sequence distribution shared by amniotes reported by Babarinde and Saitou (2016). This discrepancy could be due to the difference in functionality of old CNSs which has evolved more than 300 million years ago in amniotes serving enhanceric role in diverse variety of species such as primates, rodents, carnivores and cetartiodactyls compare 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). Regarding enhanceric epigenomic marker, Hominoidea-restricted HCNSs show depletion in tissue-specific manner especially in fetal brain compare to lincRNA and random coordinates while vista enhancer elements revealed to be significantly enriched with this marker in fetal brain as expected (Figure 4-6). Analysis of H3K4me1 promoter marker enrichment, while again showed significant depletion for HCNSs, however, the depletion showed no tissue-specificity (Figure 4-6a). These results are consistent with previous studies (Andersson et al. 2014; Leung et al. 2015) and reflects more tissue-restricted mode of function of enhancers compare to promoters. Another main difference observed for H3K4me3 and H3K4me1, is the significant enrichment of lincRNA H3K4me3 promoter marker across all tissues (Figure 4-6b)

which may reflect overlap of lincRNA with protein coding genes promoters or their transcriptional activity. These results indicate that Hominoidea-restricted HCNSs are not serving their roles as lincRNA or enhancers. Analysis of transcription level of Hominoidea-restricted HCNSs further indicates tissue-specific silenced nature of these elements in fetal brain by showing that HCNSs produce significantly less enhancer RNAs not only compare to vista enhancer elements but also to that of random coordinates (*Appendix A*19a).

Hominoidea-restricted HCNSs are likely to have suppressive silencing effects on the expression of their target genes. Specially, Figure 4-4b shows that Hominoidea-restricted HCNSs are associated with tissue-restricted suppressive effect on their target genes during embryonic developmental stage in fetal brain. This result is reproducible using expression data from various samples and various databases (Figure 4-4b and *Appendix A18*c) which clearly demonstrate significantly lower expression of Hominoidea-restricted HCNSs target genes in the tissue of fetal brain. It should be noted however that this observation is based on the assumption of target genes of regulatory elements to be identifiable considering the closest protein-coding genes. Although there have been few reported cases of long-range enhancers such as shh (Lettice et al. 2003) where regulatory elements are located far away from their target genes, however, other studies (e.g. McLean et al 2010, Babarinde and Saitou 2016) demonstrated physical distance to be a proper means for identification of target genes of regulatory elements.

It has been previously reported that Hominoidea members such as human and chimpanzee possess heterogeneous lineage-specific immune response (Barreiro et al. 2010), however, they share similar physiological and anatomical brain characteristics (Bailey and Geary 2009; Volter and Call 2012). These observations could be explained, at least in part, by my results as Hominoidea-restricted HCNSs are shown to be enriched in proximity of genes involved in nervous system but depleted for immunity and defense (Figure 4-4a). Therefore, the observed

differences in immune response and similarities in brain characteristics of Hominoidea might be related partly to HCNSs. It has also been suggested that modifications in temporal and spatial gene expression during development play crucial role in the evolution of species (Nei 2013) and minor changes in noncoding regulatory elements have been shown to have the capability to lead to major changes in gene expression pattern (Leung et al. 2015). The results of my analysis further corroborates this hypothesis by showing that target protein coding genes of HCNSs, some of which show minor differences to non-Hominoidea orthologs, 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 Hominoidea-restricted HCNSs are imposing tissue-restricted silencing effects on their proximal genes which are involved in embryonic brain development. Similar properties were also found for highly conserved noncoding sequences restricted to humans and great apes identified by Saber el al. (2016). The young Hominoidea-restricted HCNSs which have evolved less than 30 Mya are majorly 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). Significantly low GC-content of Hominoidea-restricted HCNSs suggest that these elements may be functioning as tissue-specific transcription factor binding sites. This hypothesis is in line with my findings which suggest strong tissue-specific function of Hominoidea HCNSs.

Although the silencing effect of Hominidae- and Hominoidea-restricted HCNSs are 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.

# Chapter 5. Conclusion and future directions

The genome is an elegant but cryptic source of information. The roughly three billion base pairs which constitutes the genomes of the superfamily Hominoidea and family Hominidae, directly or indirectly, encodes all the instructions for synthesizing the macro molecules which in turn lead to the emergence of unique phenotypes observed in these clades. Sequencing the genome of all members of Hominoidea, provide accurate DNA sequences for each of the ape's chromosomes. However, at present, our understanding of the protein-coding and markedly functional non-coding portions of the genomes which spatially and temporally regulate gene expression underlying the unique phenotypes observed in apes or great apes, is far from being complete. To have a better understanding of apes' genomes, and by extension, the biological events they orchestrates and the way in which they can give rise to such unique characteristics, we need a clear overview of the novel shared genomic information they enclose and the functional importance of these novel elements.

Comparative genomic approach for identification of novel protein-coding and functional noncoding sections of the genome, have been shown to be susceptible for false positive errors (Moyers and Zhang 2015). Using a strict thresholds to eliminate false positive results, here, I identified 679 novel potential regulatory elements in Hominoidea superfamily (apes) along with 1658 novel potential regulatory elements and one novel protein coding gene in Hominidae family (great apes) (Table 5-1).

By conducting a load of genomic, epigenomic and expression analysis, I confirmed the action of purifying selection and unique mode of evolution of novel regulatory elements identified here as highly conserved noncoding sequences. I also showed that the target genes of

these regulatory elements do possess unique distribution pattern in the genome and also unique expression pattern across human tissues.

Although, the computational evidence, provided here, gives strong evidence for the functional importance of the identified novel regulatory elements, further experimental verifications are required for confirmation of functionality of HCNSs. Recent advances in targeted genome editing technologies such as CRISPR/Cas9 which has made the parallel in-vivo investigation of the functionality of dozens of targets feasible, presents suitable approaches for experimental verification of the potential regulatory elements introduced here and can be the next step of this study.

In this study, I also computationally identified a protein coding gene restricted to great apes and showed its multi-step mode of evolution within 100 million years from transposable elements. This functionally unknown gene located on a medically important region, do not show any sign of domain similarity to known protein coding genes, however, there are multiple evidences of its tissue specific expression at RNA and protein level in human and great apes. To unravel the functionality of this gene named DSCR4, I conducted gene perturbation analysis followed by transcriptome profiling which provided evidence for the involvement of this gene in regulation of cell migration. Due to lack of any sort of previous knowledge on functionality of this gene, the gene perturbation approach, was the most suitable methodology feasible for investigating the functionality of this gene and provided the likely pathways and gene regulatory networks in which DSCR4 is involved. The identified biological processes in which DSCR4 is involved, should be further analyzed through direct experimental approaches as the next steps of this research.

Since no known or putative protein domain could be identified in the protein coded by DSCR4, the protein crystallographic experiments are yet another required steps for its full characterization. Since the coding sequence of DSCR4, and its proximal gene, DSCR8, with which it share a promoter, do not show any significant homology to any known protein, characterization of the protein secondary structures and the domains existing in DSCR4 and DSCR8 coded proteins, could lead to the introduction of a new family of functional proteins to which these mysterious genes belong.

Table 5-1. Novel functional coding and noncoding genomic elements identified in this study.

| *Clade* | *Novel regulatory elements* | *Novel protein-coding gene* |
|---|---|---|
| *Hominoidea superfamily (apes)* | 679 | 0 |
| *Hominidae family (great apes)* | 1658 | 1 |

# Acknowledgments

I am using this opportunity to express my gratitude to Professor Naruya Saitou who has been a wonderful Natural Selection through my graduation. I am grateful for his constructive criticism and friendly advices which shaped the future of my research life.

I am greatly thankful to Ministry of Education, Sports, Culture, Science and Technology of Japan for providing me this opportunity to pursue my graduate studies in Japan. I am also thankful to Japan Science Society and Sasagawa Company for providing me the funding for the DSCR4 functional analysis.

I would like to express my gratitude toward Professor Yoshihiro Ito, Dr. Takanori Uzawa and Ms. Marzieh Karimi at RIKEN institute for their great help and support in conducting experimental analysis.

I express my warm thanks to Dr. Isaac Babarinde and Dr. Hetteiarachchi for their support and assistance at National Institute of Genetics.

I am also sincerely grateful for invaluable assistance of Dr. Timothy Jinam, Ms. Masako Mizuguchi and Ms. Iida Ai at Satiou lab at National Institute of Genetics.

Last but not least, I would like to thank my parents, siblings and friends who made my life in Japan more pleasurable and memorable.

# References


Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):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(5689):1462-1465.

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF and others. 2012. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336(6082):736-739.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X and others. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769):503-511.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17):3389-3402.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T and others. 2014. An atlas of active enhancers across human cell types and tissues. Nature 507(7493):455-461.

Arnold K, Kiefer F, Kopp J, Battey JN, Podvinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. 2009. The Protein Model Portal. Journal of structural and functional genomics 10(1):1-8.

Arroyo JI, Hoffmann FG, Good S, Opazo JC. 2012. INSL4 pseudogenes help define the relaxin family repertoire in the common ancestor of placental mammals. Journal of molecular evolution 75(1-2):73-78.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25(1):25-29.

Babarinde IA, Saitou N. 2013. Heterogeneous tempo and mode of conserved noncoding sequence evolution among four mammalian orders. Genome Biology and Evolution 5(12):2330-2343.

Babarinde IA, Saitou N. 2016. Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics. Molecular Biology and Evolution 33(7):1807-1817.

Bai L, Morozov AV. 2010. Gene regulation by nucleosome positioning. Trends in genetics : TIG 26(11):476-483.

Bailey DH, Geary DC. 2009. Hominid brain evolution: Testing climatic, ecological, and social competition models. Human Nature 20(1):67-79.

Barreiro LB, Marioni JC, Blekhman R, Stephens M, Gilad Y. 2010. Functional Comparison of Innate Immune Signaling Pathways in Primates. PLoS Genetics 6(12):e1001249.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304(5675):1321-1325.

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT and others. 2009. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nature Genetics 41(3):359-364.

Benton MJ. 2009. Vertebrate Palaeontology: Wiley.

Bilban M, Buehler LK, Head S, Desoye G, Quaranta V. 2002. Normalizing DNA microarray data. Current issues in molecular biology 4(2):57-64.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. Genome biology 8(6):R118.

Blekhman R, Oshlack A, Chabot AE, Smyth GK, Gilad Y. 2008. Gene regulation in primates evolves under tissue-specific selection pressures. PLoS Genetics 4(11):e1000271.

Boumahdi S, Driessens G, Lapouge G, Rorive S, Nassar D, Le Mercier M, Delatte B, Caauwe A, Lenglez S, Nkusi E and others. 2014. SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. Nature 511(7508):246-250.

Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M and others. 2011. The evolution of gene expression levels in mammalian organs. Nature 478(7369):343-348.

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S. 2008. Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS One 3(5):e2209.

Cai J, Guan H, Fang L, Yang Y, Zhu X, Yuan J, Wu J, Li M. 2013. MicroRNA-374a activates Wnt/beta-catenin signaling to promote breast cancer metastasis. The Journal of clinical investigation 123(2):566-579.

Cai J, Zhao R, Jiang H, Wang W. 2008. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. Genetics 179(1):487-496.

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

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(4):1225-1234.

Call J. 2007. Apes know that hidden objects can affect the orientation of other objects. Cognition 105(1):1-25.
Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer TJ, Herrero J, Roos C, Aken B and others. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature 513(7517):195-201.
Carroll SB. 2000. Endless forms: the evolution of gene regulation and morphological diversity. Cell 101(6):577-580.
Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7(2):98-108.
Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, Brady SD, Southwick AM, Absher DM, Grimwood J, Schmutz J and others. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327(5963):302-305.
Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome research 20(3):393-402.
Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. Nature reviews, Genetics 14(9):645-660.
Chen S, Spletter M, Ni X, White KP, Luo L, Long M. 2012. Frequent recent origination of brain genes shaped the evolution of foraging behavior in Drosophila. Cell reports 1(2):118-132.
Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry 13(2):222-245.
Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA and others. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America 107(50):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. Journal of anatomy 212(4):501-543.
Darwin C. 1872. The Descent of Man, and Selection in Relation to Sex: D. Appleton.
Davidson E. 2006. PREFACE. The Regulatory Genome. Burlington: Academic Press. p ix-xi.
Dawe GS, Tan XW, Xiao Z-C. 2007. Cell Migration from Baby to Mother. Cell Adhesion & Migration 1(1):19-27.
de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome research 15(8):1061-1072.
Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, Antonarakis SE. 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 research 14(5):852-859.
Dermitzakis ET, Reymond A, Antonarakis SE. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nature reviews, Genetics 6(2):151-157.
Dixson AF. 1981. The Natural History of the Gorilla: Columbia University Press.
Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S and others. 2016. Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell 167(2):341-354.e312.
Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET and others. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nature Genetics 38(2):223-227.
Du Y, Zhang J, Wang H, Yan X, Yang Y, Yang L, Luo X, Chen Y, Duan T, Ma D. 2011. Hypomethylated DSCR4 is a placenta-derived epigenetic marker for trisomy 21. Prenatal diagnosis 31(2):207-214.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. Annual review of genomics and human genetics 10:285-311.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5:113.
Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol 61(5):717-726.
Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. Annual review of genetics 41:331-368.
Finn R, Bateman A, Clements J, Coggill P, Eberhardt R, Eddy S, Heger A, Hetherington K, Holm L, Mistry J and others. 2014. Pfam: the protein families database. Nucleic Acids Research 42(D1):D222-D230.
Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends in genetics : TIG 23(6):273-277.
Ghiasvand NM, Rudolph DD, Mashayekhi M, Brzezinski JAt, Goldman D, Glaser T. 2011. Deletion of a remote enhancer near ATOH7 disrupts retinal neurogenesis, causing NCRNA disease. Nature neuroscience 14(5):578-586.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 477(7364):295-300.
Haldane JBS. 1933. The Part Played by Recurrent Mutation in Evolution. The American Naturalist 67(708):5-19.
Halder G, Callaerts P, Gehring WJ. 1995. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. Science 267(5205):1788-1792.
Hannenhalli S, Kaestner KH. 2009. The evolution of Fox genes and their role in development and disease. Nature Reviewe Genetics 10(4):233-240.
Hao L, Ge X, Wan H, Hu S, Lercher MJ, Yu J, Chen WH. 2010. Human functional genetic studies are biased against the medically most relevant primate-specific genes. BMC evolutionary biology 10:316.
Harmston N, Baresic A, Lenhard B. 2013. The mystery of extreme non-coding conservation. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368(1632):20130021.
Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S and others. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22(9):1760-1774.
Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nature Genetics 39(9):1140-1144.
Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Molecular Biology and Evolution 32(4):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 Biology and Evolution 6(9):2527-2542.
Hettiarachchi N, Saitou N. 2016. GC content heterogeneity transition of conserved noncoding sequenecs occurred at the emergence of verterbrates. Genome Biology and Evolution.
Hiller M, Schaar BT, Bejerano G. 2012. Hundreds of conserved non-coding genomic regions are independently lost in mammals. Nucleic Acids Research 40(22):11463-11476.
Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. 2008. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Advances in bioinformatics 2008.
Huerta-Cepas J, Capella-Gutierrez S, Pryszcz LP, Denisov I, Kormes D, Marcet-Houben M, Gabaldon T. 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(Database issue):D556-560.
Jacob F. 1977. Evolution and tinkering. Science 196(4295):1161-1166.
Janes DE, Chapus C, Gondo Y, Clayton DF, Sinha S, Blatti CA, Organ CL, Fujita MK, Balakrishnan CN, Edwards SV. 2011. Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. Genome Biology and Evolution 3:102-113.
Jang GB, Kim JY, Cho SD, Park KS, Jung JY, Lee HY, Hong IS, Nam JS. 2015. Blockade of Wnt/beta-catenin signaling suppresses breast cancer metastasis by inhibiting CSC-like phenotype. Scientific reports 5:12465.
Jareborg N, Birney E, Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome research 9(9):815-824.
Jiao X, Katiyar S, Liu M, Mueller SC, Lisanti MP, Li A, Pestell TG, Wu K, Ju X, Li Z and others. 2008. Disruption of c-Jun Reduces Cellular Migration and Invasion through Inhibition of c-Src and Hyperactivation of ROCK II Kinase. Molecular Biology of the Cell 19(4):1378-1390.
Jiao X, Katiyar S, Willmarth NE, Liu M, Ma X, Flomenberg N, Lisanti MP, Pestell RG. 2010. c-Jun Induces Mammary Epithelial Cellular Invasion and Breast Cancer Stem Cell Expansion. The Journal of Biological Chemistry 285(11):8218-8226.
Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. Nature 413(6855):514-519.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome research 20(10):1313-1326.
Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J and others. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458(7236):362-366.
Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. Proceedings of the National Academy of Sciences of the United States of America 99(9):6147-6151.
Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. Genome research 15(10):1373-1378.
Kenigsberg E, Tanay A. 2013. Drosophila functional elements are embedded in structurally constrained sequences. PLoS Genetics 9(5):e1003512.

Kim SY, Pritchard JK. 2007. Adaptive Evolution of Conserved Noncoding Elements in Mammals. PLoS Genetics 3(9):e147.
Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S and others. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465(7295):182-187.
Kimura M. 1983. The Neutral Theory of Molecular Evolution: Cambridge University Press.
King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188(4184):107-116.
Klenke C, Janowski S, Borck D, Widera D, Ebmeyer J, Kalinowski J, Leichtle A, Hofestädt R, Upile T, Kaltschmidt C and others. 2012. Identification of Novel Cholesteatoma-Related Gene Expression Signatures Using Full-Genome Microarrays. PLoS ONE 7(12):e52718.
Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. Genome research 19(10):1752-1759.
Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genetics 4(8):e1000144.
Kritsas K, Wuest SE, Hupalo D, Kern AD, Wicker T, Grossniklaus U. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. Genome research 22(12):2455-2466.
Krummel MF, Bartumeus F, Gérard A, 2016, T cell migration, search strategies and mechanisms. Nature reviews. Immunology, 16(3):193-201
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.
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(20):2685-2686.
Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ and others. 2015. Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317-330.
Lamason RL, Mohideen M-APK, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE and others. 2005. SLC24A5, a Putative Cation Exchanger, Affects Pigmentation in Zebrafish and Humans. Science 310(5755):1782-1786.
Lee AP, Kerk SY, Tan YY, Brenner S, Venkatesh B. 2011. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. Molecular Biology and Evolution 28(3):1205-1215.
Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human molecular genetics 12(14):1725-1735.
Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen C-A, Lin S, Lin Y, Qiu Y and others. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518(7539):350-354.
Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences of the United States of America 103(26):9935-9939.
Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, Lu SJ, Li XM, Yu Q, Zheng X and others. 2010. A human-specific de novo protein-coding gene associated with human brain functions. PLoS computational biology 6(3):e1000734.
Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, Merkurjev D, Zhang J, Ohgi K, Song X and others. 2013. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature 498(7455):516-520.
Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, 3rd, Zody MC and others. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438(7069):803-819.
Lohoff FW, Weller AE, Bloch PJ, Nall AH, Ferraro TN, Kampman KM, Pettinati HM, Oslin DW, Dackis CA, O'Brien CP and others. 2008. Association between the catechol-O-methyltransferase Val158Met polymorphism and cocaine dependence. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 33(13):3078-3084.
Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. Science 333(6045):1019-1024.
Madri JA, Graesser D. 2000. Cell migration in the immune system: the evolving inter-related roles of adhesion molecules and proteinases. Developmental immunology 7(2-4):103-116.
Martin-Ordas G, Call J. 2009. Assessing generalization within and between trap tasks in the great apes. International Journal of Comparative Psychology 22(1):43-60.
Matsunami M, Saitou N, 2013, Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain, Genome Biology and Evolution, 5(1):140-50.
McBride T, Arnold SE, Gur RC. 1999. A comparative volumetric analysis of the prefrontal cortex in human and baboon MRI. Brain, behavior and evolution 54(3):159-166.

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genetics 5(12):e1000762.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology 28(5):495-501.
Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Research 38(suppl 1):D204-D210.
Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. Nature protocols 8(8):1551-1566.
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 Research 36(11):3746-3756.
Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A and others. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447(7141):167-177.
Milhavet O, Gary DS, Mattson MP. 2003. RNA interference in biology and medicine. Pharmacological reviews 55(4):629-648.
Mohr SE, Smith JA, Shamu CE, Neumuller RA, Perrimon N. 2014. RNAi screening comes of age: improved techniques and complementary approaches. Nature reviews. Molecular cell biology 15(9):591-600.
Morgan G, Wooding FB. 1983. Cell migration in the ruminant placenta: a freeze-fracture study. Journal of ultrastructure research 83(2):148-160.
Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. Molecular Biology and Evolution 32(1):258-267.
Mulcahy NJ, Call J. 2006. How great apes perform on a modified trap-tube task. Animal Cognition 9(3):193-199.
Mulcahy NJ, Call J, Dunbar RI. 2005. Gorillas (Gorilla gorilla) and orangutans (Pongo pygmaeus) encode relevant problem features in a tool-using task. Journal of comparative psychology (Washington, D.C. : 1983) 119(1):23-32.
Nakamura A, Hattori M, Sakaki Y. 1997. A novel gene isolated from human placenta located in Down syndrome critical region on chromosome 21. DNA research : an international journal for rapid publication of reports on genes and genomes 4(5):321-324.
Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505(7485):635-640.
Nei M. 2007. The new mutation theory of phenotypic evolution. Proceedings of the National Academy of Sciences of the United States of America 104(30):12235-12242.
Nei M. 2013. Mutation-Driven Evolution: OUP Oxford.
Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. Science 302(5644):413.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302(1):205-217.
Olson LE, Roper RJ, Sengstaken CL, Peterson EA, Aquino V, Galdzicki Z, Siarey R, Pletnikov M, Moran TH, Reeves RH. 2007. Trisomy for the Down syndrome 'critical region' is necessary but not sufficient for brain phenotypes of trisomic mice. Human molecular genetics 16(7):774-782.
Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer EL. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Research 38(Database issue):D196-203.
Paffenholz R, Bergstrom RA, Pasutto F, Wabnitz P, Munroe RJ, Jagla W, Heinzmann U, Marquardt A, Bareiss A, Laufs J and others. 2004. Vestibular defects in head-tilt mice result from mutations in Nox3, encoding an NADPH oxidase. Genes & development 18(5):486-491.
Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D. 2003. Control of leaf morphogenesis by microRNAs. Nature 425(6955):257-263.
Patterson F, Linden E. 1981. The Education of Koko: Holt, Rinehart, and Winston.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD and others. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118):499-502.
Pilbeam D. 1996. Genetic and morphological records of the Hominoidea and hominid origins: a synthesis. Molecular phylogenetics and evolution 5(1):155-168.
Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A and others. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443(7108):167-172.
Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21(5):676-679.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science 314(5800):786.
Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O/'Connor TD, Santpere G and others. 2013. Great ape genetic diversity and population history. Nature 499(7459):471-475.
Prescott SL, Srinivasan R, Marchetto MC, Grishina I,Narvaiza I and others. 2015, Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest, Cell, 163(1):68-83.
Prufer K, Munch K, Hellmann I, Akagi K, Miller JR, Walenz B, Koren S, Sutton G, Kodira C, Winer R and others. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486(7404):527-531.
Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C and others. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505(7481):43-49.
Pybus M, Dall'Olio GM, Luisi P, Uzkudun M, Carreno-Torres A, Pavlidis P, Laayouni H, Bertranpetit J, Engelken J. 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Research 42(Database issue):D903-909.
Ram G, Chinen J, 2011, Infections and immunodeficiency in Down syndrome, Clinical and Experimental Immunology, 164(1):9-16.
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 Genetics 10(7):e1004525.
Saber MM, Adeyemi Babarinde I, Hettiarachchi N, Saitou N. 2016. Emergence and Evolution of Hominidae-Specific Coding and Noncoding Genomic Sequences. Genome Biology and Evolution 8(7):2076-2092.
Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. Science 312(5780):1614-1620.
Saitou N. 2014. Introduction to Evolutionary Genomics: Springer London.
Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic regulation of nucleosome positioning in the human genome. Cell 132(5):887-898.
Schrauf C, Call J. 2009. Great apes' performance in discriminating weight and achromatic color. Animal Cognition 12(4):567-574.
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 Research 42(Database issue):D922-925.
Semendeferi K, Lu A, Schenker N, Damasio H. 2002. Humans and great apes share a large frontal cortex. Nature neuroscience 5(3):272-276.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S and others. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research 15(8):1034-1050.
Sokol S, Christian JL, Moon RT, Melton DA. 1991. Injected Wnt RNA induces a complete body axis in Xenopus embryos. Cell 67(4):741-752.
Sotomayor M, Weihofen WA, Gaudet R, Corey DP. 2012. Structure of a force-conveying cadherin bond essential for inner-ear mechanotransduction. Nature 492(7427):128-132.
Stray-Gundersen K., 1995, Babies with Down Syndrome: A New Parents' Guide, Woodbine House.
Strick PL, Dum RP, Fiez JA. 2009. Cerebellum and nonmotor function. Annual review of neuroscience 32:413-434.
Sumiyama K, Saitou N. 2011. Loss-of-function mutation in a repressor module of human-specifically activated enhancer HACNS1. Molecular Biology and Evolution 28(11):3005-3007.
Suzuki R, Saitou N. 2011. Exploration for functional nucleotide sequence candidates within coding regions of mammalian genes. DNA Res 18:177-187.
Sweetman D, Munsterberg A. 2006. The vertebrate spalt genes in development and disease. Developmental biology 293(2):285-293.
Symmons O, Spitz F. 2013. From remote enhancers to gene regulation: charting the genome's regulatory landscapes. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368(1620):20120358.
Takahashi M, Saitou N. 2012. Identification and characterization of lineage-specific highly conserved noncoding sequences in Mammalian genomes. Genome Biology and Evolution 4(5):641-657.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution 30(12):2725-2729.
Tay SK, Blythe J, Lipovich L. 2009. Global discovery of primate-specific genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America 106(29):12019-12024.
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 research 22(22):4673-4680.

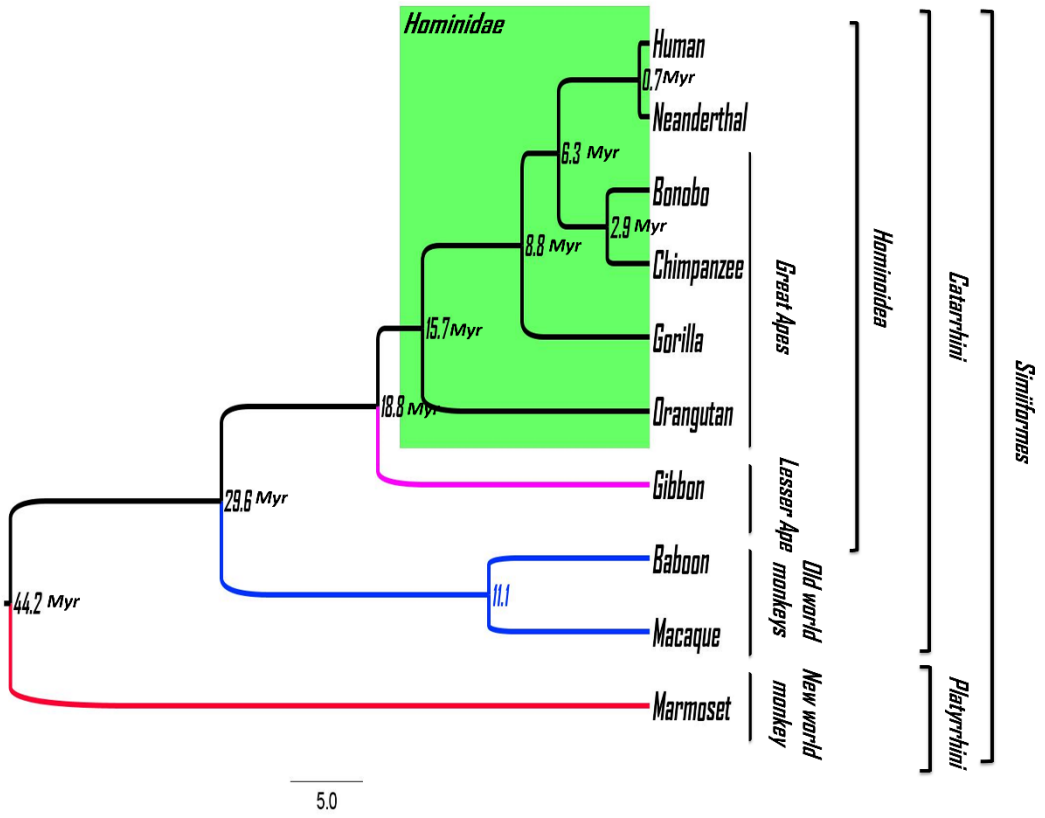
Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Alba MM. 2009. Origin of primate orphan genes: a comparative genomics approach. Molecular Biology and Evolution 26(3):603-612.
Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A and others. 2015. Proteomics. Tissue-based map of the human proteome. Science 347(6220):1260419.
Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Research 35(Database issue):D88-92.
Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. Nature 461(7261):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. Animal Cognition 15(5):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 Research 41(Database issue):D358-365.
Weijer CJ, 2009, Collective cell migration in development., Journal of cell science, 15: 3215-3223
Weng JK, Li Y, Mo H, Chapple C. 2012. Assembly of an evolutionarily new pathway for alpha-pyrone biosynthesis in Arabidopsis. Science 337(6097):960-964.
Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AFA, Finn RD. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Research 41(D1):D70-D82.
Wisdom R, Johnson RS, Moore C. 1999. c-Jun regulates cell cycle progression and apoptosis by distinct mechanisms. The EMBO Journal 18(1):188-197.
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. Proceedings of the National Academy of Sciences of the United States of America 106(18):7273-7280.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nature reviews, Genetics 8(3):206-216.
Wright R, Basson M, D'Ari L, Rine J. 1988. Increased amounts of HMG-CoA reductase induce" karmellae": a proliferation of stacked membrane pairs surrounding the yeast nucleus. The Journal of Cell Biology 107(1):101-114.
Wu DD, Zhang YP. 2013. Evolution and function of de novo originated genes. Molecular phylogenetics and evolution 67(2):541-545.

Xiao Y, Gong Y, Lv Y, Lan Y, Hu J, Li F, Xu J, Bai J, Deng Y, Liu L and others. 2015. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. Scientific reports 5:10889.
Xie C, Zhang YE, Chen JY, Liu CJ, Zhou WZ, Li Y, Zhang M, Zhang R, Wei L, Li CY. 2012. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. PLoS Genetics 8(9):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 Biology 8(10):e1000494.
Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. Genome research 18(9):1446-1455.

# Appendix

Appendix A1. Molecular phylogeny of simians

Hominidae is one of the two living families of ape superfamily Hominoidea; Hylobatidae or lesser apes constitute the other family. Divergence times (measured as million years ago) were retrieved from Time tree knowledge-base. Pink, blue and red colored branches respectively represent lesser apes (Hylobatidae), old world monkeys (Cercopithecidae) and new world monkeys.



Appendix A2.

**a) Properties of Hominoid-specific genes identified by Xie et al. (2012) over time using Ensembl genomic database.**

| Gene | Ensembl 54 (2009) | Ensembl 75 (2014) | Ensembl 82 (2015) |
|---|---|---|---|
| ENSG00000174407 | Protein coding§<br>Cat, dolphin † | Protein coding<br>gibbon | Antisense gene<br>- |
| ENSG00000204091 | Protein coding<br>tarsier, mouse lemur, megabat, dolphin, alpaca | Antisense gene<br>- | Antisense gene<br>- |
| ENSG00000212736 | Protein coding<br>- | Do not exist | Do not exist |
| ENSG00000167747 | Protein coding<br>Megabat, dolphin | Protein coding<br>Gibbon, Megabat , dolphin | Protein coding<br>Gibbon, Megabat, Vervet-agm, dolphin |
| ENSG00000214112 | Protein coding<br>hyrax, megabat, tarsier, dolphin, elephant, cat | Do not exist | Do not exist |
| ENSG00000221891 | Protein coding<br>- | Processed Pseudogene | Processed Pseudogene |

§ ***Gene type.***

† ***Species other than Hominidae in which significant conservation and coding potential identified***

**b) Properties of Hominidae-specific genes identified in current study over time using Ensembl genomic database.**

| Gene | Ensembl 54 (2009) | Ensembl 75 (2014) | Ensembl 82 (2015) |
|---|---|---|---|
| ***ENSG00000184029*** | ***Protein coding***§<br>**-**† | ***Protein coding***<br>**-** | ***Protein coding***<br>**-** |

§ ***Gene type.***

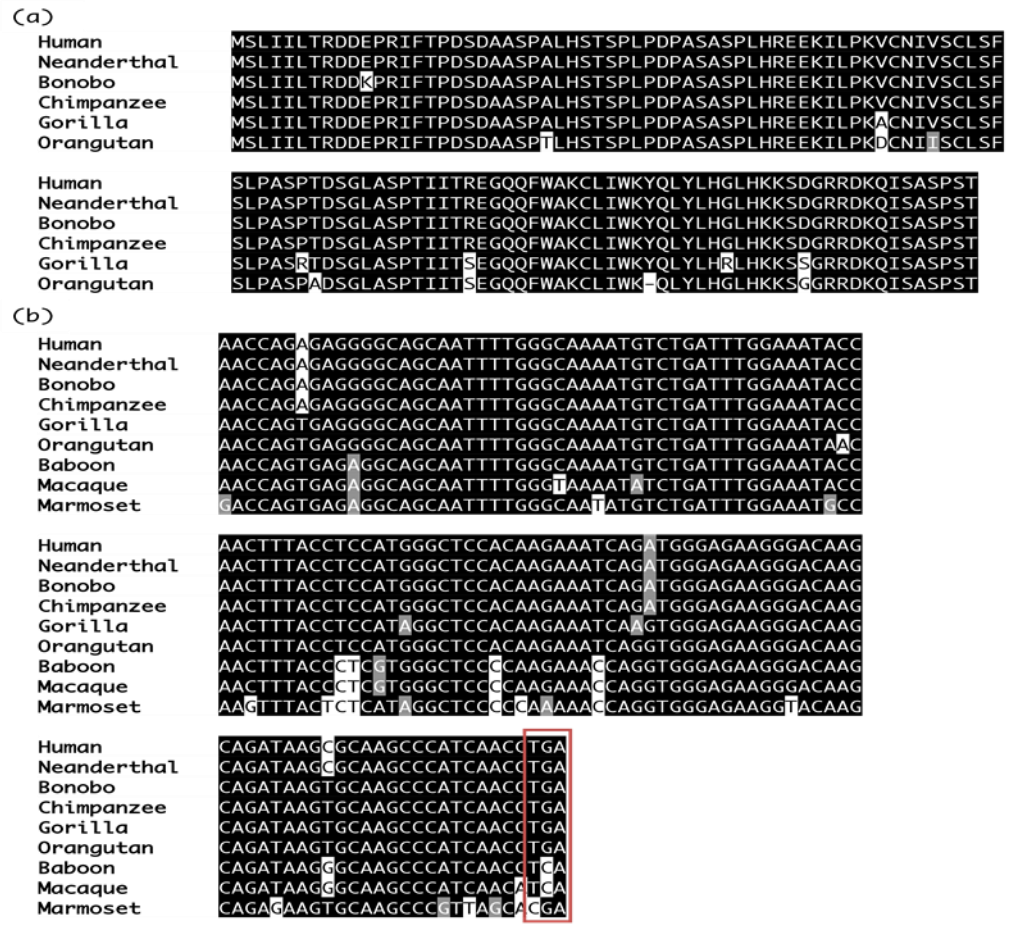
† ***Species other than Hominidae in which significant conservation and coding potential identified***

Appendix A3. Non-Hominidae species considered for identification of Hominidae-specific genes

| *Species* | | *Build* | *Species* | | *Build* |
|---|---|---|---|---|---|
| Alpaca | Vicugna pacos | vicPac1 | **Marmoset** | Callithrix jacchus | C_jacchus3.2.1 |
| Anole Lizard | Anolis carolinensis | AnoCar2.0 | **Medaka** | Oryzias latipes | MEDAKA1 |
| Armadillo | Dasypus novemcinctus | Dasnov3.0 | **Megabat** | Pteropus vampyrus | pteVam1 |
| Atlantic Cod | Gadus morhua | gadMor1 | **Microbat** | Myotis lucifugus | Myoluc2.0 |
| Bushbaby | Otolemur garnettii | OtoGar3 | **Mouse** | Mus musculus | GRCm38 |
| Caenorhabditis elegans | Caenorhabditis elegans | WBcel235 | **Mouse Lemur** | Microcebus murinus | micMur1 |
| Cat | Felis catus | Felis_Catus_6.2 | **Nile tilapia** | Oreochromis niloticus | Oreni1.0 |
| **Cave fish** | Astyanax mexicanus | AstMex102 | **Opossum** | Monodelphis domestica | BROADO5 |
| **Chicken** | Gallus gallus | Galgal4 | **Panda** | Ailuropoda melanoleuca | ailMel1 |
| Chinese softshell turtle | Pelodiscus sinensis | PelSin_1.0 | **Pig** | Sus scrofa | Sscrofa10.2 |
| Ciona intestinalis | Ciona intestinalis | KH | **Pika** | Ochotona princeps | pika |
| Ciona savignyi | Ciona savignyi | CSAV2.0 | **Platyfish** | Xiphophorus maculatus | Xipmac4.4.2 |
| Coelacanth | Latimeria chalumnae | LatCha1 | **Platypus** | Ornithorhynchus anatinus | OANA5 |
| Common Shrew | Sorex araneus | COMMON_SHREW1 | **Rabbit** | Oryctolagus cuniculus | OryCun2.0 |
| Cow | Bos taurus | UMD3.1 | **Rat** | Rattus norvegicus | Rnor_5.0 |
| Dog | Canis lupus familiaris | CanFam3.1 | **Rock Hyrax** | Procavia capensis | proCap1 |
| Dolphin | Tursiops truncatus | turTur1 | **Sheep** | Ovis aries | Oar_v3.1 |
| Duck | Anas platyrhynchos | BGI_Duck_1.0 | **Sloth** | Choloepus hoffmanni | choHof1 |
| Drosophila | Drosophila melanogaster | BDGP5 | **Spotted gar** | Lepisosteus oculatus | LepOcu1 |
| Elephant | Loxodonta africana | loxAfr3 | **Squirrel** | Ictidomys tridecemlineatus | Spetri2 |
| Ferret | Mustela putorius furo | MusPutFur1.0 | **Stickleback** | Gasterosteus aculeatus | BROADS1 |
| Flycatcher | Ficedula albicollis | FicAlb1.4 | **Tarsier** | Tarsius syrichta | tarSyr1 |
| Fugu | Takifugu rubripes | FUGU4 | **Tasmanian Devil** | Sarcophilus harrisii | DEVIL7.0 |
| Gibbon | Nomascus leucogenys | NLeu1.0 | **Tetraodon** | Tetraodon nigroviridis | TETRAODON8 |
| Guinea Pig | Cavia porcellus | CavPor3 | **Tree Shrew** | Tupaia belangeri | TREESHREW |
| Hedgehog | Erinaceus europaeus | HEDGEHOG | **Turkey** | Meleagris gallopavo | UMD2 |
| Horse | Equus caballus | EquCab2 | **Wallaby** | Macropus eugenii | Meug_1.0 |
| Kangaroo Rat | Dipodomys ordii | dipOrd1 | **Xenopus** | Xenopus tropicalis | JGI_4.2 |
| Lamprey | Petromyzon marinus | Pmarinus_7.0 | **Yeast** | Saccharomyces cerevisiae | R64_1-1 |
| Lesser hedgehog tenrec | Echinops telfairi | TENREC | **Zebra Finch** | Taeniopygia guttata | taeGut_3.2.4 |
| Macaque | Macaca mulatta | MMUL_1 | **Zebrafish** | Danio rerio | Zv9 |

Appendix A4. *DSCR4 sequence alignment.*

(a) DSCR4-coded protein's multiple sequence alignment. (b) Multiple sequence alignment of homologous DNA sequences to human DSCR4 exon 3 protein coding sequence. Common disabler is marked in red rectangle.



Appendix A5. Prediction of structures in DSCR4 coded protein.

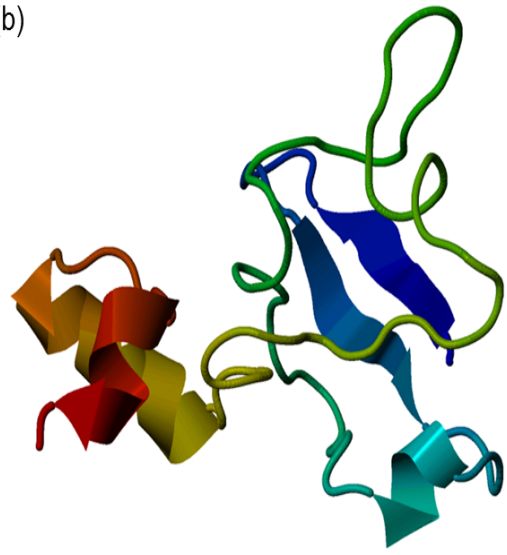
(a) Secondary structure analysis of DSCR4 protein using chou and Fasman secondary structure prediction algorithm. (b) DSCR4 secondary structure prediction using 2kjda protein as template. (c) DSCR4 secondary structure prediction using I-TASSER

(a)

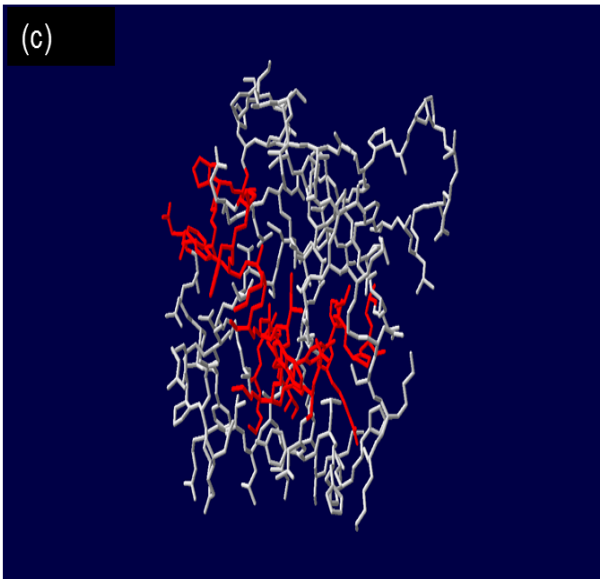

**Secondary Structure:**

```
                  *         *         *         *         *
Query 1    MSLIILTRDDEPRIFTPDSDAASPALHSTSPLPDPASASPLHREEKILPKVCNIVSCLSF 60
Helix 1    HHHHHH               HHHHHH             HHHHHHHHHHHHHHHHHHHH  60
Sheet 1    EEEEEE                                        EEEEEEEEEEEEE  60
Turns 1           TT  T    TTT    T     T   T          TT                60

                  *         *         *         *         *
Query 61   SLPASPTDSGLASPTIITREGQQFWAKCLIWKYQLYLHGLHKKSDGRRDKQISASPST 118
Helix 61                   HHHHHHHHHHHHHHHHHHHHHHHHHHH    HH          118
Sheet 61                EEEEEEEEEEEEEEEEEEEEEE                        118
Turns 61          T   TT    TT T                    T      T     T    118


Total Residues: H: 62    E: 42    T: 20
      Percent: H: 52.5  E: 35.6  T: 16.9
```

(b)



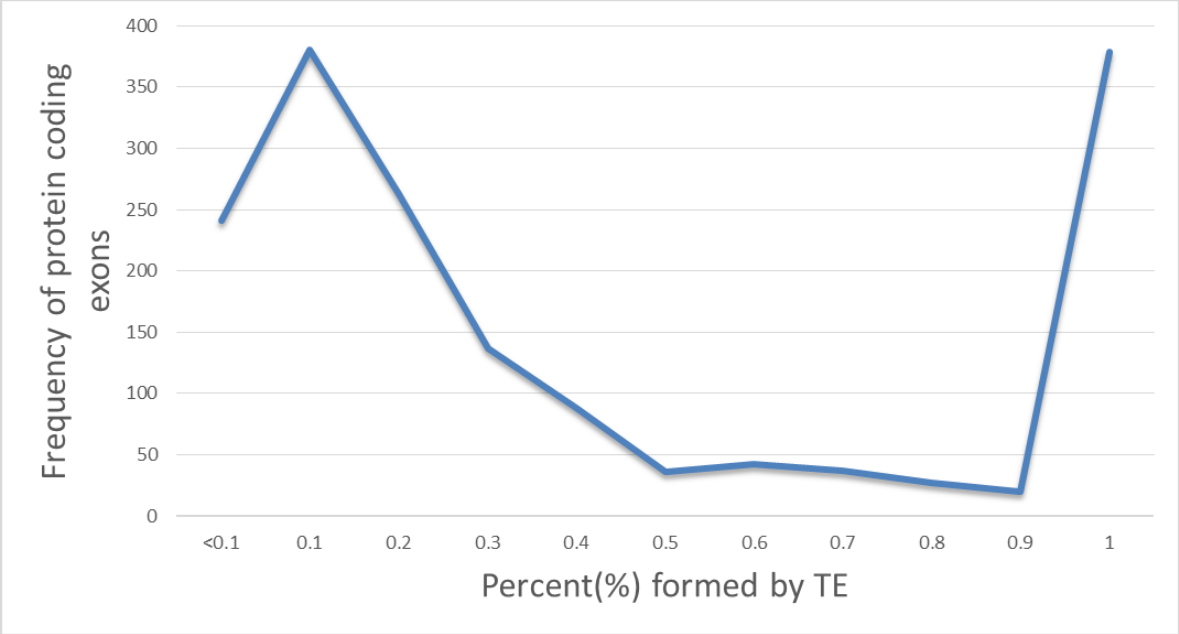
(c)



*Appendix A6. Contribution of transposable elements in DSCR4 exons formation*

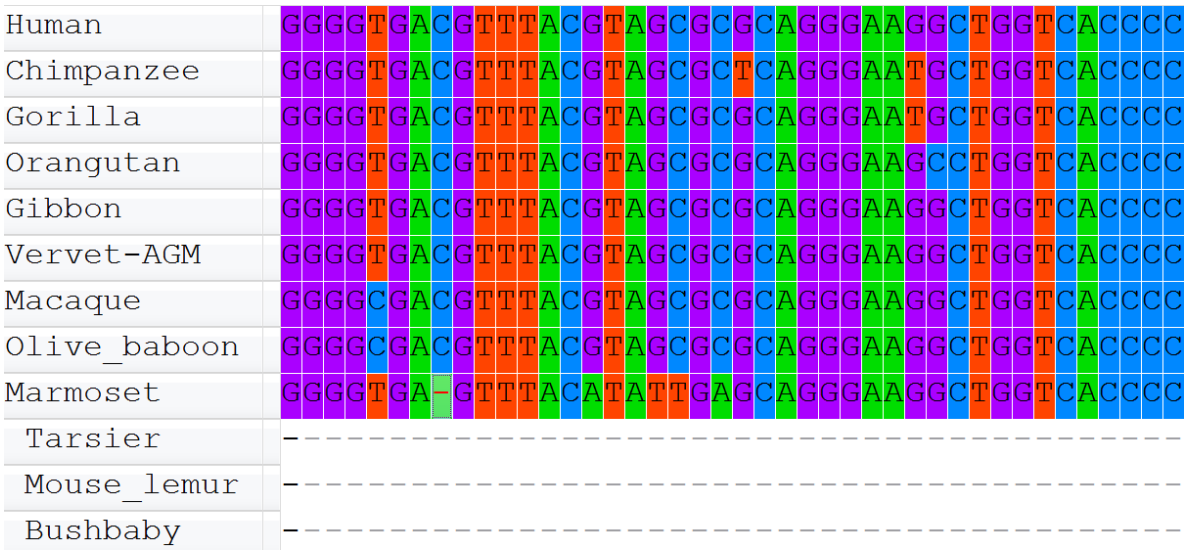
| | Gene coordinate | | | TE match in human genome | | | TE properties | |
|---|---|---|---|---|---|---|---|---|
| **Genomic region** | **Chromosome** | **Start** | **Length** | **SW SCORE** | **Start** | **End** | **Class** | **Family** |
| DSCR4/ DSCR8 promoter | 21 | 39493455 | 92 | 2102 | 39493281 | 39494097 | LTR9 | LTR/ERV1 |
| DSCR4 EXON1 | 21 | 39493222 | 233 | 2102 | 39493281 | 39494097 | LTR9 | LTR/ERV1 |
| DSCR4 EXON1 | 21 | 39493222 | 233 | 940 | 39493141 | 39493280 | LTR16A | LTR/ERVL |
| DSCR4 EXON2 | 21 | 39492401 | 102 | 1849 | 39492309 | 39492672 | MLT2C1 | LTR/ERVL |
| DSCR4 EXON3 | 21 | 39426313 | 763 | 460 | 39426799 | 39426943 | LTR79 | LTR/ERVL |

*Appendix A7. Contribution of transposable elements to human protein coding genes' exon* formation.



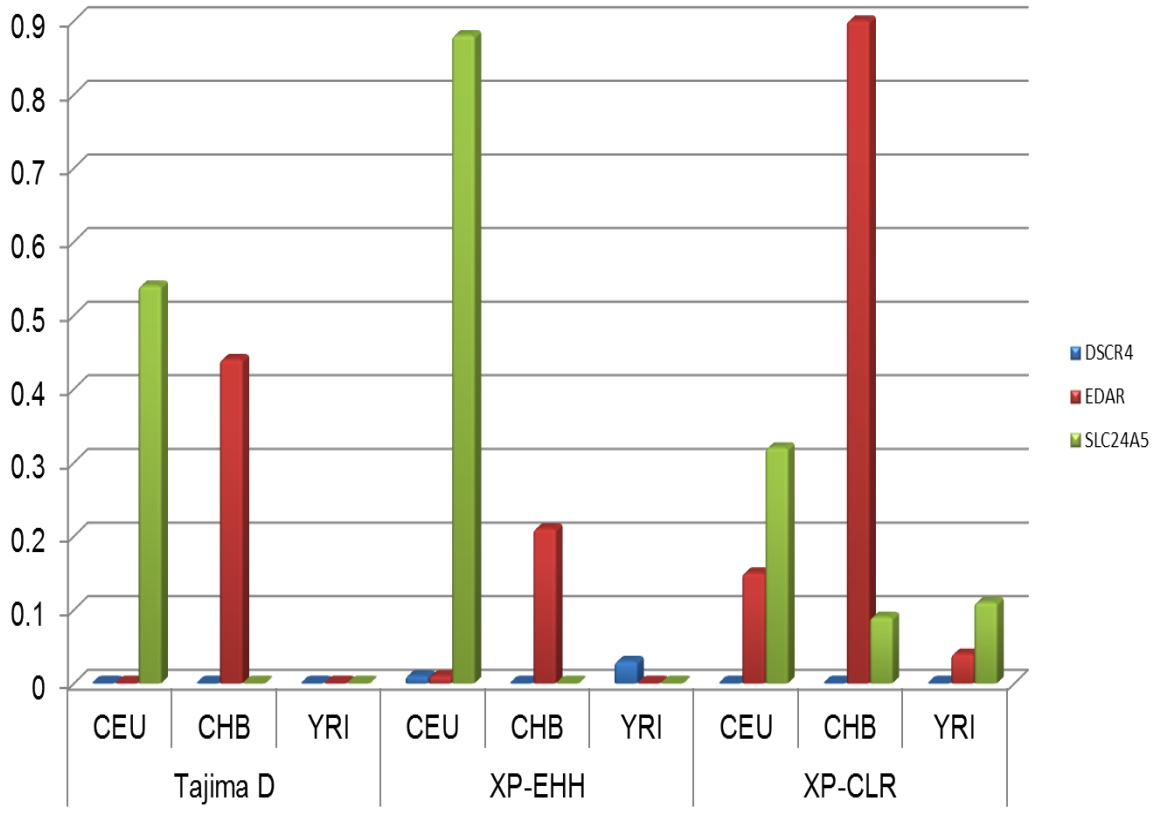
*Appendix A8. Multiple alignment of DSCR4 core promoter.*

Analysis of the core promoteric region of DSCR4/8 bidirectional promoter reveals that DSCR4 promoteric region has retrotransposed 43-73 million years ago in common ancestor of simians.



*Appendix A9a. DSCR4 Analysis of selection based on population genomic variation data.*

The percentage of windows under positive selection based on Tajima D and XP-CLR as well as percentage of SNPs under positive selection based on XP-EHH in European, Asian and African populations for HS gene, DSCR4 along with SLC24A5 and EDAR genes which have been shown to be under positive selection respectively in European and Asian populations, are shown as bar chart. (CEU: Utah residents with Northern and Western European ancestry from the CEPH collection, CHB: Han Chinese in Beijing, China, YRI: Yoruba in Ibadan, Nigeria). DSCR4 doesn't show signs of positive selection in any of the investigated populations.



*Appendix A9b. DSCR4 Analysis of selection of DSCR4 gene based on the ratio of non-synonymous to synonymous site changes in humans and great apes.*

| | Sd | Sn | S | N | ds | dn | dn/ds |
|---|---|---|---|---|---|---|---|
| Human-Chimpanzee | 2 | 0 | 83.33 | 273.66 | 0.0244 | 0 | 0 |
| Human-Gorilla | 1 | 6 | 83 | 274 | 0.0121 | 0.0222 | 1.834711 |
| Human-Orangutan | 2 | 6 | 82.5 | 271.5 | 0.0246 | 0.0224 | 0.910569 |

Sd. Number of synonymous changes

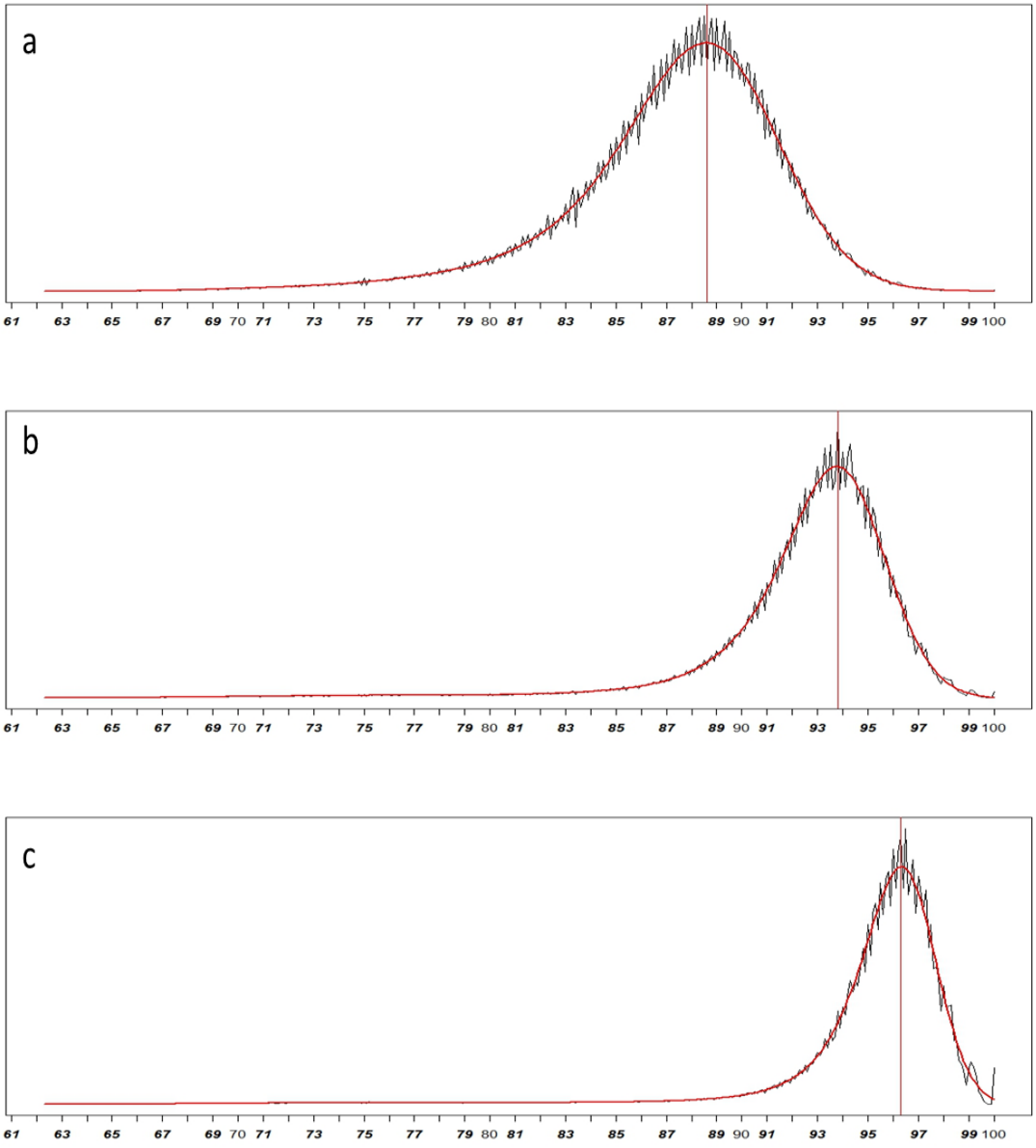
Sn. Number of non-synonymous changes

S. Number of synonymous sites

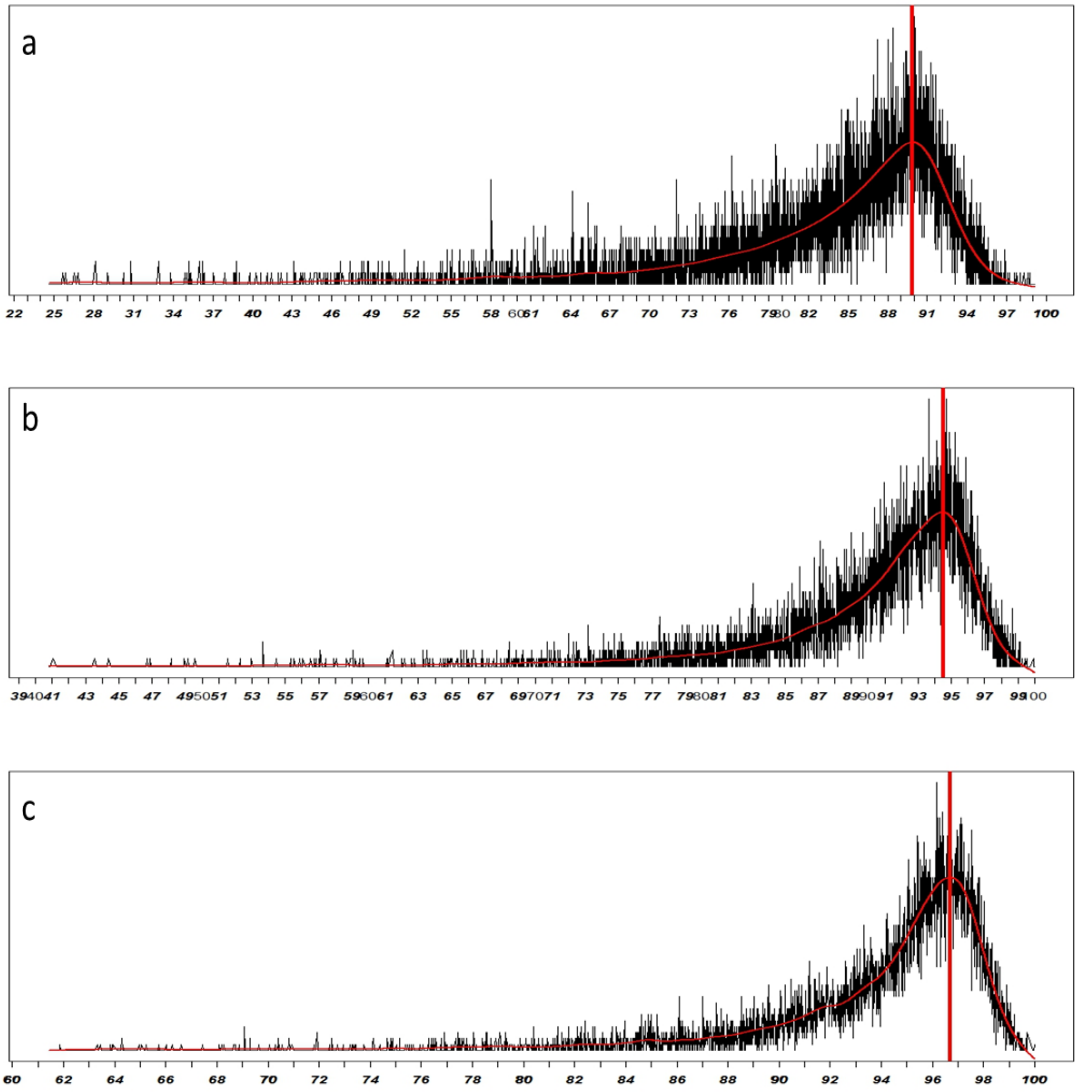
N. Number of non-synonymous sites

*Appendix A10a. Neutrally evolving sequences' conservation level in Non-coding sequences.*

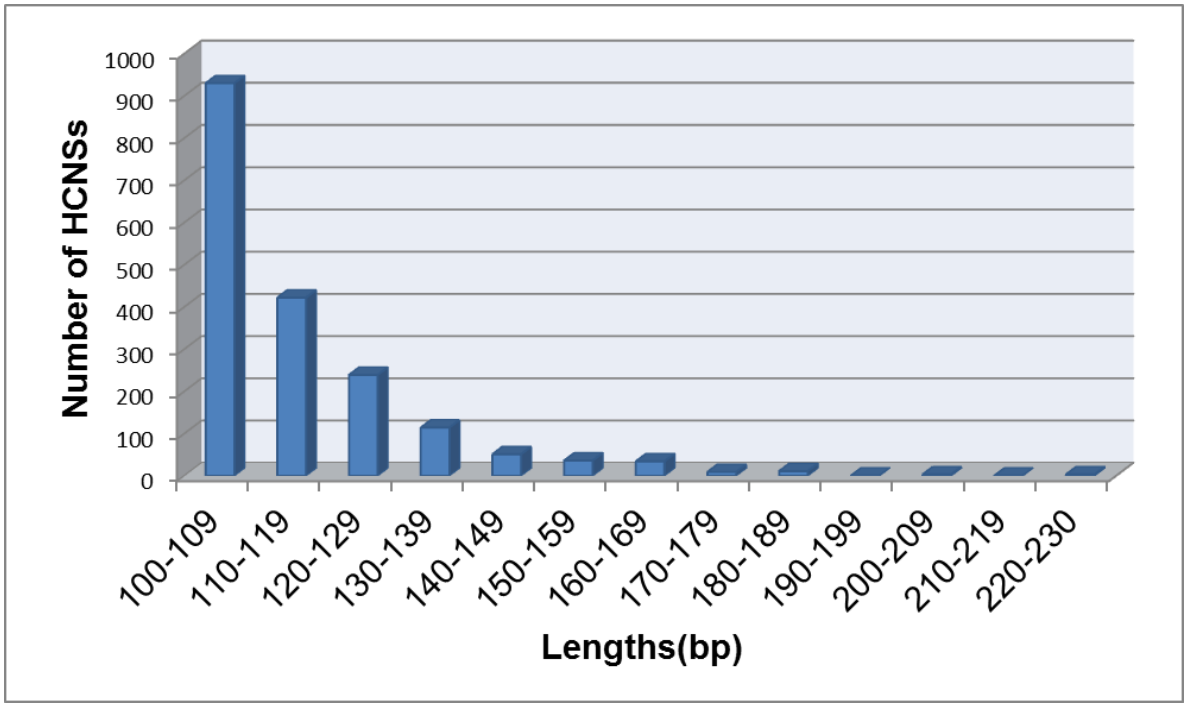
Human-marmoset, human-macaque and human-gibbon homologous sequences' conservation plot based on non-coding DNA conservation level (a, b and c, respectively).



*Appendix A10b. Neutrally evolving sequences' conservation level in coding sequences.*

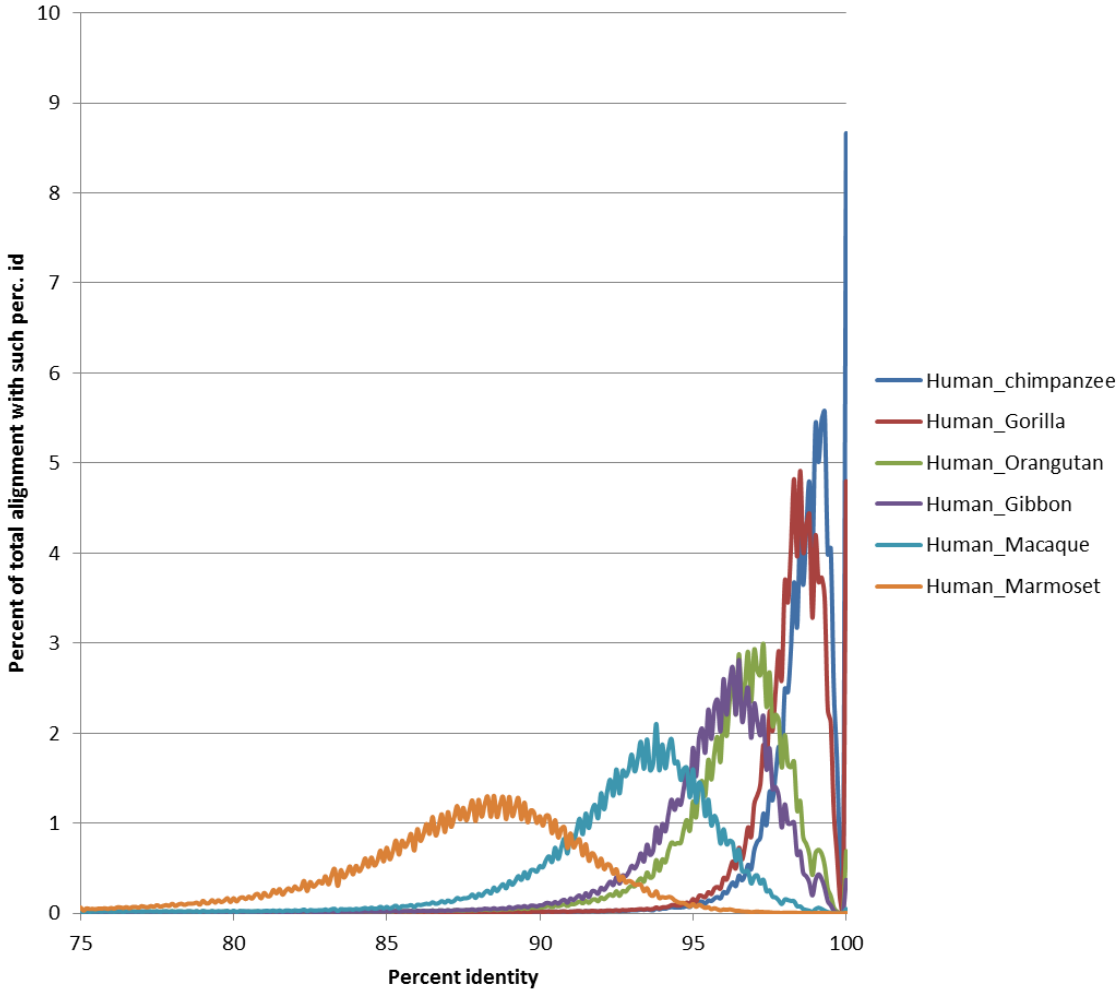


*Appendix A11. Length distribution of HS HCNSs.*



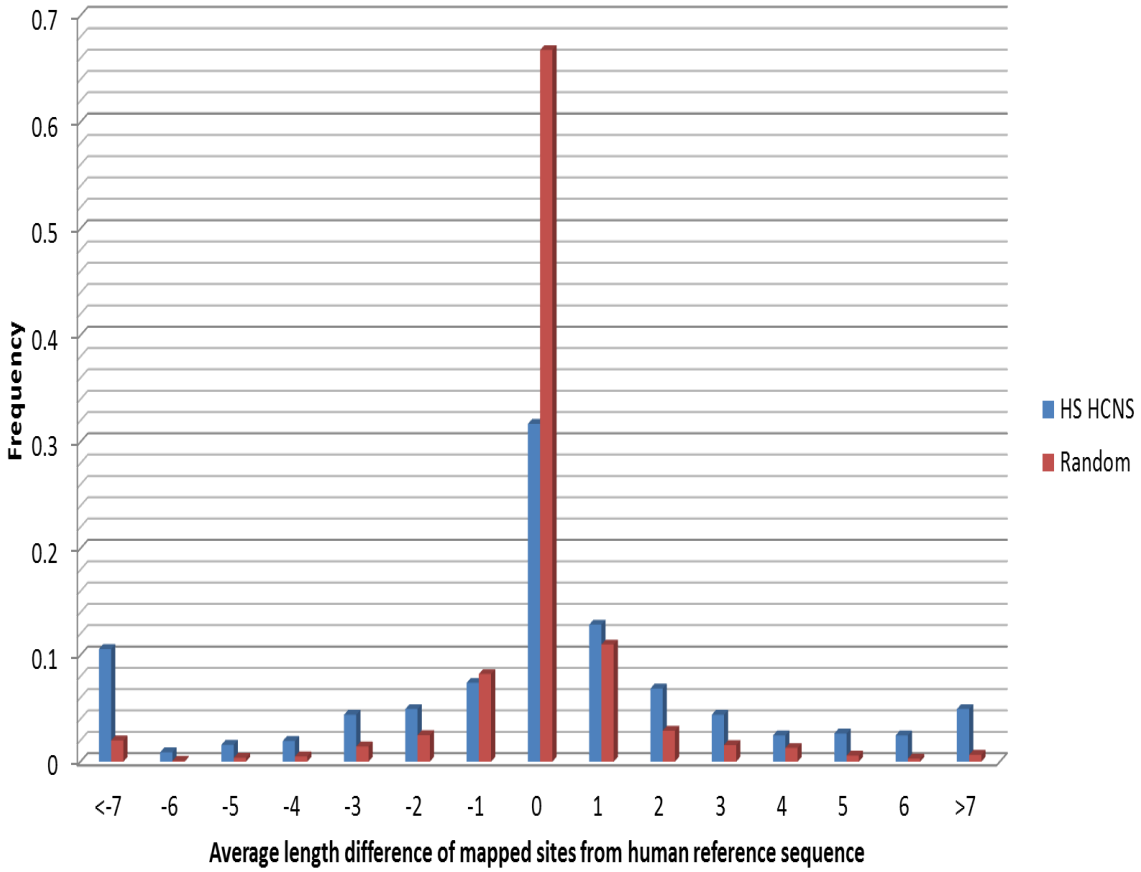
*Appendix A12. Noncoding identity distribution within Catarrhini infra-order.*

Distributions of whole-genome noncoding sequence identities are represented for Catarahini members.



*Appendix A13a. Evolutionary origin of Hominidae-specific HCNSs (HS HCNSs).*

Average length difference of sequences mapped to HS HCNSs in gibbon and rhesus macaque from human reference sequence is significantly higher than random sequences of the same number and size.

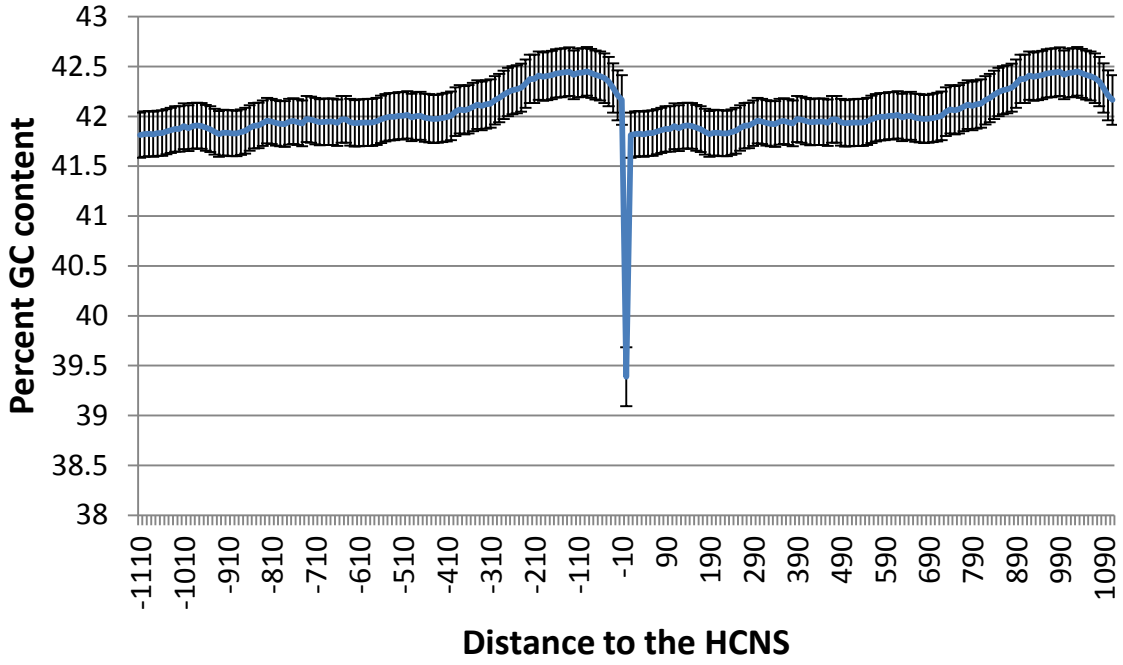


*Appendix A13b. Examples of Hominidae-specific HCNSs under strong accelerated evolution in Hominidae common ancestor*

| HS HCNS properties | | | Genetic distance | Likely target gene | | |
|---|---|---|---|---|---|---|
| Chromosome | Start | Length | α distance | Gene name | Start | End |
| 3 | 43045007 | 107 | 0.19 | KRBOX1 | 42850938 | 43097363 |
| X | 152417953 | 165 | 0.12 | MAGEA1 | 152481522 | 152486115 |
| 11 | 49108620 | 100 | 0.11 | TRIM64C | 49075266 | 49080664 |
| 2 | 131062944 | 105 | 0.1 | CCDC115 | 131095814 | 131099922 |
| X | 119157282 | 110 | 0.1 | RHOXF2B | 119205848 | 119211707 |
| 1 | 13214501 | 116 | 0.1 | PRAMEF26 | 13216356 | 13219581 |
| 16 | 33769118 | 101 | 0.09 | RP11-812E19.9 | 33647044 | 33647696 |
| 16 | 33744583 | 100 | 0.09 | RP11-812E19.9 | 33647044 | 33647696 |
| 16 | 32922979 | 100 | 0.09 | TP53TG3 | 32684852 | 32688053 |
| 3 | 14132416 | 112 | 0.08 | TPRXL | 13978756 | 14124311 |
| 7 | 62537666 | 109 | 0.08 | AC006455.1 | 62809239 | 62812151 |
| 5 | 70988362 | 107 | 0.08 | CARTPT | 71014990 | 71016875 |
| 11 | 104943827 | 124 | 0.08 | CASP1 | 104896170 | 104972158 |
| 14 | 74764127 | 133 | 0.07 | ABCD4 | 74752126 | 74769759 |
| 18 | 76749706 | 105 | 0.07 | SALL3 | 76740275 | 76762677 |
| 7 | 62778744 | 123 | 0.06 | AC006455.1 | 62809239 | 62812151 |
| 10 | 27618391 | 120 | 0.06 | PTCHD3 | 27687116 | 27703297 |
| 2 | 132093021 | 108 | 0.06 | PLEKHB2 | 131862420 | 132111282 |
| 7 | 48121826 | 120 | 0.06 | UPP1 | 48128225 | 48148330 |
| 17 | 26048502 | 109 | 0.06 | NOS2 | 26083792 | 26127525 |
| 8 | 21426860 | 107 | 0.06 | GFRA2 | 21547915 | 21669869 |
| 2 | 132106021 | 135 | 0.06 | PLEKHB2 | 131862420 | 132111282 |
| 6 | 155929553 | 107 | 0.06 | NOX3 | 155716504 | 155777037 |
| 7 | 63149048 | 112 | 0.06 | AC073188.1 | 62858414 | 62858860 |

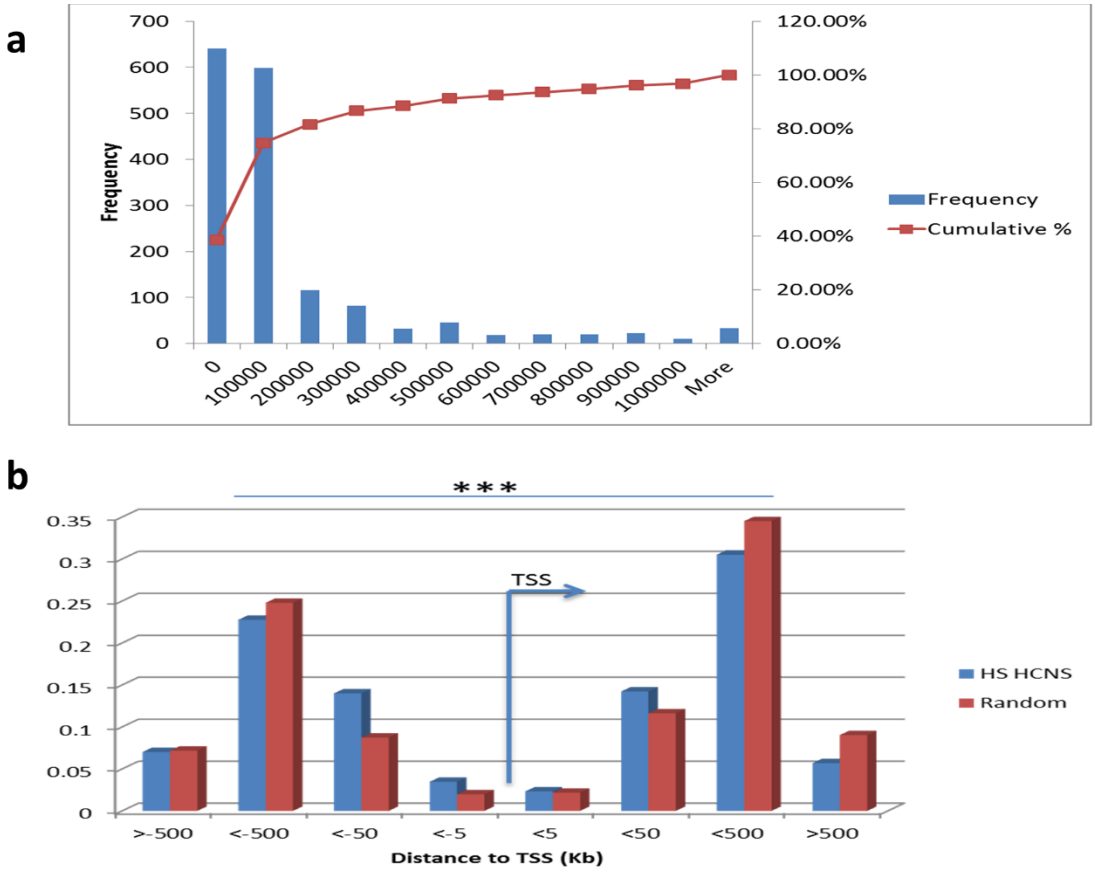
*Appendix A13c. The GC content of the CNS and CNS flanking regions of CNSs.*

Using sliding windows of 200bp size and sliding steps of 10bp, the percent GC contents of the Hominidae-specific HCNS and flanking regions were computed. Position 0 is the 100bp in the center of the CNSs.



*Appendix A14a. Proximity of Hominidae-specific HCNSs to genes and transcription start sites (TSSs)*

(a) The proximity of the HS HCNSs to genes. The horizontal axis represents the distance of the HCNS to the closest protein coding gene. (b) The proximity of the HCNSs to Transcription start site (TSS) compared to random expectations using GREAT online software. Hominidae-specific HCNSs are significantly overrepresented within range of <50 Kb from TSSs, and underrepresented for distance ranges > 50 kb from TSSs (p-value < 2.2e-16, chi-square test).

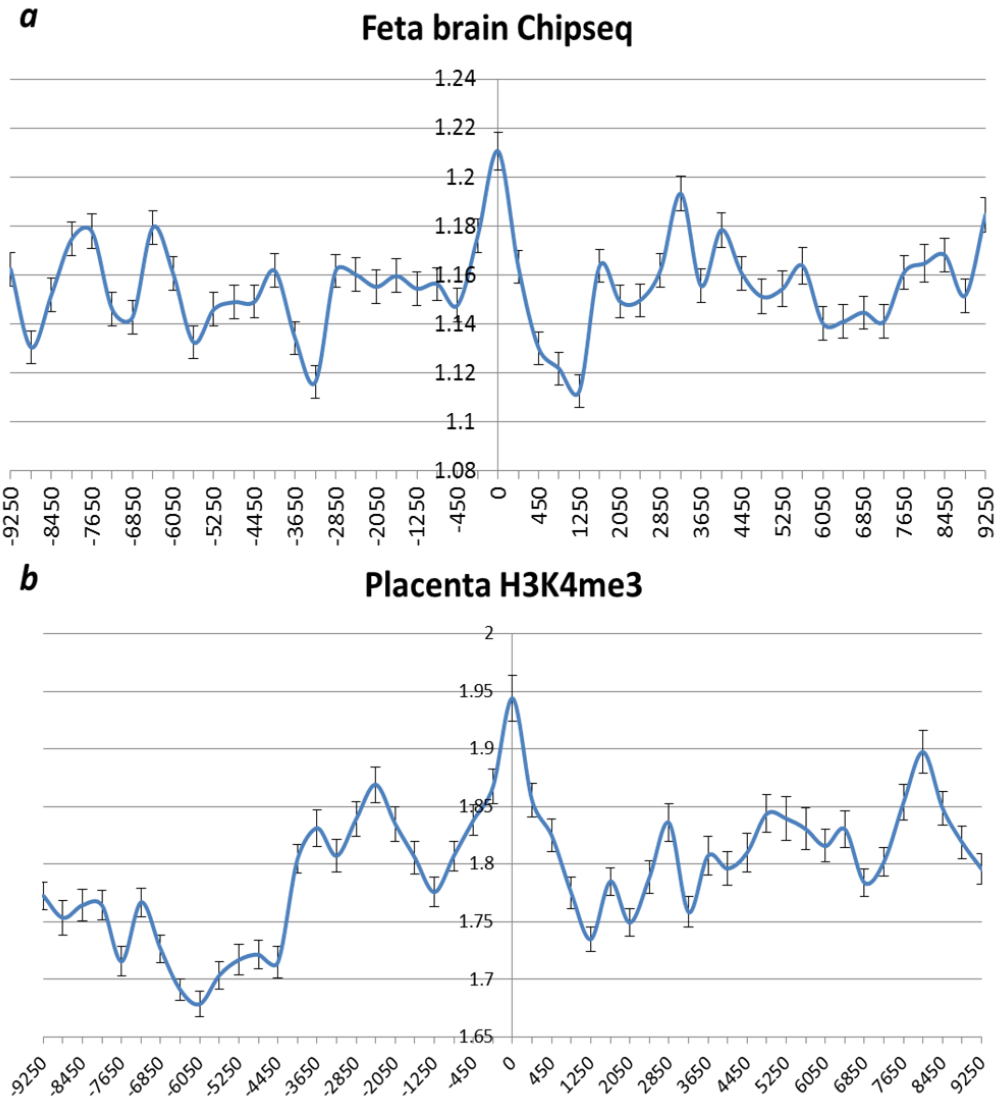


Appendix A14b. Hominidae-specific HCNSs' target genes with expression switch on the branch connecting great apes and macaque.

| Organ | Gene ID | HS HCNS per target gene |
|---|---|---|
| Brain | ENSG00000183098, ENSG00000179163, ENSG00000105675 | 1 |
| Cerebellum | ENSG00000183098, ENSG00000112679, ENSG00000169851, ENSG00000169851, ENSG00000086827, ENSG00000146938, ENSG00000169504, ENSG00000169504, ENSG00000124198, ENSG00000165194, ENSG00000165300, ENSG00000165300, ENSG00000213949, ENSG00000162415, ENSG00000083123, ENSG00000105675, ENSG00000184220, ENSG00000146085, ENSG00000146085, ENSG00000111261, ENSG00000114480, ENSG00000114480, ENSG00000060718, ENSG00000143061, ENSG00000143061, ENSG00000174989, ENSG00000165194, ENSG00000165194, ENSG00000165194, ENSG00000204262, ENSG00000196950, ENSG00000166897, ENSG00000206052, ENSG00000166266, ENSG00000169851, ENSG00000087502, ENSG00000133687, ENSG00000169851, ENSG00000108018 | 2.6 |
| Heart | ENSG00000149418 | 1 |
| Liver | ENSG00000183098, ENSG00000112139, ENSG00000131142, ENSG00000144028, ENSG00000140479, ENSG00000139263, ENSG00000197766, ENSG00000172348, ENSG00000166377, ENSG00000077380, ENSG00000122641, ENSG00000122641, ENSG00000106070, ENSG00000113273, ENSG00000110900, ENSG00000110900, ENSG00000171714 | 1.4 |
| Testis | ENSG00000183098, ENSG00000184226, ENSG00000184226, ENSG00000153822, ENSG00000086827, ENSG00000179163, ENSG00000124198, ENSG00000112996, ENSG00000215009, ENSG00000186094, ENSG00000163528, ENSG00000141622, ENSG00000135541, ENSG00000185008, ENSG00000198597, ENSG00000138316, ENSG00000085382, ENSG00000170417, ENSG00000071242, ENSG00000071242, ENSG00000166377, ENSG00000197603, ENSG00000157330, ENSG00000187391, ENSG00000145439, ENSG00000174429, ENSG00000125999, ENSG00000143222, ENSG00000169946, ENSG00000143507, ENSG00000065325, ENSG00000141568, ENSG00000188517, ENSG00000173200, ENSG00000147202, ENSG00000204262, ENSG00000106070, ENSG00000132842, ENSG00000164326, ENSG00000180537, ENSG00000087502, ENSG00000184178, ENSG00000185008, ENSG00000163995, ENSG00000038532, ENSG00000152936, ENSG00000152936, ENSG00000152936 | 1.4 |

Appendix A14c. Analysis of tissue specificity of Hominidae-specific HCNSs.

(a) Analysis of the Epigenome roadmap chromatin immunoprecipitation data reveals intensified chipseq signal within HS HCNSs compared to flanking regions in fetal brain. (b) Analysis of the Epigenome roadmap data regarding H3K4me3 epigenetic mark showed intensified signal within HS HCNS in placenta. H3K4me3 is associated with active promoter regions.



*Appendix A15a*

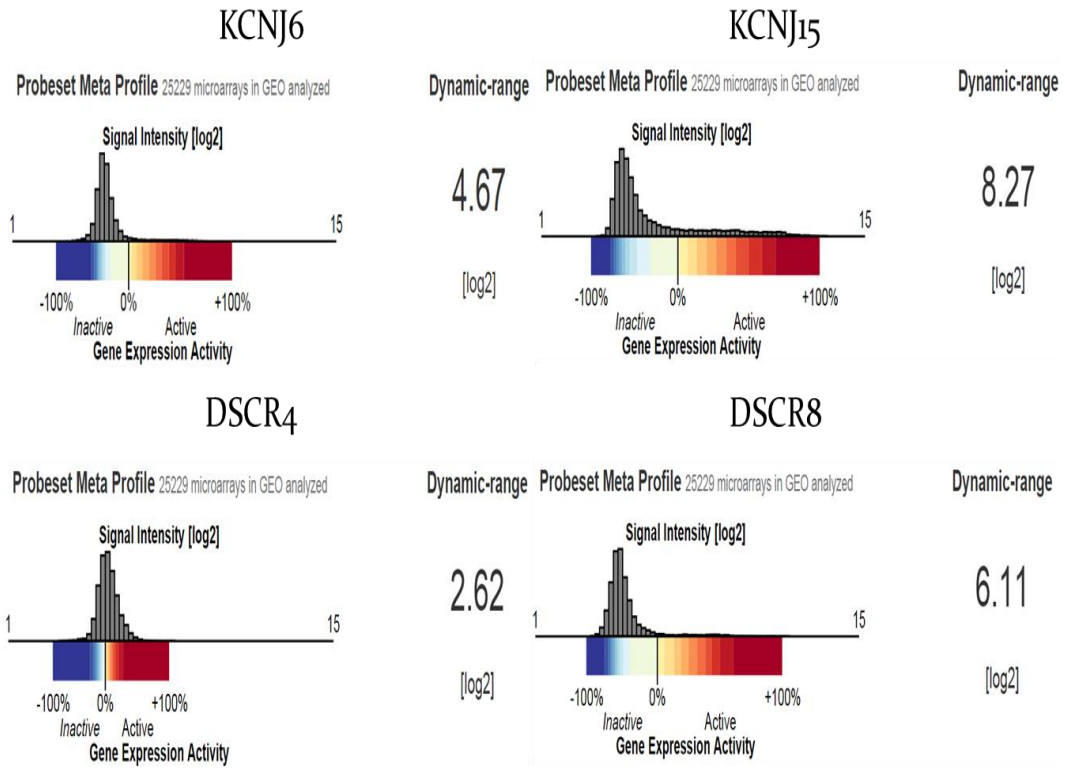
(a) Gene ontology analysis of conserved noncoding sequences under accelerated evolution in Human (HACNs).

| Biological Process | Binom Fold Enrichment |
|---|---|
| regulation of type B pancreatic cell development | 6.67 |
| negative regulation of transcription by competitive promoter binding | 4.99 |
| forebrain ventricular zone progenitor cell division | 4.64 |
| regulation of transcription involved in cell fate commitment | 3.48 |
| positive regulation of filopodium assembly | 3.35 |
| neuron recognition | 3.13 |
| neuron fate specification | 2.84 |
| regulation of filopodium assembly | 2.68 |
| neural crest cell migration | 2.39 |
| forebrain neuron differentiation | 2.32 |
| cell recognition | 2.3 |
| cell fate specification | 2.2 |
| neural crest cell development | 2.19 |
| negative regulation of neuron differentiation | 2.18 |
| forebrain generation of neurons | 2.18 |
| neuron fate commitment | 2.13 |
| homophilic cell adhesion | 2.08 |

(b) Gene ontology analysis of human genome regions under accelerated evolution (HARs).

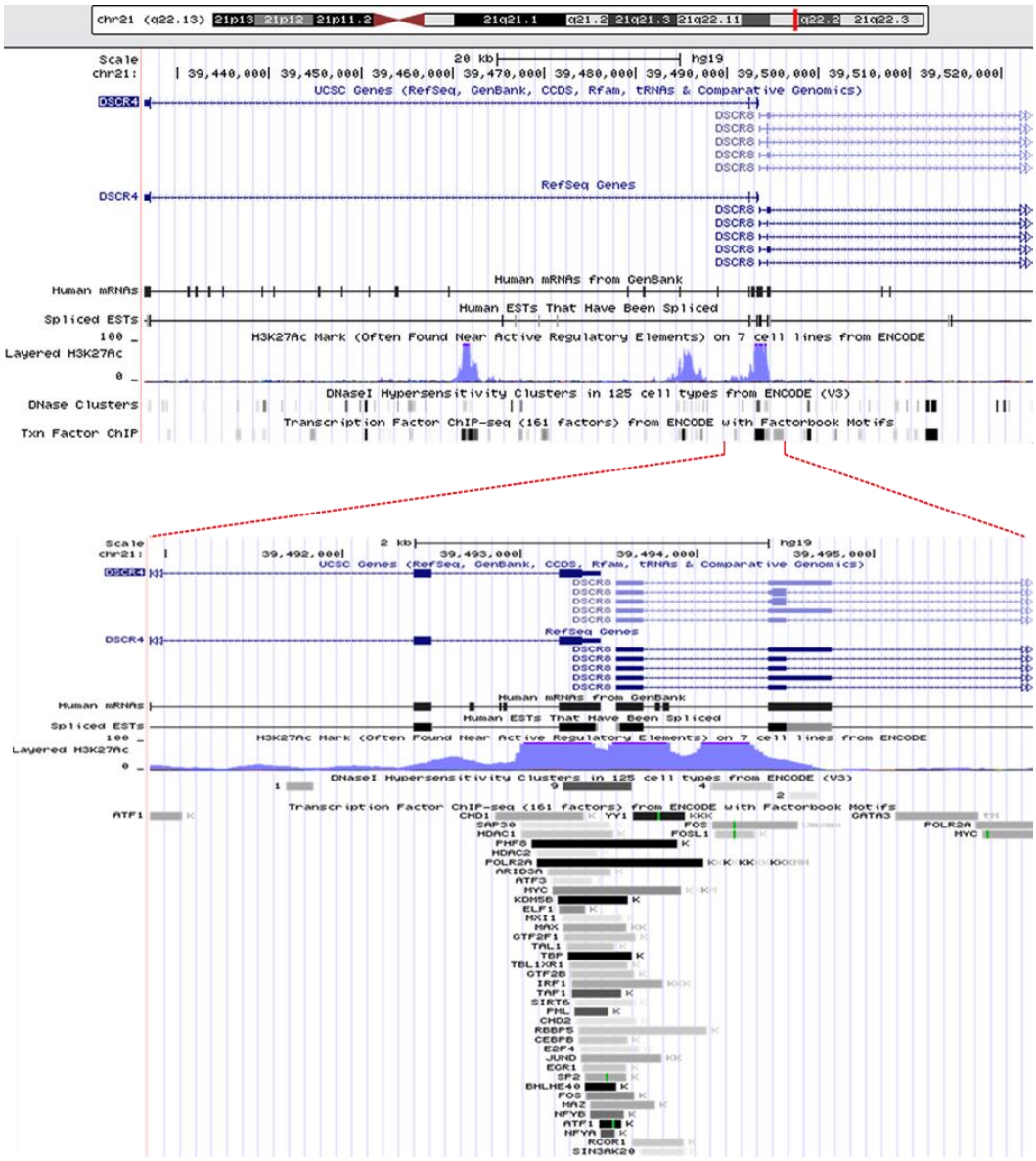
| Biological Process | Binom Fold Enrichment |
|---|---|
| - | - |

*Appendix A15b. Absolute gene expression profiles of DSCR4 and DSCR8 genes along with their flanking genes based on 25229 Affymetrix Human Genome U133 Plus 2.0 Array analysis.*



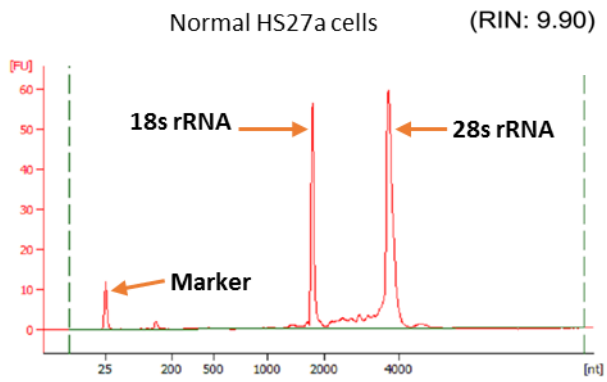
*Appendix A15c. DSCR4 gene regulation.*

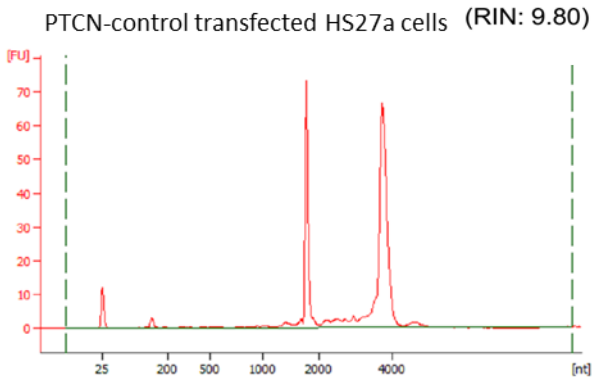
Encode data suggest existence of three active regulatory elements within DSCR4gene. These regions share three characteristics: h3k27 acylation epigenetic mark that is found near active regulatory elements, forming open chromatin region which is a characteristic shared by several classes of transcription factor binding sites and acting as binding site for several transcription factors.

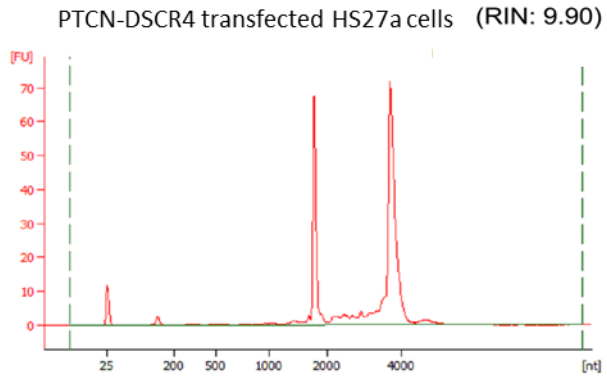


Appendix A16. Quality check analysis of extracted RNA samples from PTCN-DSR4 transfected, PTCN-control transfected and normal HS27a cells using Agilent 2100 bioanalyzer.

Three peaks indicating a marker, 18s ribosomal RNA (rRNA) and 28s rRNA, clearly visible in all three samples indicate the high quality of extracted total RNAs. The RNA Integrity Number (RIN) (calculated out of 10), also confirms that there are no sign of degradation in extracted RNA samples.

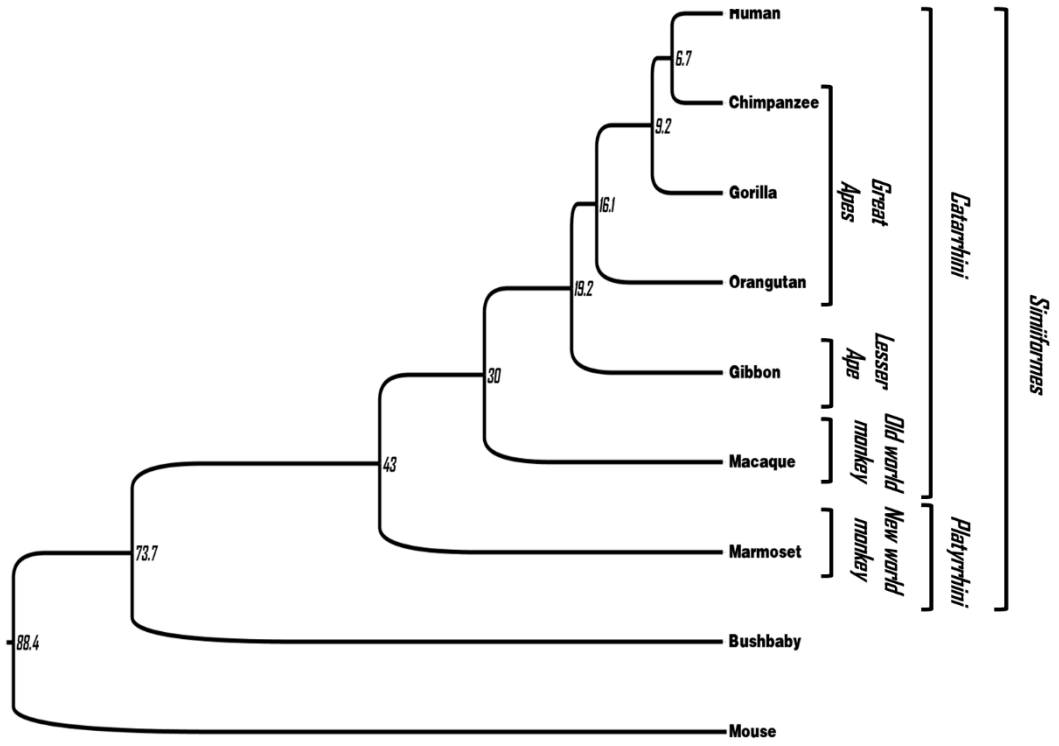


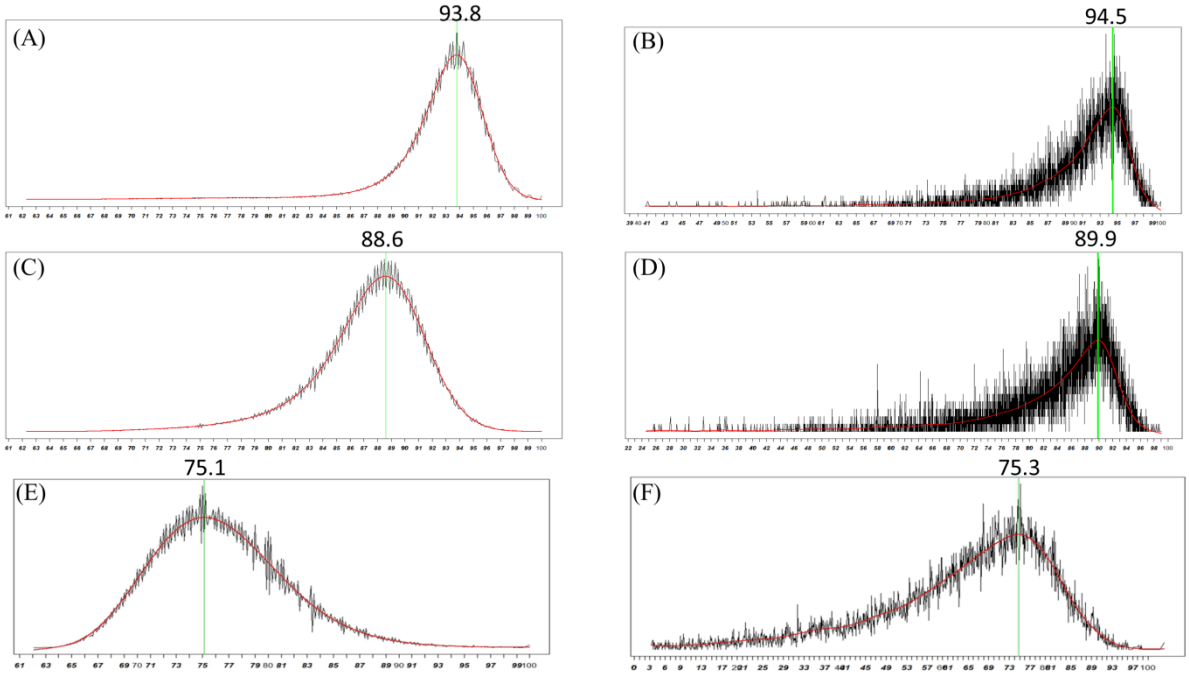
*Appendix A17a. Phylogenetic tree and divergence times of lineages in primate order.*

The evolutionary relationships of Hominoidea, including humans, great apes and lesser apes with other primate members and their divergence times are shown. Rhesus macaque representing family Cercopithecidae (Old world monkeys), marmoset representing parvorder Platyrrhini (New world monkeys) and bushbaby representing family Galagidae are the three species with closest evolutionary relationships with Hominoidea and have been used as outgroups in this study.



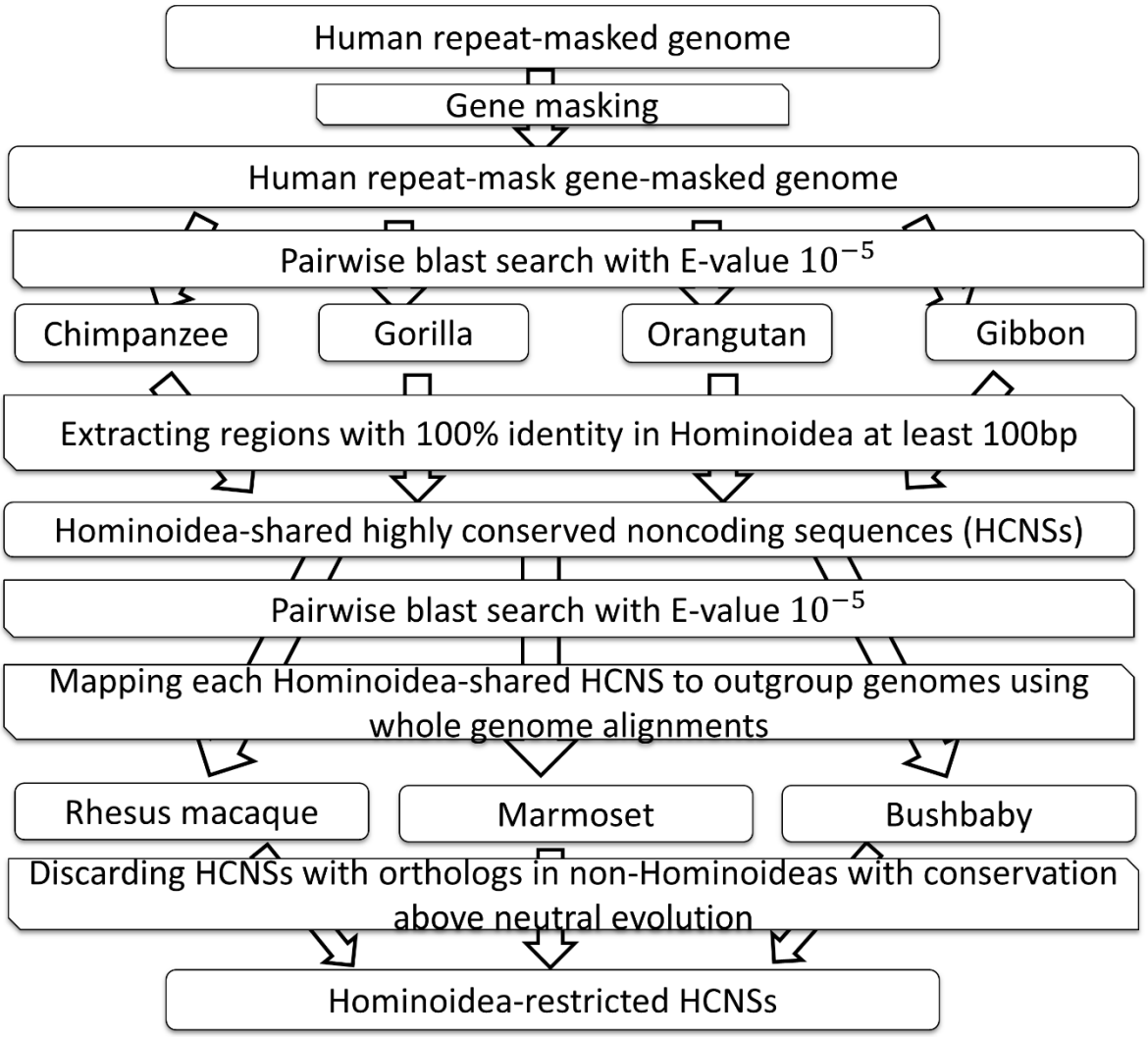
Appendix A17b. Neutral evolution rates in protein coding and non-coding sequences.

Human-rhesus macaque, human-marmoset and human-bushbaby protein coding sequences' synonymous sites substitution rate were calculated for genes with one-to-one orthology (B, D and F respectively). Human-rhesus macaque, human-marmoset and human-bushbaby nucleotide substitution rate in non-coding non-repetitive sequences were calculated using whole-genome DNA sequence alignments (A, C and E, respectively). The mode of the plots represent neutral evolution threshold. The neutral evolution rates are similar in coding and noncoding sequences, with slight abundance of conservation in protein coding synonymous sites, expected due to the action of negative selection on some of the coding synonymous sites.



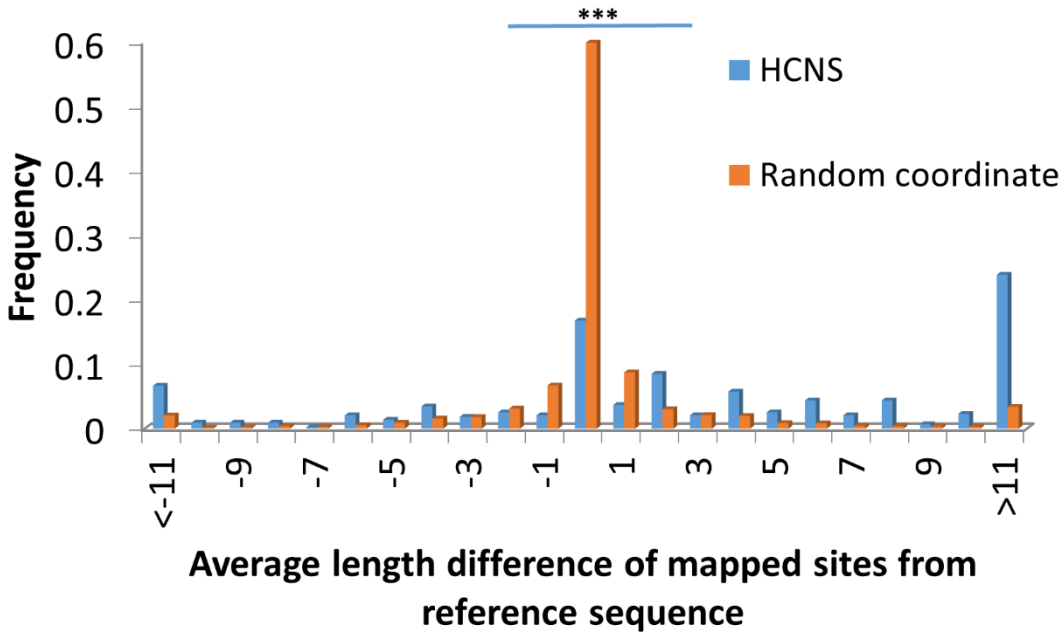
*Appendix A17c. Hominoidea-restricted HCNS identification pipeline.*

Repeat-masked and CDS-masked human genome were used as reference and whole gnome pairwise BlastN searched were conducted between Hominoidea. The non-coding sequences at least 100bp long with absolute conservation across all Hominoidea were defined as Hominoidea-shared HCNSs. In the next step, each of Hominoidea-shared HCNSs were searched using BlastN and also mapped to outgroup genomes using whole genome alignment data. Discarding Hominoidea-shared HCNSs with orthologous sequences in outgroup species with conservation above neutral evolution threshold, 679 Hominoidea-restricted HCNSs were identified.



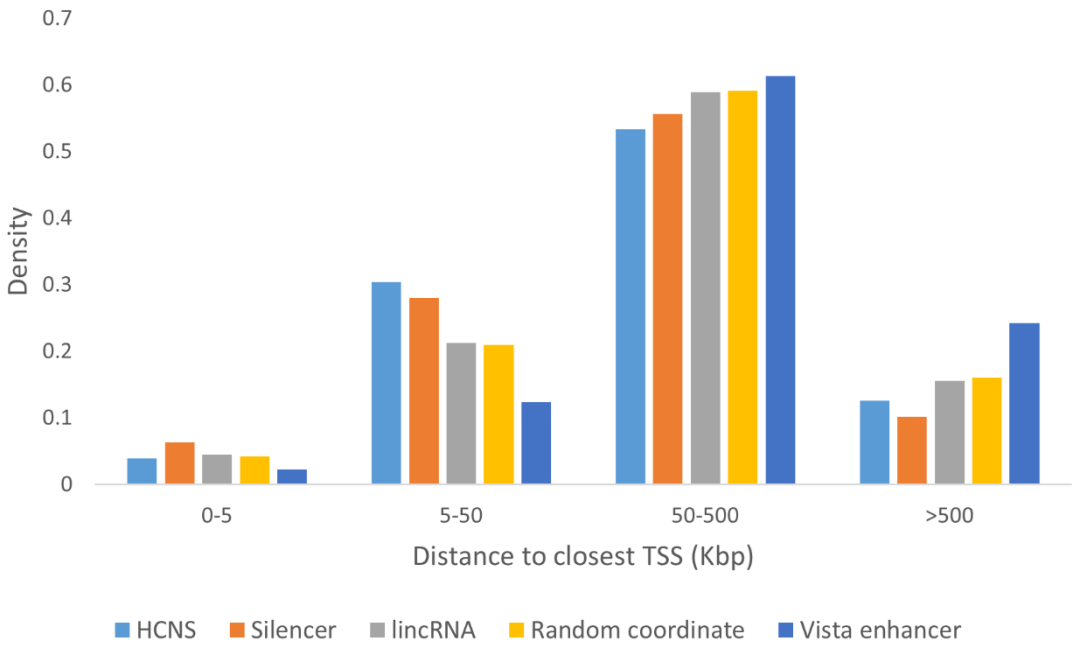
*Appendix A18a. Rate of insertion and deletion at HCNS ancestral sequences.*

Eighty three percent Hominoidea HCNSs' ancestral sequences have experienced insertion or deletion, however, only forty percent of sequences under neutral evolution show such characteristic. HCNSs have significantly higher rate of insertions and deletions longer than 10 nucleotides compare to neutrally evolving random sequences. (Chi square P value: <0.0001)



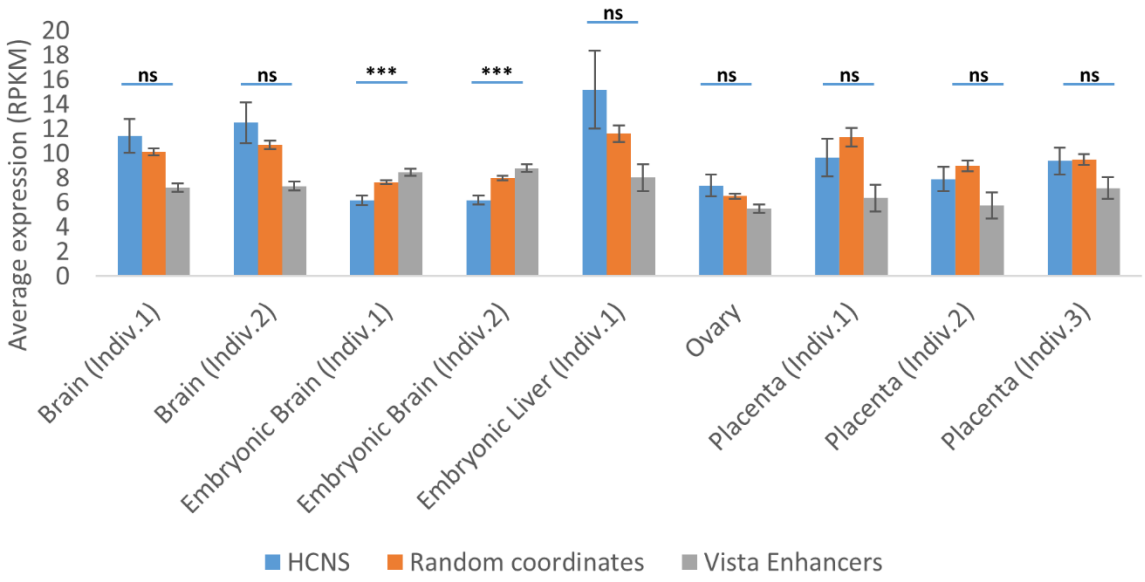
*Appendix A18b. Genomic distribution enrichment analysis using GREAT tool.*

HCNSs are enriched in close proximity of protein coding gene's transcription start sites. The enrichment is strongest at distances between 5-50 kb at upstream and downstream flanking regions compared to vista enhancers and random coordinates and similar to intergenic silencers. HCNSs and intergenic silencer elements are underrepresented at distances farther than 50 kb from their target protein coding genes' transcription start sites. Chi square P values are <0.0001 for pairwise comparison of HCNSs with random coordinates, vista enhancers and lincRNAs.



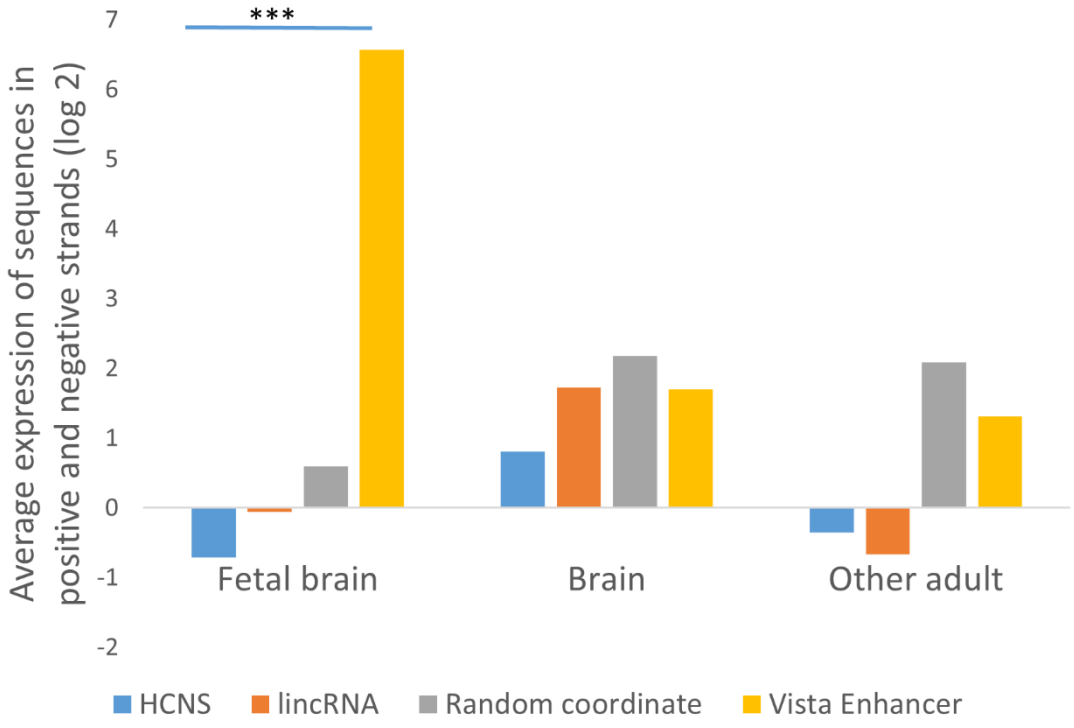
*Appendix A18c. Enrichment of HCNS-target genes' expression across human tissues.*

Analysis of the average gene expression of HCNS target genes across human tissues consistently reveals lower expression of HCNS target genes compare to target genes of random coordinates and vista enhancer elements in embryonic brain. No significant difference were observed across other tissues. ns (non-significant); ***P value < 0.001 (Mann–Whitney U test).



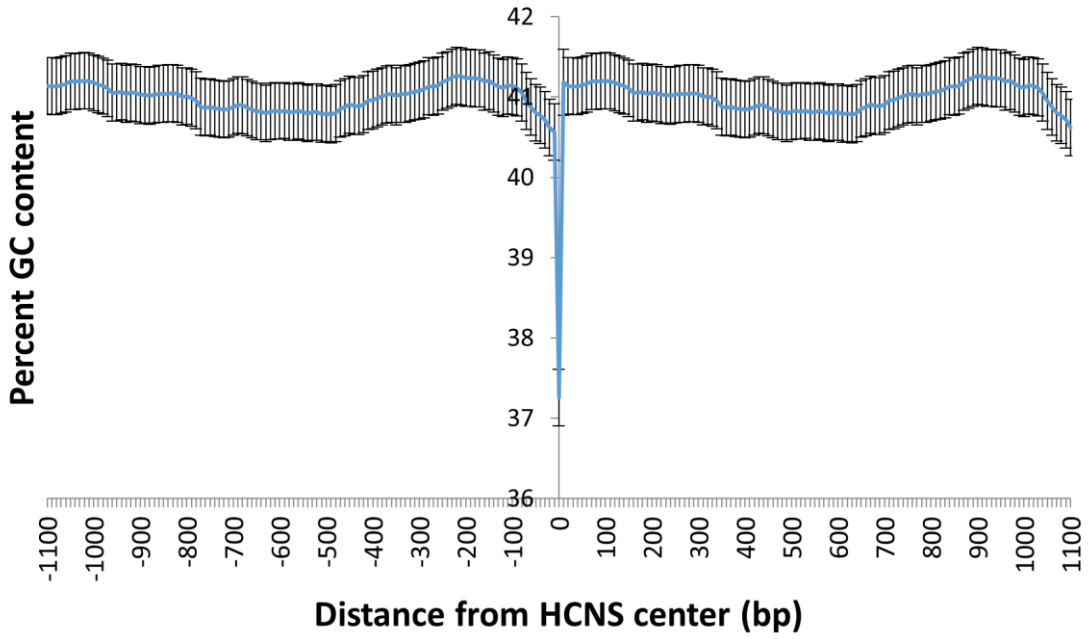
*Appendix A19a. Enrichment of HCNS eRNA expression in human tissues.*

Enhancer eRNA expression analysis across human tissues reveals that HCNSs have significantly lower eRNA expression levels compare to vista enhancers and random sequences in fetal brain. ***P value < 0.001 (Mann–Whitney U test).



*Appendix A19b. Enrichment of HCNS eRNA expression in human tissues.*

Using sliding windows of 100bp size and sliding steps of 10bp, the percent GC contents of the Hominoidea-restricted HCNSs and flanking regions were computed. Position 0 is the 100bp in the center of the CNSs. HCNSs have significantly lower GC levels compare to their upstream and downstream flanking regions and also compare to whole genome average.



Appendix A20. DNA sequence of top HCNSs under strong accelerated evolution in Hominidae common ancestor. (The genomic locations are presented according to GRCH37 In the format of: Chromosome|genomic start coordinate:genomic end coordinate)

>Human|X:119157282:119157391

TTACTCCAGCCCACATGGCAGCTGCCCAGGCAGCCCAGAAGGAGGCCTAAGAAAGCCCTAGAAGCAAA
GCCATAAGAATAAGTCAGTTTTCCTAGAGGTCAAACAAGGCT

>Human|X:152417953:152418117

GCAGGGATTGTGTTCTTTGGACCATCCTAATAGTCACCTCCTTGGACCTTCTTGCCTAGCAGTCTCCAGC
ATAGTAGTAATTACAGGAAGAATTGGAAAGGAGCATGGCACAGTGCCCATAGGCAGGTGCATGCTGTC
CTTACCCACCAGGAAAGTAGCCCCTTC

>Human|11:49108620:49108719

ACAGACAAATCAGGAAGACCTGCAATATCCATGAGGAATCAGAATCCTTTCAAGATGAAGCAGAGAGG
GTGAACACAATATATTATGGATTTGAGAACCC

>Human|11:78300454:78300554

GGACCAGGCTTATAAAGCTCACTAAAAGTGTGACAATGAAACATTACAGTATATGCCTCCCAGTGTGAT
GCAGCAGGACATACTTAGTATACCTAGAATAT

>Human|1:13214501:13214616

TATGAATGAATCCAGTCCAGAAATGCCCACCCTGCCCCCTGCTGGCTCCTGGGGCTCTGCTCTTTGGGG
GAATCATGATGAAATTGTGGCAGAGAGTAGAAGTTGAGCCCCATTGC

>Human|11:104943827:104943950

AAAAACTTTCTTGTGTCATTACCTTGTTATCTTTGGAGAATAAGGACCAGGAGTCACAAAAACATCCACC
TTTTGACATGTGGGCCACCAATGTTACTTCCTGTGCTCTAAGCGATGAGTTAAA

>Human|16:32922979:32923078

GTTAAGACTGAATATTTGGTGTAATGGTGCATCACATTATTCTAGCTTCCATACTAGTTGTTTTATTTGTT
TGTTTTCTCCTTTGTTGGCATTTGGTTTC

>Human|16:33744583:33744682

GAAACCAAATGCCAACAAAGGAGAAAACAAACAAATAAAACAACTAGTATGGAAGCTAGAATAATGT
GATGCACCATTACACCAAATATTCAGTCTTAAC

>Human|16:33769118:33769218

GGATCTTCAATAGAAACACTCTTGTTTACAGATTTGCTCTGTGATGTGTGATTAGAGATGATTTTCTCAT
CTCAGGAACAATAAGAATCAGAAGCTGAAAC

>Human|3:14132416:14132527

CACCAAAATAGCAGACTTTCAACAATCTCTGTTCTTAGGACATTGATGGGATTTAAAGTCTTTTCTCTGA
ATCCCTGAAGATAGTTATGTAGTTAAAACTCATGCCAGAAAA

>Human|3:43045007:43045113

TAGATGATAGACAGAATCAAACTCAAGTCACTCCCTGCCAGAGTCATCTGGCAGCAGAGTGGTAGCAC
ACTGGTAGCAGAGTGAAGGCAGCCTGCTACCAGATAAAA

>Human|2:131062944:131063048

AGACTCACCAGGAAATGGAGAGCCAAGTAGAGAAACCAAGGCAGCTATCCAAGAATAAGACATCATTC
AATGGGTGGAAGACTTTTGCAAAGTCAAAATTGAATA

>Human|5:70988362:70988468

GCTAGGGTCCTCCAAGGAGTGGATGAAGTATCCTAACAGGGCATATTGGGCAATGCCAGCTGCAAAGA
AGAGGCCACCCATTGATCCATATTTAGGGAAAAGAGAAC

>Human|7:62537666:62537774

GCCCCCAAAATTCCACTCCTGGCCTGCATGGGTCCCAGTATCTGCCAGGCCTGCTGTTGACACCAAAAC
AGCCATTGCTGTGGCCCTCAGCCCTGCCTCTCAAGGCCTT

>Human|9:89404350:89404457

ATGCTGAAAAATAATATGTTGGCCTAGATAATACCTAAACAATACCTACATGTTGACCTCAACAATACCA
AATTGCATGTTGGCCTAGACAATATCTAAACAATACCT