Original research article
iDarwin volume 2, pages 3-33
Published on May 1, AS 0022 (2022 AD)

Primate deep conserved noncoding sequences and non-coding RNA: their possible relatedness to Central Nervous System

Nilmini Hettiarachchi[1, 2]

[1]Bioinformatics Institute and [2]School of Biological Sciences

University of Auckland, 3a Symonds Street, Auckland Central 1010, New Zealand

Email: nilminihett@gmail.com; n.hettiarachchi@auckland.ac.nz

Abstract

Background Conserved Noncoding Sequences (CNSs) are extensively studied for their regulatory properties and functional importance to organisms. Many features such as location, proximity to the likely target gene, lineage specificity, functionality of likely target genes and nucleotide composition of these sequences have been investigated, thus have provided very meaningful insight to signify underlying evolutionary importance of these sequences. Also thorough investigation around how to assign function to noncoding regions of eukaryote genomes is another area that is studied. On one hand evolutionary analyses, including

signatures of selection or conservation which can indicate the presence of constraint, suggest that sequences that are evolving non neutrally are candidates for functionality. On the other hand evidence that is based on experimental profiling of transcription, methylation, histone modifications and chromatin state also solidify the functional importance of CNSs. While these types of data are very important and are associated with function in most cases, this is not always the case. Evolutionary conservation of noncoding elements that are identifiable in more than one species, is still being used as the initial guideline in investigating function via experiments. As there may be patterns that are often specifically associated with potentially functional elements, fully understanding the experimental profiles of conserved noncoding regions may help us construed functionality of conserved noncoding regions easily.

Results In an effort to integrate experimental profile data, we investigated evidence of expression of conserved noncoding sequences (CNSs). For CNSs from ten primates, we assessed transcription, histone modifications, level of evolutionary constraint or accelerated evolution, possible target genes, tissue expression profiles of likely target genes and clustering patterns. In total we found 153,475 CNSs conserved across all ten primates. Of these 59,870 were overlapping non coding regions of ncRNA genes. H3K4Me1 marks (often associated with active enhancers) were highly correlated with CNSs whereas H4K20Me1 (linked to, e.g. DNA damage repair) had high correlation with conserved ncRNA regions (ncRNA-gene-CEs). Both CNSs and conserved ncRNA showed evidence of being under purifying selection. The CNSs in our dataset overall exhibited lower allele frequencies, consistent with higher levels of evolutionary constraint. We also found CNSs and ncRNA-gene-CEs produce mutually exclusive groups. The analyses also suggest that both types of conserved elements have undergone accelerated evolution, which we speculate may indicate changes in regulatory requirements following divergence events. Finally, we find that likely target genes for hominoidae specific, primate and mammalian common CNSs and ncRNA-gene-CEs are predominantly associated with brain-related function in humans.

Conclusion The deep conserved primate CNSs and ncRNA gene-CEs signify functional importance suggesting ongoing recruitment of these sequences into brain-related functions, consistent with King and Wilson's hypothesis that regulatory changes may account for rapid changes in phenotype among primates.

Keywords: CNSs, ncRNA, histone modification, accelerated evolution, purifying selection, brain evolution

Introduction

Conservation in the non-coding regions of genomes are elaborately studied over more than a decade and still it keeps revealing intriguing information about the non-coding regions of genomes, which we once considered as "junk" DNA. It is very clear now that once referred to as "junk" shapes the lineage and species individuality.

Assessing the functional potential of non-protein coding regions of genomes is a tasking process. Two broad approaches focus on either evolutionary signals of conservation and non-neutrality (Sandelin et al. 2004; Woolfe et al. 2005; Vavouri et al. 2007; Elgar 2009; Lee et al. 2011; Matsunami and Saitou 2012; Takahashi and Saitou 2012; Hettiarachchi et al. 2014; Hettiarachchi et al. 2016; Babarinde 2021) or 'activity' signals, such as expression, epigenetic or chromatin states. These approaches have led to strikingly different conclusions regarding the extent of the genome which is functional. Evolutionary analyses estimate that ~10% of the genome is under constraint, whereas biochemical activities have been attributed to ~80% of the human genome (The encode project consortium 2012). This has led to major debate regarding the identification of biological function (Graur et al. 2013). On one hand, it is clear that evolutionary analyses may be conservative, in that they frequently rely on conservation of an element in more than one species in order to predict function. Thus, they may miss species-specific elements or activities. In contrast, a difficulty with assessing functional regions through measurement of biochemical activities is that it is not clear how to separate bona fide functional activities from background non-functional regions. While clearly conservation reliant computational approach and biochemical measurements are both useful in identifying functional elements, the debate that followed the publication of the ENCODE project (The ENCODE Project Consortium 2012) suggests that a cautious approach is required to extend evolutionary analyses to include biochemical activity data (Graur et al. 2013; Doolittle 2013). In an attempt to understand whether biochemical activity data relate to evolutionarily derived functional annotation, we here integrate some of these data into an analysis of conserved noncoding elements from primates. We initially used annotation data to divide these conserved noncoding sequences into CNSs and ncRNA and downstream use expression data on these sequences and find that distinct types of histone modification are associated with CNSs and conserved ncRNAs. For CNSs, it is thought that they may act on adjacent protein-coding genes (Woolfe et al. 2005; Elgar 2009; Matsunami and Saitou 2012), whereas the trans-acting nature of ncRNAs suggests that need not be the case for this class of noncoding elements. We nevertheless applied this criterion, as some enhancers show evidence of transcription (Andersson et al. 2014; Melamed et al. 2016; Tippens et al. 2018), blurring the lines between ncRNAs and cis-acting regulatory elements. Further Awan et al. (2017)

have found that primate specific Long ncRNAs (lncRNAs) and microRNAs (miRNAs) are two important RNA classes with regulatory functions while Saber et al. (2016) have investigated the importance of highly conserved coding and noncoding regions in higher order primates.

In this study we have extensively assessed and analysed primate common mammalian common and Hominoidae specific, CNSs and ncRNA sequences with respect to their functional significance through histone modifications, genomic location, target gene expression patterns and clustering patterns. It is vital that we understand the evolutionary importance and dynamics of conserved noncoding regions in humans and our closest relatives. The functional importance of these regions are worth investigating further as most of these regions are vital for proper functioning of the primate genomes. Therefore our motivation behind this study solely lies in producing a very thorough set of conserved noncoding sequences that can be at a later stage experimentally verified for its vital

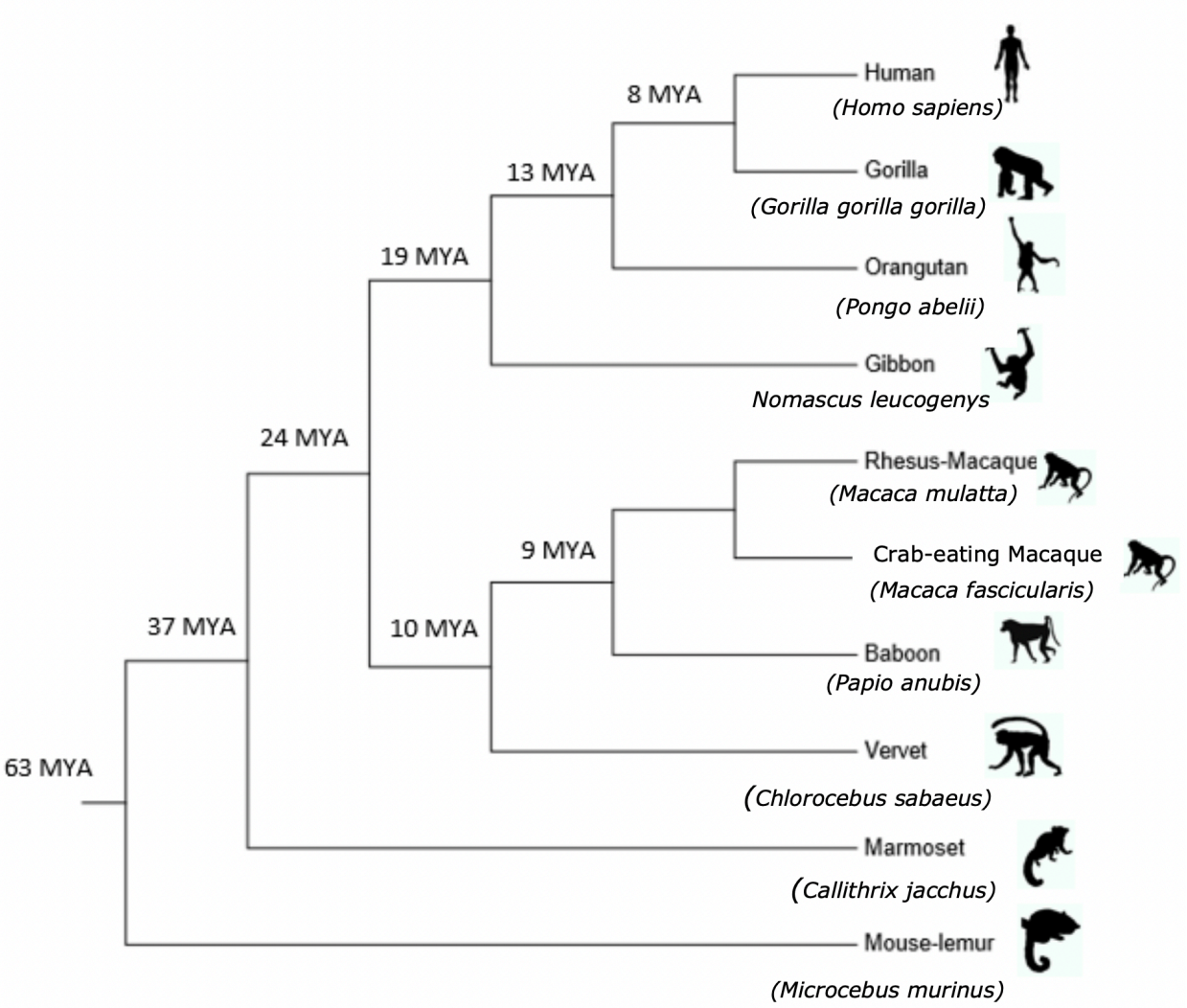


Figure 1. Phylogenetic relationships between the primate species used in the study. (MYA – Million Years Ago). The data for these species were downloaded from Ensembl release 94. The divergence times are according to Goodman et al. (1998) and Glazko and Nei (2003).

functionality in primates. While we would be hesitant to assign function based on the non-evolutionary components of our analyses (histone modification, genomic location), these may nevertheless be helpful parameters in determining the function of genomic elements where comparative data are absent.

Materials and Methods

Genomes used in the analysis

Primate repeat masked genomes of Human (*Homo sapiens)*, gorilla (*Gorilla gorilla gorilla*), orangutan (*Pongo abelii*), gibbon (*Nomascus leucogenys*), rhesus macaque (*Macaca mulatta*), crab eating macaque (*Macaca fascicularis*), baboon (*Papio anubis*), vervet (*Chlorocebus sabaeus*), marmoset (*Callithrix jacchus*), mouse lemur (*Microcebus murinus*) were downloaded from Ensembl release 94 (Yates et al. 2020). All genomes used were above minimum of 5X coverage. Figure 1 depicts the phylogenetic relationships between the species used in the study. The divergence times used for the work are according to Goodman et al. 1998; Glazko and Nei 2003.

Masking coding regions

CDS coordinates were downloaded for respective genomes from Ensembl BioMart release 94 (Yates et al. 2020). The CDS coordinates of different transcripts were merged to remove overlapping or redundant coordinates. If a pair of coordinates were overlapping they were merged into one continuous length. The merged CDs coordinates were masked in the respective genomes, which resulted in only noncoding genomes.

Homology search

After masking coding regions we searched for genomic locations that are conserved across all 10 primate species used in the study. BlastN 2.6.0 (Altschul et al. 1997) with parameter e < 0.001 was used to determine homologous regions in a whole genome chain search starting from the most closely related species pair used in the study (human and gorilla). The resulting sequence hits were filtered to obtain the best hit with lowest e-value (multiple hits per sequence occurrence were filtered only to keep the best hit scenario with lowest e-value) and these sequences (the reference genome sequences [human]) was used as the query to search in the evolutionarily next closest species. Likewise a chain search was followed step by step through the 10 primate species considered in the study. Mitochondrial DNA was removed from the analysis.

In setting the percentage conservation threshold for the CNSs the methodology introduced by Babarinde and Saitou (2013) was followed. Initially the nucleotide sequences of 1:1 orthologs of protein coding genes of human and mouse lemur were extracted from Ensembl release 94 (Yates et al. 2020). After Blastn search ($e < 0.001$) of these gene nucleotide sequences the average conservation percentage identity was obtained. This average percentage identity was used as the cut off threshold to extract the primate common CNSs. In other words the primate common CNSs had to be present in all 10 species used in the analysis while having a percentage identity exceeding the protein coding gene conservation level of the two most distant species used in the study (human and mouse lemur).

The determined commonly conserved sequences were classified into two groups based on how many conserved regions overlapped with annotated noncoding RNA gene regions in human. The annotation data for ncRNA genes is based on Ensembl version 94 (Yates et al. 2020). Any conserved noncoding sequence that overlapped with annotated ncRNA genes were classified as primate common ncRNA-gene-CEs (ncRNA-gene-Conserved Elements) and remaining sequences are referred to as CNSs (Conserved Noncoding Sequences).

Determining histone modification and DNaseI hypersensitive site overlaps for primate common CNSs and ncRNA-gene-CEs

The chip-seq data for histone modifications (H3K4Me1, H3K4Me2, H3K4Me3, H3K9ac, H3K9Me3, H3K27ac, H3K27Me3, H3K36Me3, H3K79Me3, H4K20Me1) were downloaded from UCSC (http://genome.ucsc.edu/encode/downloads.html) which is based on a direct submission from ENCODE production data (Human Genome Build 37 (hg19)). Specifically cell-line GM12878 data for histone modification broad peaks were considered. GM12878 cell-line is mentioned as having a relatively normal karyotype. Also DNaseI hypersensitive site hotspot data for 2 replicates were downloaded from the above mentioned link.

Since the originally submitted ENCODE data was based on Grch37 human genome build, the coordinates needed to be lifted to the latest version of human genome which is used in the analysis (Grch38). UCSC browser lift-over (https://genome.ucsc.edu/cgi-bin/hgLiftOver?hgsid=696327251_5kBwkhosCaqhOqgcpSUNVALQIBaf) was used to convert the ENCODE data coordinates to the latest coordinates.

The converted coordinates of histone modifications and DNaseI hypersensitive sites were checked for overlaps with the primate common CNSs and ncRNA-gene-CEs. The median signal strength value was considered as the most reliable measure of

the functional signature strength across CNSs and ncRNA-gene-CEs, since the distribution of the signal strength values were not normally distributed. Normality test for skewness and kurtosis was performed with shapiro-wilks test.

In order to determine the difference between conserved regions with respect to neutrally evolving regions in the genome, we also picked random samples of sequence coordinates from rest of the noncoding regions of the human genome. We made 25 coordinate datasets for each histone modification. These random regions were picked on a very stringent criteria that they do not overlap with coding regions, already determined conserved noncoding regions in the current study or repeat regions of the human genome. These random data sets contain same number of sequences as the original dataset of CNSs and ncRNA-gene-CEs. Also the random sequences were extracted in such a way to have the same length and same chromosome as a particular CNS or conserved ncRNA.

Selection pressure on primate common ncRNA-gene-CEs and CNSs

We downloaded 1000GP human single nucleotide polymorphism data (ftp://ftp.ensembl.org/pub/release-94/variation/vcf/homo_sapiens/) for Yoruba population in Ibadan, Nigeria (YRI), Han Chinese in Beijing (CHB), China and individuals with Northern and Western European ancestry (CEU) collected from Utah. Then we determined how many SNPs overlap with primate common CNSs and primate common ncRNA-gene-CEs. Random coordinate sets from human genome were picked in order to compare between conserved regions and non-conserved regions. The random samples have the same number of instances, same length, and were picked from the same chromosome as the CNSs or the ncRNA-gene-CEs. The random samples were normalized and tested for statistical significance by chi-square test.

Evolutionary rates of CNSs and ncRNA-gene-CEs

The identified CNSs and ncRNA-gene-CEs were concatenated into two sequence groups and the evolutionary rates for all primate common CNSs and ncRNA-gene-CEs at each branch were estimated by constructing neighbour-joining trees (Saitou and Nei 1987) with 1000 bootstrap replications. Evolutionary distances were calculated with maximum composite likelihood method with MEGA 5.0 (Tamura et al. 2011). The objective of this analysis was to determine if these CNSs and ncRNA-gene-CEs have gone through a fast-evolving phase before stabilizing and which branches show accelerated evolution.

Functional classification of genes in close proximity to CNSs and ncRNA-gene-CEs

First we determined the orthologous genes that are closest to CNSs and ncRNA-gene-CEs. These closest genes in the reference genome (human) were considered as the likely target gene, if they had an orthologous gene in the most distant outgroup species used in the study (mouse lemur). For these gene sets we determined functional enrichment via PANTHER14.1 classification system (Thomas et al. 2003).

Likely target gene tissue expression patterns

In order to determine if the likely target gene tissue expression patterns follow the same trend as Gene Ontology (GO) analysis we retrieved Riken Fantom 5 project data for adult human and fetal data for 76 tissues through EBI (https://www.ebi.ac.uk/gxa/home) expression atlas. For this analysis we used 8089 likely target genes for CNSs and 7421 likely target genes for ncRNA-gene-CEs that were found in the expression data. The genes that had any sign of expression were considered for further analyses. The number of genes expressed for brain and nervous system related functions and several tissues related to housekeeping functions were considered as two potential groups for better comparison. Also we constructed 10 random gene sets for each group (CNSs and ncRNA-gene-CEs) for clarity on random expectation. The statistical significance of the data and random samples were determined by one-sample t-test with statistical significance level of 0.05.

Clustering patterns of CNSs and ncRNA-gene-CEs

We determined the tendency of primate common CNSs and conserved ncRNA occurring in clusters. This analysis was conducted by considering groups of CNSs and ncRNA-gene-CEs that clustered with one likely target gene (considering closest orthologous gene as described before). This analysis determines how many CNSs or conserved ncRNAs are associated with its closest orthologous target gene and their occurrence trends in the reference genome. We classified the CNSs and ncRNA-gene-CEs into two groups for further comparisons, that is genes with 1-2 CNSs / 1-2 ncRNA-gene-CEs and genes with >6 CNSs / >6 ncRNA-gene-CEs.

Brain related tissue expression levels for genes associated with CNSs and ncRNA-gene-CE clusters

Furthermore we determined if genes with many CNSs or ncRNA have difference in their expression level in diverse brain related tissues. Here we used genes expressed in four tissues namely amygdala, caudate nucleus, cerebellum and spinal cord. To have an elaborate view on tissue expression we considered genes that showed any kind of expression without a baseline TPM (Transcripts per Million) threshold. The objective of this analysis is to see any differences or similarities in gene expression patterns for CNSs and conserved ncRNA in clusters and how the 2 groups are functioning on the genomic level. This helps to shed light upon

understanding the evolutionary dynamics of CNSs and conserved ncRNA with respect to expression. Statistical tests for significance were determined by Mann-Whitney U test.

CNSs and ncRNA-gene-CEs relation to experimentally verified enhancers, promoters and CTCF binding sites

We determined if our primate common CNSs and ncRNA-gene-CEs are actually related to predicted enhancer, promoter and CTCF binding sites in (ensemble.org/info/genomeensembl.org/info/genome/funcgen/regulatory_build.html). Based on the premise that, if CNSs are actually governing gene regulation we expect more enhancers to be associated with CNSs. Also checking the magnitude of ncRNA-gene-CEs association to enhancers helps to elucidate the functional landscape of conserved ncRNA as well. CTCF binding sites are considered important and conserved across lineages functioning as insulator elements, blocking enhancer activity functioning as domain barriers (Hou et al. 2008; Cuddapath et al. 2009; Herold et al. 2012). Promoters help to get an understanding on which regions have higher tendency to get transcribed, by comparing association of our CNSs and ncRNA-gene-CEs to documented promoter regions and CTCF binding sites also help to differentiate CNSs and ncRNA-gene-CEs and their specific role in a genome.

We compare our dataset against random regions with same length and genomic location (chromosome) as CNSs or ncRNA-gene-CEs of the reference genome. Statistical significance was determined by one sample t-test.

Also to be confident that the pattern of representation is not dependent on the initial number difference in CNSs and ncRNA-gene-CEs, we further randomly picked 20 samples of 59,870 CNSs (same number as identified ncRNA-gene-CEs) and searched for the signals of enhancers, CTCF binding and promoters in the 20 sampled datasets. Also we determined the functional categories for the likely target genes of CTCF binding, enhancer and promoter overlapping CNSs and ncRNA-gene-CEs.

Determining the presence of primate deep conserved elements in mammals for more ancient conservation

We tried to determine how many of the primate common conserved sequences are found in mammalian common ancestor. For this analysis we searched primate common CNSs and ncRNA-gene-CEs separately in rat (*Rattus norvegicus*) and mouse genomes (*Mus musculus*) in Ensembl release 94 (with Blastn search ($e < 0.001$). The sequence homologs that were found in rat and mouse both were presumed to have originated in mammalian common ancestor. For the mammalian common sequences the likely target gene GO was determined via PANTHER14.1. Also we looked at the tissue expression patterns of these genes via Riken Fantom 5 project data for adult human and fetal expression atlas through EBI (https://www.ebi.ac.uk/gxa/home).

Have Hominoid specific conserved non coding elements originated to render specific functions to hominoid lineage?

We determined the Hominoid lineage specific elements that are unique to human, gorilla, orangutan and gibbon by searching for homologs of this group in all primate outgroup species. Any sequence instance that was also found in any of the outgroup species were removed from the final data set or any further analyses. Finally we considered only the sequences above 90% percentage identity for further analyses. These hominoid unique elements were checked for their closest likely target gene related functions and tissue expression patterns.

Do CNSs and ncRNA-gene-CEs belong to independent categories?

Since we observed similar functional categories for CNSs and ncRNA-gene-CEs we identified in the study, it was important to know if there were any CNSs and ncRNA cross homology sequences causing this scenario. Therefore we searched all the conserved noncoding sequences against itself with Blastn with e<0.001. After removing self-hits and one copy of reciprocal hits, the remaining hits with a CNS and ncRNA combinations were extracted (4,772 sequence instances). The coordinates of these CNS, ncRNA cross homology entities were searched for their features with Ensembl feature category based on annotation.

Results

Identification and characterization of a highly conserved CNS and ncRNA dataset spanning primates

To generate a conservative, high confidence set of CNS and ncRNA sequences, we screened for conserved noncoding regions present in 10 primates. This yielded 153,475 primate common conserved noncoding regions. Of these, 59,870 are candidate ncRNA gene overlaps (overlapping with annotated ncRNAs) while 93,605 sequences were found in other non-coding (Intergenic, Untranslated Regions [UTR], introns of protein coding genes) regions of the human genome (reference genome used in the study). Based on these assignments, we will refer to these conserved noncoding regions as ncRNA-gene-CEs and CNSs, respectively throughout the results section.

Epigenetic profiles of primate CNS and ncRNA gene-CEs

H3K4Me1 is a histone modification associated with active enhancer regions (Koenecke et al. 2016; Barakat et al. 2018). The median signal strength of H3K4Me1 located inside CNSs were slightly higher but not statistically significant compared to random coordinates (Figure 2A). Also another histone modification that is associated with activated enhancer regions is

H3K27ac. The CNSs have significantly higher levels of H3K27ac compared to random samples ($p < 0.0001$). H3K79 methylation is known to be associated with developmentally regulated gene expression in multicellular eukaryotes. For this histone modification, both primate common CNSs and ncRNA-gene-CEs have high levels of signal strength compared to random samples. H3K9ac which is a histone modification mark predominantly found in active promoter regions showed high signal strength in CNSs with respect to random sample ($p = 0.003$). In general according to our results, CNSs show a higher tendency to be active enhancer elements associated with regulatory activities.

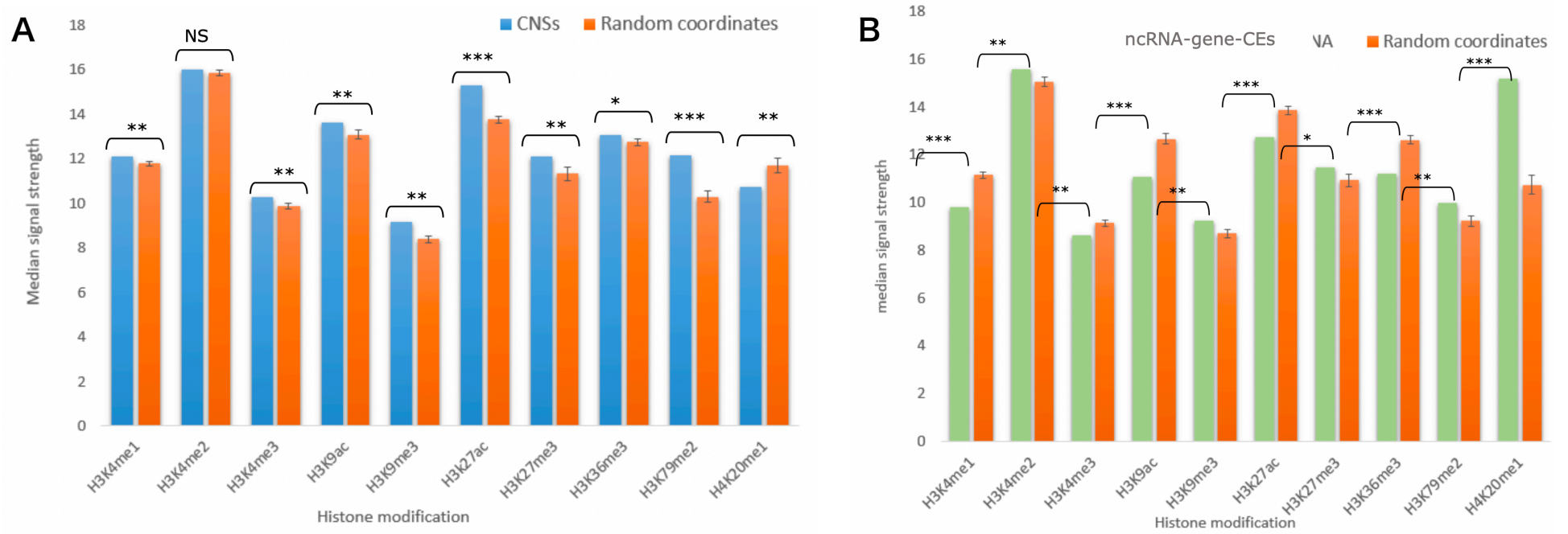


Figure 2. (A) Histone modification median signal strength for primate common CNSs. (B) Histone modification median signal strength for primate common ncRNA-gene-CEs. Specifically CNSs show higher signal strength for H3K27ac (histone marks associated with active enhancer regions) compared to conserved ncRNA and also random expectation. Conserved non-coding RNA shows significantly higher signal strength for H4K20me1. Statistical significance was determined via one-sample t-test. ($P < 0.0001$ - ***, $P > 0.001$ = **, $P > 0.01$ = *).

H3K4Me1 in ncRNA-gene-CEs were statistically significantly lower than random expectation ($p < 0.0001$), signifying that these primate common ncRNA-gene-CEs may have less chance of functioning as active enhancers. Also H3K27ac, another active enhancer histone modification showed significantly less strength in ncRNA-gene-CEs compared to random samples. This result implies that the ncRNA-gene-CEs in this case may not be associated with active enhancers as CNSs.

One interesting result is that, H4K20Me1 shows significantly higher levels of association with primate ncRNA-gene-CEs compared to CNSs (Figure 2B). This histone modification is involved in DNA replication and damage repair. *CCND1* is the first long ncRNA that was identified as being transcribed in response to DNA damage signals (Wang et al. 2008; Klein and Assoian 2008). *LincROR* has been identified as a P53 repressor in response to DNA damage (Zhang et al. 2013). In addition several other studies have revealed ncRNA association to DNA repair and replication (Ge and Lin 2014; Hawley et al. 2017). The fact that we found primate common ncRNA-gene-CEs had a high signal strength for H4K20Me1 further signifies association of ncRNA-gene-CEs with DNA damage repair. The number of histone modification sites found inside CNSs and ncRNA-gene-CEs are given in Table 1.

Histone modification	N* CNSs	N* ncRNA-gene-CEs
H3K4Me1	144	100
H3K4Me2	108	96
H3K4Me3	79	53
H3K9ac	44	32
H3K9Me3	44	20
H3K27ac	45	50
H3K27Me3	36	23
H3K36Me3	31	54
H3K79Me2	9	9
H4K20Me1	8	15

N*: Number of regions located inside

Table 1. Number of histone modification signals located inside CNSs or ncRNA-gene-CEs

DNaseI hypertensive sites (DHSs) in conserved noncoding regions

We found that CNSs have significantly less abundance of DNaseI hypersensitive sites associated with them ($p<0.0001$), compared to random expectation. Whereas ncRNA-gene-CEs show an opposite trend where they have more DNaseI sites in them ($p<0.0001$), compared to random samples (Figure 3). DHSs signify open chromatin regions according to many reports (Song et al. 2011; Madrigal and Krajewski 2012; Li and Cui 2018). Therefore this result implies that the primate common CNSs we identified are located in regions with less open chromatin conformation compared to random regions in the human genome. But

primate common ncRNA-gene-CEs have more DHSs compared to the random expectation implying a higher tendency for ncRNA-gene-CEs to be in open chromatin conformation.

Purifying selection on conserved noncoding regions

In order to test for the level of selection of primate common CNSs and ncRNA-gene-CEs we carried out the derived allele frequency analysis. Functional regions under purifying selection are expected to have lower levels or lower frequency of derived alleles than regions that are neutrally evolving. Our results show that the primate common CNSs and ncRNA-gene-CEs both have a high percentage of lower derived alleles than neutrally evolving random sequences. Alleles with lower frequency levels (DAF< 0.1) in both CNSs and ncRNA-gene-CEs show significantly higher level of abundance than expected. (Z-test p<0.00001).

Also it is important to note that high frequency alleles are less abundant in evolutionarily conserved regions than neutrally evolving background sequences of the human genome. Abundance levels of SNPs with derived allele frequency>0.9 are significantly higher (z-test p<0.00001) in neutrally evolving regions compared to conserved regions in our results. Also it is clear that ncRNA-gene-CEs are under slightly higher levels of purifying selection than CNSs. This scenario is uniform with both Yoruba and Han Chinese population (Figures 4 and 5). With Yoruba population, CNSs had a percentage level of 47.65% (Figure 4A) for lower allele frequency whereas ncRNA-gene-CEs had 49.81% (Figure 4B). Han Chinese and European (Figures 5 and 6) SNP data also showed that ncRNA had a higher percentage level (24.3% and 50.02%, respectively) compared to CNSs for lower derived alleles (<0.1). For Yoruba population the statistical significance levels of CNSs and ncRNA-gene-CEs SNP

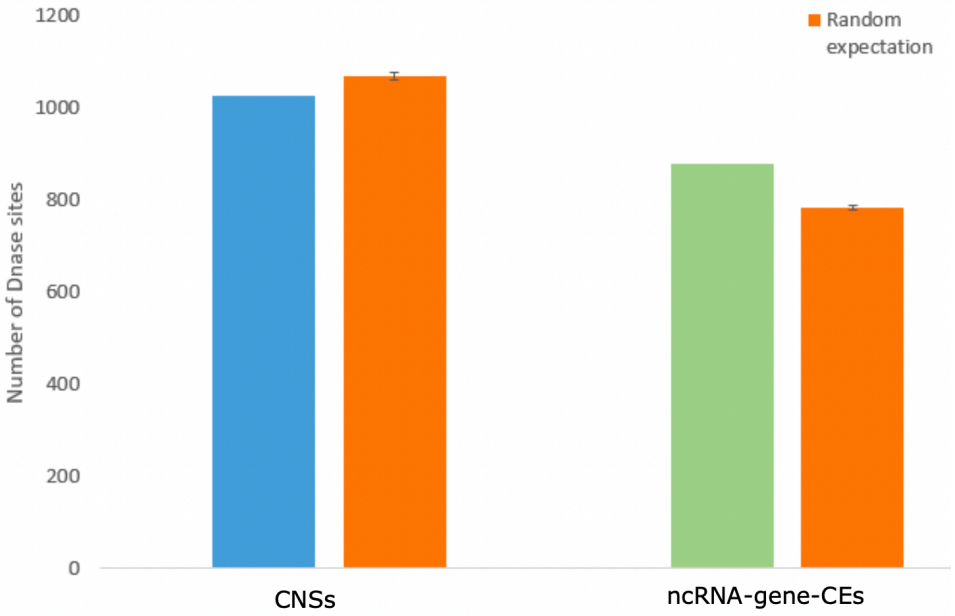


Figure 3. Dnase I hypersensitive sites associated with primate common CNSs and ncRNA-gene-CEs. The sites that are found inside conserved regions were considered. The random expectation was determined by 25 random samples from human genome that are not conserved and are presumably neutrally evolving. The results were statistically significant (p<0.0001).

distribution from the random expectation are p < 2.05854E-49 and p <8.91508E-74 (chi-square test) respectively. In other words the observed SNP distribution in CNSs and ncRNA-gene-CEs has a very low probability of being a chance distribution.

These results show that generally, conserved non-coding regions are under purifying selection and some conserved non-coding regions, such as primate common ncRNA-gene-CEs in this analysis are under a slightly higher selective pressure than CNSs. It is already well known that conserved non-coding regions are functionally important and disruption or mutations of these conserved regions can lead to detrimental effects.

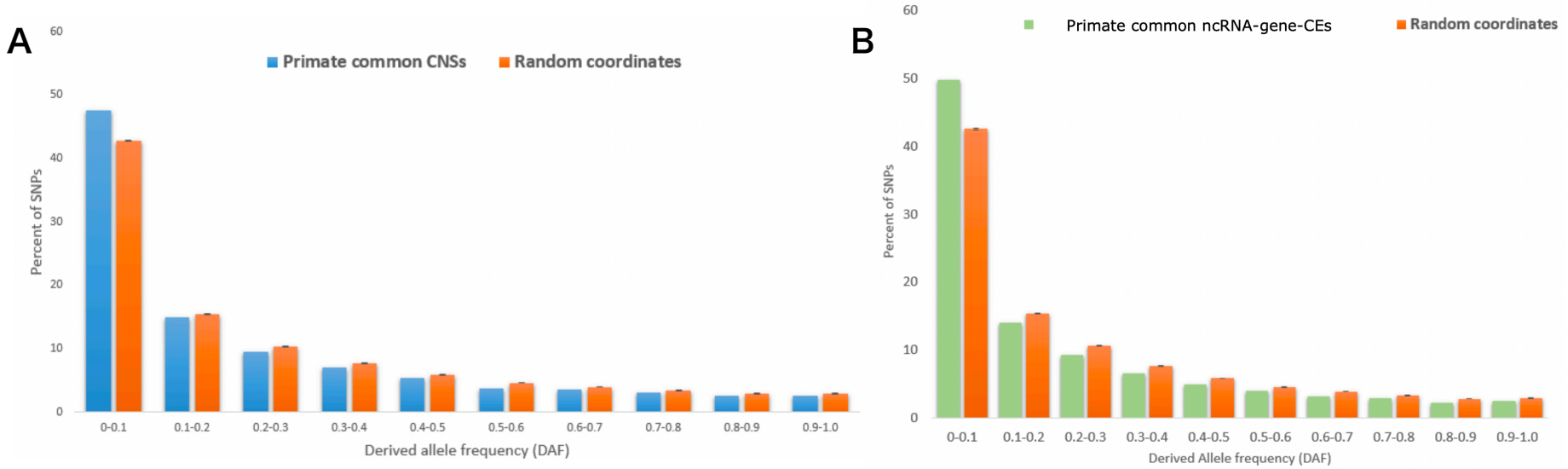


Figure 4. Derived allele frequency analysis for primate common CNSs and primate common ncRNA-gene-CEs using Yoruba population. (A) Derived Allele Frequencies for CNSs with Yoruba population. (B) Derived Allele Frequencies for ncRNA-gene-CEs with Yoruba population data. Random coordinates from the human genome build Grch38 were picked and considered as neutrally evolving regions compared to conserved regions. The 1000GP SNP data was used for this analysis. CNSs and conserved non-coding RNA both have higher percentage of SNPs for lower derived allele frequency level <0.1 compared to random expectation. The results are statically significant at p <0.05 (z-test).

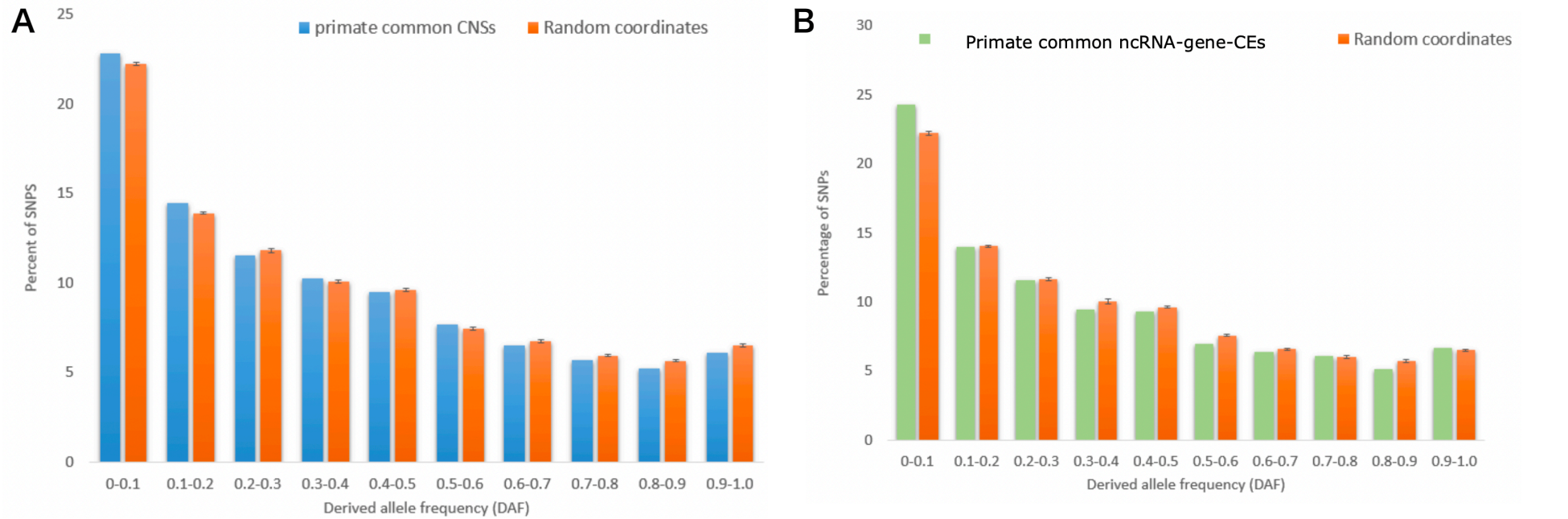


Figure 5. Derived allele frequency for CNSs and ncRNA-gene-CEs with Han Chinese population. (A) Derived Allele Frequencies for CNSs with Han Chinese population (B) Derived Allele Frequencies for ncRNA-gene-CEs with Han Chinese population data.

Primate common ncRNA-gene-CEs also show a higher level of selection or in other words, less tolerant of higher frequency mutations. Majority of the primate common ncRNA-gene-CEs we identified in this analysis were long non-coding RNA. It has been shown in many studies that long non-coding RNA can also function in gene regulation controlling gene expression (Bao et al. 2013), therefore it can be expected that conserved ncRNA to be under purifying selection.

Accelerated evolution of CNSs and ncRNA-gene-CEs

We found primate common CNSs and ncRNA-gene-CEs have gone through a phase of accelerated evolution in several different stages before getting stabilized. The human-gorilla common ancestor sequences showed 4 times faster evolutionary rate compared to human-gorilla-orangutan common ancestor for both CNSs and ncRNA-gene-CEs (Figure 7A, 7B). Also we observed that Hominoidae common ancestor showed accelerated evolution in CNS sequences after diverging from Hominoidae-

old world monkey common ancestor. Also old world monkeys show about 2 fold acceleration after diverging from Hominoidae-old world monkey common ancestor.

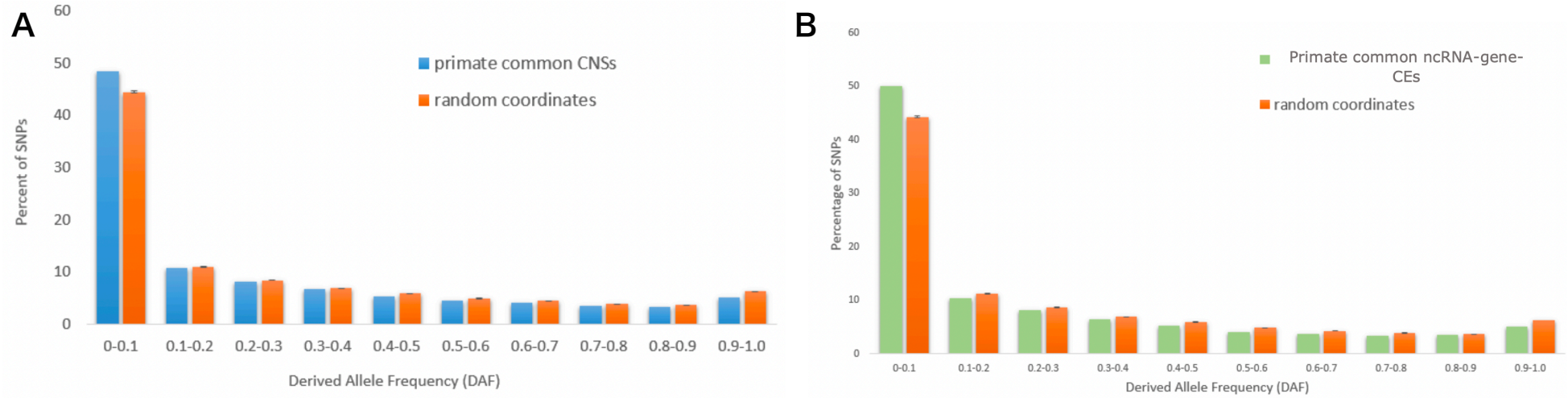


Figure 6. Derived allele frequency for CNSs and ncRNA-gene-CEs with European population. **(A) Derived Allele Frequencies for CNSs with European population. (B) Derived Allele Frequencies for ncRNA-gene-CEs with European population data.**

Functional classification of genes associated with CNSs and ncRNA-gene-CEs

Functional classification for the closest orthologous gene for CNSs showed a pattern where majority of the genes were associated with neuron development (Table 2A). Surprisingly the closest orthologous gene to ncRNA-gene-CEs also followed the same pattern (Table 2B) per gene ontology. The genes closest to ncRNA-gene-CEs were also involved in neuron development and brain related functions. Even though the fold enrichment (genes observed, over expected number) was slightly higher for genes closest to CNSs compared to ncRNA-gene-CEs, this uniform result signifies that these ncRNA-gene-CEs may be involved in cis regulation of these genes. It definitely is intriguing that these primate common CNSs and ncRNA-gene-CEs follow similar patterns in their potential functions.

Likely target gene tissue expression patterns for primate common CNSs

(A)

GO term	Fold enrichment	P-value
Regulation of cell differentiation	1.72	1.13E–03
Neuron differentiation	1.71	3.43E–07
Regulation of developmental processes	1.70	1.97E–04
Cell morphogenesis involved in neuron differentiation	1.70	5.20E–04
Nervous system development	1.62	5.59E–08

(B)

GO term	Fold enrichment	P-value
Positive regulation of transcription by RNA polymerase II	1.58	1.04E–05
Neuron differentiation	1.56	7.16E–05
Generation of neurons	1.54	5.88E–05
Neurogenesis	1.52	9.06E–05
Nervous system development	1.47	5.16E–05

Table 2. GO terms related with closest genes for CNSs and ncRNA-gene-CEs. (A)Gene Enrichment values for likely target genes for CNSs (B) Gene Enrichment values for likely target genes forncRNA-gene-CEs.

Species range	CNSs	ncRNA-gene-CEs
Hominoidae specific	9,118	3,363
Primate specific	93,605	59,869
Mammalian common	35,972	23,549

Table 3. Number of CNSs and ncRNA-gene-CEs found in hominoidae, primate, and mammal common ancestors.

(A)

GO term	Fold enrichment	P-value
Regulation of chemotaxis	3.19	1.62E–03
Glutamate receptor signaling pathway	2.77	4.60E–04
Synaptic transmission	2.70	4.61E–02
Regulation of cell morphogenesis	2.70	1.53E–03
Negative regulation of signal transduction	1.98	5.01E–04

(B)

GO term	Fold enrichment	P-value
Regulation of synapse structure or activity	4.32	9.50E–04
Positive regulation of synaptic transmission	3.28	1.17E–03
Regulation of axogenesis	3.14	1.05E–03
Glutamate receptor signaling pathway	2.95	4.05E–04
Synaptic transmission	2.89	2.09E–04

Table 4. GO terms related with closest genes for mammalian common CNSs and conserved ncRNA. (A) Gene Enrichment values for likely target genes for CNSs (B) Gene Enrichment values for likely target genes forncRNA-gene-CEs.

Majority of the closest proximity genes for both CNSs and ncRNA-gene-CEs were related to brain and nervous system related tissues. Amygdala, brain meninx, occipital lobe and spinal cord among others showed the highest fraction of expressed genes for both categories of CNSs and ncRNA-gene-CEs (Figure 8).The expressed gene fractions are statistically significant in all cases compared to random samples ($p<0.00001$).This pattern is same as what was observed in the GO analysis where high level of enrichment was shown for nervous system development. Also

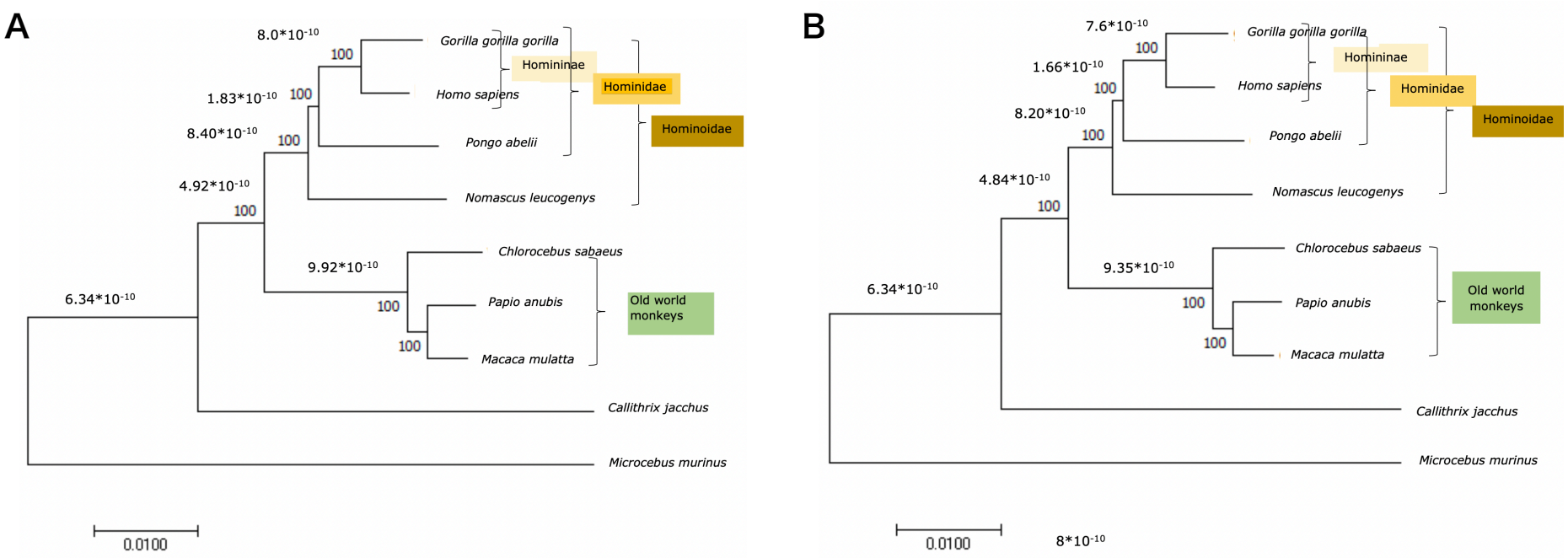


Figure 7. (A) CNSs show faster evolution at different stages in phylogeny. (B) Evolutionary rates of ncRNA-gene-CEs in primate lineage. The branches with faster substitution (>=8*10-10per site per year) are highlighted by a red bar. Primate common CNSs went through acceleration at Human-Gorilla, Hominoidae and Old world monkey common ancestor.Conserved ncRNA gene sequences also show acceleration during primate evolution. The mutation rate is as per site per year.

random gene sets that are not closest to CNSs or ncRNA-gene-CEs, showed significantly lower levels of genes expressed in brain related tissues compared to closest likely target genes. Also there is a significant difference in CNS and ncRNA percentage gene expression for brain related tissues, that is CNSs have a slightly higher tendency to harbour brain or nervous system related genes as the adjacent likely target gene compared to ncRNA-gene-CEs($P = 0.00799$, one tail t-test).

About 40% of primate common sequences are found in Mammalian common ancestor

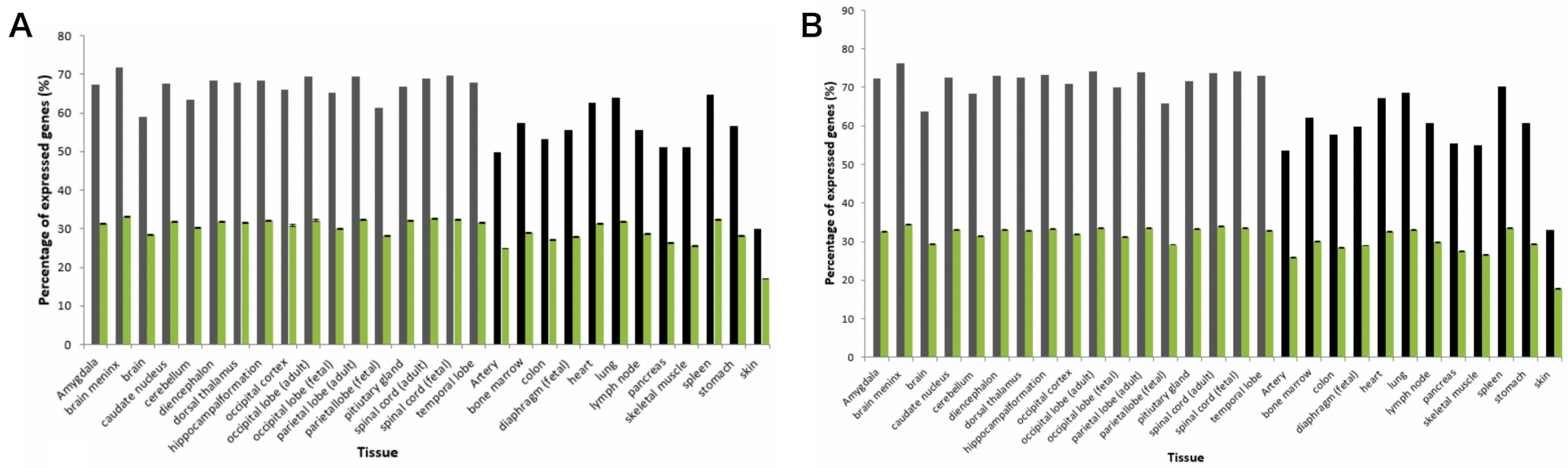


Figure 8. Primate common CNSs and ncRNA-gene-CEs closest gene expression patterns. (A) The percentage of expressed genes for CNS. (B) The percentage of expressed genes for ncRNA-gene-CEs. Grey, black bars represent brain related and house-keeping tissues respectively. Green bars signify the random expectation. Statistical significance of all results is $p<0.00001$.

We found only 35,972 CNSs and 23,549 ncRNA-gene-CEs in the mammalian common ancestor (Table 3). This means only 38% of CNSs and 39% of the ncRNA-gene-CEs found as primate common sequences already existed in the mammalian common ancestor, which are very ancient sequences. Therefore majority (about 60%) of primate common sequences we identified in this study appear to have solely originated in the primate common ancestor.

The closest likely target gene functional classification for these sequences revealed that mammalian common ncRNA-gene-CEs (while assuming a cis activity for these sequences) are related to regulation of synapse structure and activity (Table 4).

Mammalian common CNSs too show similar pattern of relatedness to central nervous system but with less fold-enrichment meaning that these ancient noncoding conservation may not be the sole determinant of brain function in primates. i.e primate evolution has taken a new turn by recruiting or generating new regulatory elements to suits its lineage necessities.

Category	Number	Proportion
miRNA target site	449	88%
TF binding related	43	9%
Histone related	15	3%
RNApolIII	1	0%

Table 5. Broad feature categories of CNS-ncRNA cross homology sequences. This analysis took a total of 4,772 cross homology elements and searched for coordinate overlaps in Ensembl feature categories.

(A)

GO term	Fold enrichment	P-value
Neuron development	2.04	2.68E–05
Neuron differentiation	1.69	7.85E–04
Chemical synaptic transmission	1.58	5.12E–04
Trans synaptic signaling	1.58	5.12E–04
Intracellular signal transduction	1.42	6.03E–05

(B)

GO term	Fold enrichment	P-value
Immune effector process	0.31	2.05E–04
Phagocytosis	0.23	3.90E–04
Defense response to bacterium	0.06	1.30E–06
Regulation of lymphocyte activation	<0.01	5.53E–06

Hominoidae specific CNSs and ncRNA-gene-CEs are less associated with immune responses but more with neuron development

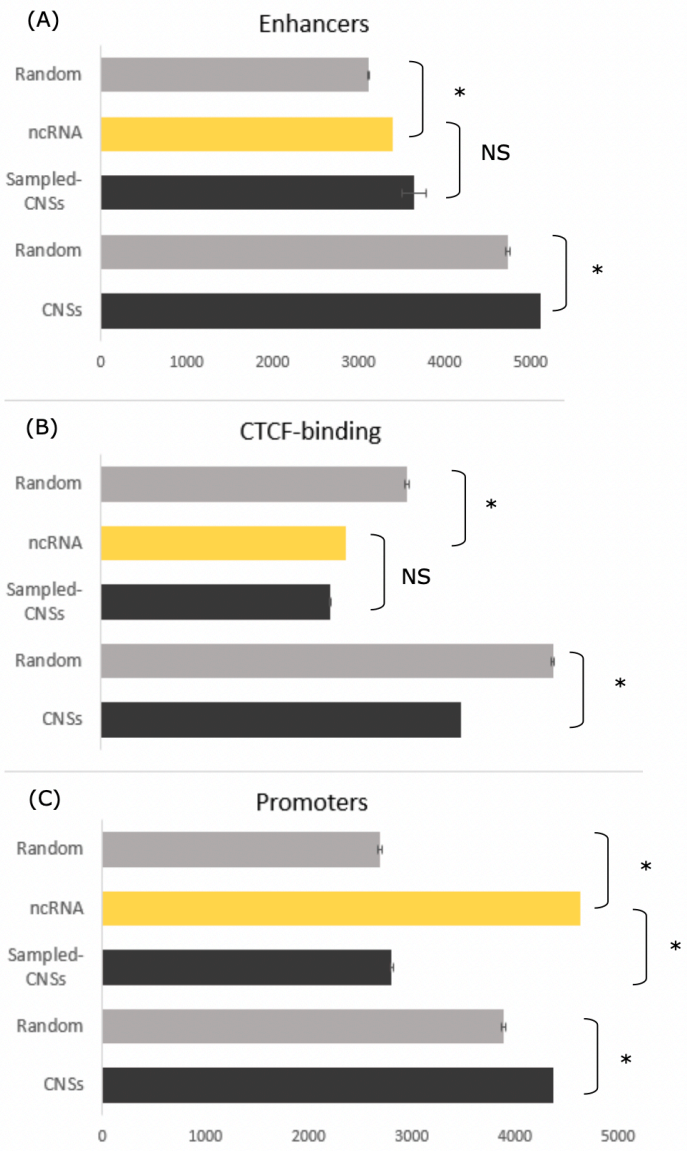
We identified a total of 12,481 elements that presumably originated and are conserved only in the hominoidae lineage. Of the total elements 3,363 were ncRNA-gene-CEs, while 9,118 were CNSs. In general terms these lineage specific elements also cluster close to genes that are predominantly related with neuron development and differentiation (Table 4).

CNSs and ncRNA-gene-CEs are mostly mutually exclusive groups of conserved elements

We found that 4,772 conserved elements showed CNS-ncRNA cross homology, In other words only 3.10% of all the identified conserved non coding elements fell into this category. The feature categories of these cross homology elements were predominantly categorized as miRNA target sites (Tables 5 and 6). Also it's important to note that miRNA target site coordinates are all non-overlapping and are located in diverse locations in the reference genome. This result shows that the CNS, ncRNA-gene-CEs predominantly fall into mutually exclusive groups which further clarifies that the previously observed similar functional categorization of the CNS and ncRNA-gene-CEs could not have been a result of cross homology sequences.

Table 6. GO terms related with closest genes for hominoidae specific conserved elements. (A) Overrepresented functional classifications of likely target genes. (B) Underrepresented GO terms.

CNSs and ncRNA-gene-CEs both show less enrichment in CTCF binding and higher enrichment in enhancer regions compared to random regions in the human genome



We see significantly higher levels of enhancers in CNSs and ncRNA-gene-CEs compared to random expectation ($p<0.05$) (Figure 9A). This further establishes the fact that primate common CNSs and ncRNA-gene-CEs play a significant role as enhancer elements. Interestingly we find that both CNSs and ncRNA-gene-CEs show less enrichment for CTCF binding compared to other random regions in the human genome ($p<0.05$), signifying these identified conserved regions have a less tendency to be insulator or repressive domains (Figure 9B). But one important feature to note is that more promoter regions are associated with ncRNA-gene-CEs in our dataset compared to CNSs, reflecting more transcriptional activity ($p<0.05$) of ncRNA-gene-CEs compared to CNSs (Figure 9C).

In comparison the randomly picked (same number of CNSs as ncRNA-gene-CEs in the original data set – sampled 20 times with replacement) CNSs with ncRNA-gene-CEs, showed that number of CNSs that harboured enhancer were not statistically significant compared to ncRNA-gene-CEs, meaning that these ncRNA-gene-CEs and CNSs have equal tendency to have enhancer elements (Figure 9A). Also sampled CNSs showed no statistical significance in abundance of CTCF binding sites compared to ncRNA-gene-CEs, signifying

Figure 9. Enhancer, CTCF binding and promoter elements in primate common CNSs and ncRNA-gene-CEs according to Ensembl regulatory track data. (A) Enhancer elements overlapping CNSs, sampled CNSs, ncRNA-gene-CEs and random regions. (B) CTCF binding elements overlapping CNSs, sampled CNSs, ncRNA-gene-CEs and random regions. (C) Promoter elements overlapping CNSs, sampled CNSs, ncRNA-gene-CEs and random regions. Statistically significant ($p<0.05$) combinations are depicted by * and non-significance by NS.

Feature Category	Number CNSs	Number ncRNA gene-CEs
Enhancers	5,110	3,403
CTCF binding	4,189	2,840
Promoters	4,366	4,632

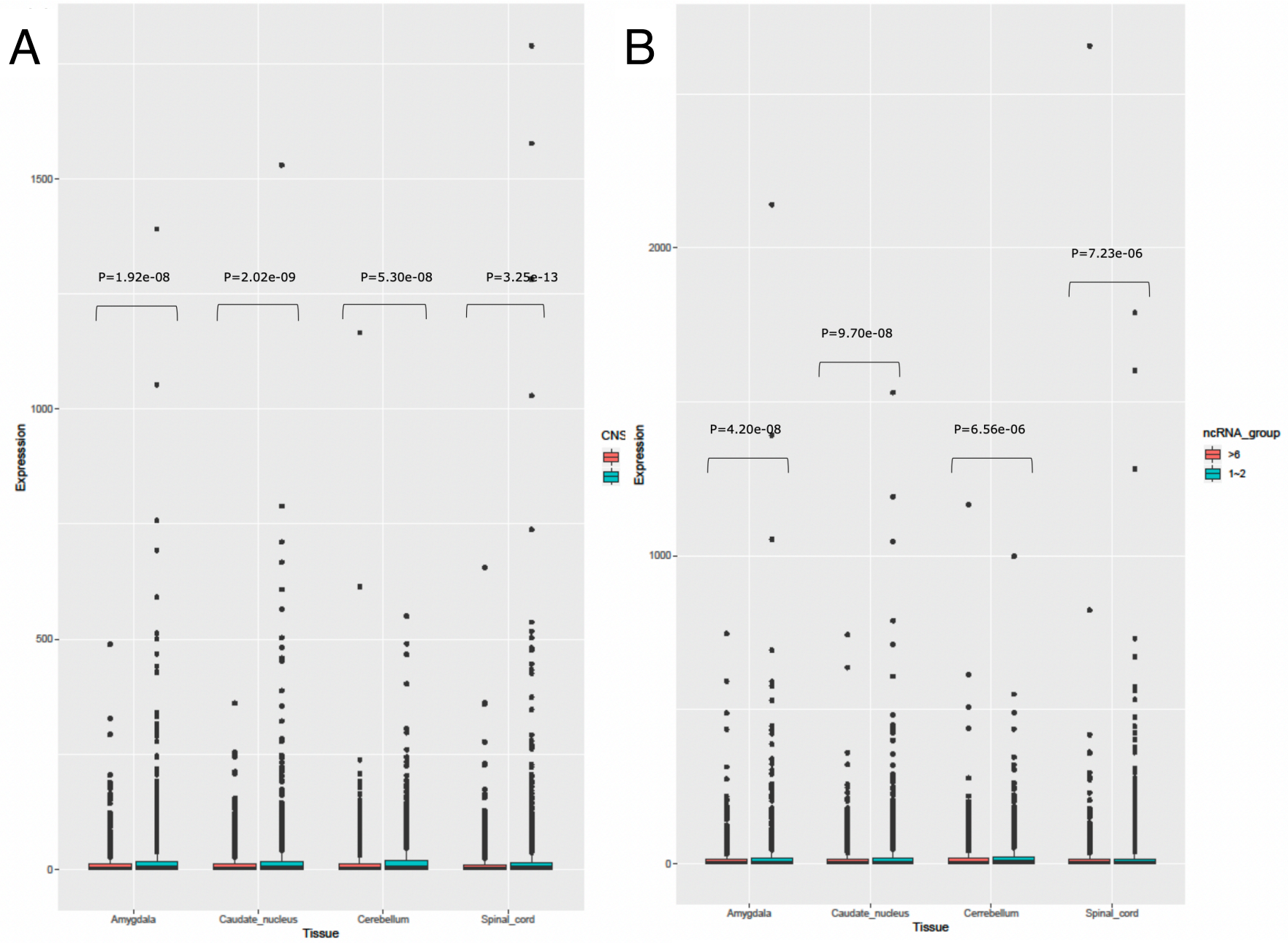
Table 7. CNSs and ncRNA-gene-CEs overlapping Ensembl verified enhancers, CTCF bindng sites and promoters.

that CNSs and ncRNA-gene-CEs have equally less tendency to function as insulator elements (Figure 9B). But promoter regions were significantly less abundant in CNSs than ncRNA-gene-CEs signifying a difference in transcription activity in CNSs and ncRNA-gene-CEs (p is < .00001) (Figure 9C) which also agrees with the comparison we made to randomly picked non conserved regions of the genome to primate common CNSs and ncRNA-gene-CEs. Also the enhancer and CTCF overlapping CNSs showed higher enrichment related to nervous system and brain related structural and functional GO terms while promoter overlapping CNSs and ncRNA-gene-CEs were more related with housekeeping structural and functional terms (Table 7). This further clarifies experimentally verified enhancer related CNSs are related to brain and nervous system function as we have noted earlier.

Expression levels of genes associated with clustered CNSs and ncRNA-gene-CEs reveal significant difference based on clustering pattern

We found 3380 and 3457 likely target genes in total, associated with 1-2 CNSs and >6 CNSs respectively. 1-2 CNS group showed more target genes being expressed in amygdala and spinal cord (Figure 10). Also in general 1-2 CNS group showed more genes being expressed compared to genes associated with more than 6 CNSs. Also we closely observed the expression levels of these genes and found that there is a significant difference in the gene expression levels for each of the 4 tissues for the 2 groups. More expression levels were seen for genes associated with 1-2 CNSs for all for tissues (amygdala, caudate nucleus, cerebellum and spinal cord) compared to CNSs in clusters of more than 6 conserved regions (Figure 10A).

Figure 10 (see next page) (A) Expression levels for genes associated with clustered CNSs.The CNS_group represents grouping pattern that was considered, i.e 1-2 means genes with 1or 2 CNSs, whereas >6 means genes with more than 6 CNSs. The statistical significance was determined by Mann-whitney U-test. The x-axis represents gene expression level for each of these groups for several brain related tissues (absolute values are used in TPM (Transcripts per Million)).
(B) Expression levels for genes associated with clustered conserved ncRNA. The ncRNA group represents grouping pattern that was considered,i.e 1-2 means genes with 1or 2 conserved ncRNAs, whereas >6 means, genes with more than 6 conserved ncRNAs. The x-axis represents gene expression level for each of these groups for several brain related tissues.


A
B
P=1.92e-08
P=2.02e-09
P=5.30e-08
P=3.25e-13
P=4.20e-08
P=9.70e-08
P=6.56e-06
P=7.23e-06
Expression
Tissue
Amygdala
Caudate_nucleus
Cerebellum
Spinal_cord
Cerrebellum
CNS
ncRNA_group
>6
1~2
0
500
1000
1500
2000

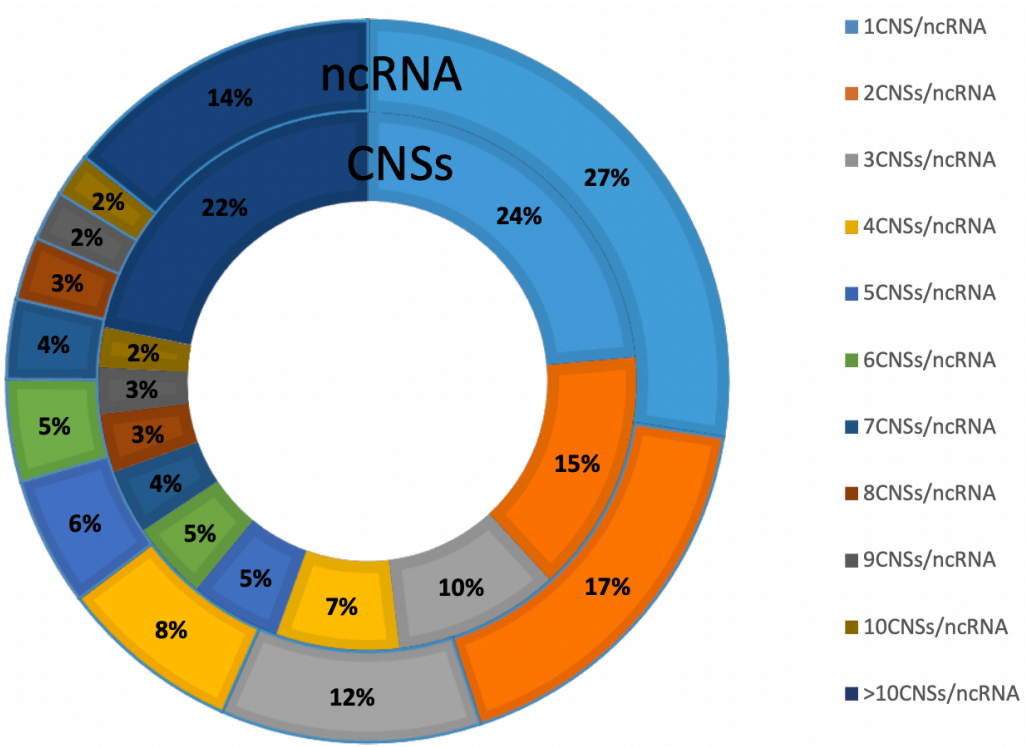


Figure 11. percentage occurrence of gene-CNS gene-ncRNA clusters of primate common conserved noncoding regions. The figure colour legend represents all cases of gene-CNS or ncRNA associations. i.e 1 gene - 1CNS/1 ncRNA association, 1 gene -multiple CNSs/ncRNA associations and percentage occurrence.

ncRNA-gene-CEs also followed the same pattern as CNSs where 1-2 conserved ncRNA cluster related genes showed higher expression levels compared to larger clusters (Figure 10B). Also further investigation showed that the genes with 1-2 CNSs or ncRNA-gene-CE clusters were very much similar in their expression levels across other tissues we considered for this analysis. It appears that multiple long blocks of ncRNA-gene-CEs or CNSs are not necessarily required to achieve high expression levels of the target genes.

Clustering patterns of CNSs and ncRNA-gene-CEs with likely target genes

In majority of the cases, one conserved noncoding region was associated with one likely target gene, for both CNSs and ncRNA-gene-CEs (Figure 11). CNSs and ncRNA-gene-CEs showed no significant difference regards to occurrence in clustering, implying that it is equally likely for CNSs and conserved ncRNA to occur in clusters.

Discussion

We have identified a set of highly conserved ncRNAs and CNSs from primates. Many are also found in other mammals, and our analyses indicate these are under selective constraint, giving us confidence that they are likely to be functional elements.

We assessed the probable function of our CNS and ncRNA dataset by looking at the closest adjacent gene. This is frequently done for CNSs, on the basis that their mode of action is local, and in cis (Nelson and Wardle 2013, Vavouri et al. 2007), while we expected no clear signal for candidate ncRNAs on the basis that these are more likely to act as trans elements. We were therefore surprised to find that the GO term enrichment of adjacent genes for both CNSs and our candidate ncRNA-gene-CEs was similar. One possibility is that the expression from these ncRNA-gene-CEs may be of type seen for some enhancer elements, where expression may open up genomic regions, making them accessible for cis-action. This phenomenon is well documented (Shen et al. 2018; Kim and Shiekhattar 2015), and our data suggest this may be prevalent across transcribed noncoding elements, thus blurring the boundaries between DNA-based cis-acting elements and trans-acting ncRNA genes.
Despite these apparent similarities, we nevertheless find distinct and distinctive epigenetic signatures that appear to delineate the expressed and unexpressed elements in our study. These observations in turn suggest that it may be possible to use epigenetic signatures in the identification of functional ncRNAs and CNSs. This may provide a major additional tool in noncoding element identification, particularly as it is difficult to confidently identify these without comparative genomic data. The comparative approach aids in building high-confidence annotations as sequence conservation is a strong signal of functional constraint. However, for species specific regulatory elements, which may be central to phenotypic change (King & Wilson 1975), epigenetic signatures may help in the identification of newly-emerged noncoding sequences.

We identified 153,475 conserved noncoding sequences that are common in the 10 primate species used in the study. We designated 59,870 sequences as ncRNA-gene-CEs and 93,605 sequences as CNSs. CNSs showed higher signal strength for H3K4Me1 and H3K27ac which are histone modifications related with active enhancer regions. Therefore we assume that many of these CNS elements might be functioning as active enhancers. It has been found that conserved elements are associated with active enhancer regions (Heintzman et al. 2009; Creyghton et al. 2010). Blum et al. (2012) found that 34% of their myoblast enhancers and 36% of myotube enhancers overlapped conserved noncoding sequences. It is currently known that even some of the ncRNA can also act as enhancer elements (Kim et al. 2010; Natoli and Andrau. 2012; Chen et al. 2017), but our primate common ncRNA-gene-CEs showed less association with active enhancer element modifications, signifying that these conserved elements may be related to other functions. Also we found that CNSs were enriched with histone modifications associated with regulation of development and active promoters. These findings with regards to CNSs are in congruence with already documented evidence (Inada et al. 2003; Bernat et al. 2006; Hettiarachchi et al. 2016).We found that primate common ncRNA-

gene-CEs had significantly higher levels of H4K20Me1 which is a histone modification associated with DNA repair than for developmental gene expression and regulation. There are several studies documenting that small and long ncRNA can play a role in DNA damage repair process (Sharma and Misteli. 2013; Michelini et al. 2017). Also D'Alessandro and Fagagna (2018) reported that especially long ncRNA plays a role in genome stability. With regards to the primate common CNSs and ncRNA-gene-CEs we notice an interleaved functional arena with regards to gene regulation, whereas DNA damage repair is concerned we see that predominantly ncRNA play a key role, thus implying a slight differentiation in the functions related to genome stability. Also we found that primate common ncRNA-gene-CEs have a higher tendency to be located in open chromatin regions compared to rest of the genome. This easily accessible open conformation agrees with facilitating the regulatory functions that may be related to ncRNA.

Both CNSs and ncRNA-gene-CEs showed lower frequency of derived alleles compared to random regions in human genome signifying purifying selection. It has been known for a while that conserved non-coding sequences are under purifying selection and are not merely mutational cold spots (Drake et al. 2006; Katzman et al. 2007; Sakuraba et al. 2008). The primate common ncRNA-gene-CEs showed a higher percentage level of lower allele frequency compared to CNSs, which in a way signifies that the ncRNA-gene-CEs are under a strongly constrained selective pressure. It is expected that majority of conserved regions to be under purifying selection given the assumption that they have been conserved across long divergence times due to their functional importance, but our result here shows that even among conserved regions some of them can be under a stronger selective pressure which could be due indispensable functions that cannot be compromised in the genome.

Yet another interesting result of this study is that CNSs and ncRNA-gene-CEs have gone through a phase of accelerated evolution in several different stages during evolution. The human-gorilla common ancestor sequences showed 4 times faster evolutionary rate compared to human-gorilla-orangutan common ancestor for both ncRNA-gene-CEs and CNSs. This result suggests that certain mutational changes occurred at a faster rate in the human-gorilla common ancestor sequences to facilitate the functional requirements of the species after divergence. Also we observed that Hominoidae common ancestor showed accelerated evolution in CNSs after diverging from Hominoidae-old world monkey common ancestor. Also old world monkeys show about 2 fold acceleration after diverging from Hominoidae-old world monkey common ancestor. Going a step further from widely accepted belief that conserved regions are under constant purifying selection it is important to understand that some conserved regions change faster prior to getting stabilized and go through purifying selection to maintain what was achieved by accelerated

mutation events. This scenario is possible where species after divergence adopt new set of physiological or morphological features that cannot be regulated by already existing functional elements in a genome. Particularly Homininae also referred to as "African great apes" consisting human, chimpanzee and gorilla did particularly experience an increase in brain size and bipedalism (Lovejoy Co. 1980). Without enough evidence it is hard to point that the accelerated evolution in conserved regions might be responsible for special characteristics of this lineage, but it has been found that certain CNSs are related to neuronal adhesion of human and chimpanzee (Prabakar et al. 2006).

It is known through experimental evidence that CNSs, actually do regulate the closest gene in most scenarios (Sumiyama et al. 2002; Bhatia et al. 2014). Functional classification for the ncRNA-gene-CEs and CNSs showed that both groups are enriched with genes involved in regulating transcription and neuronal functions. Adopting the most widely accepted classical understanding this implies that ncRNA-gene-CEs and CNSs both might be responsible for similar functions. A recent study revealed that especially long ncRNA (lncRNA) are associated with development of central nervous system and neurodegenerative disorders (Wei et al. 2018). Despite the previous belief that lncRNA does not have a function and is a result of mistakes during transcription can now be challenged with computational and also some experimental evidence. This result further clarifies that CNSs that were identified as ncRNA-gene-CEs in this study might actually have the potential to be related to gene regulation prior any experimental evidence.

The GO for CNSs and ncRNA-gene-CEs showed evidence that likely target gene of these regions are enriched in brain and nervous system related tissues especially Amygdala, brain meninx, occipital lobe and spinal cord (fetal and adult both). Several studies document that CNSs are related to brain and nervous system tissues (Meyer et al. 2017). But our finding shows that ncRNA-gene-CEs related genes can also be related to nervous system development rather than house-keeping functions.

We found that 60% of the primate common elements have solely originated in primate common ancestor and could be accounting for necessary functional requirements in the primate lineage. One interesting feature is that the newly originated elements also predominantly cluster around genes related to central nervous system. This indirectly implies that further adoption of conserved elements was required for brain function or re-wiring of molecular pathways related to brain function in the primate lineage. One significant feature that should be noted is, there has been a gradual expansion in the neocortex volume in higher primates compared to primitive primates such as insectivores and prosimians. Also further expansion in the primate lineage leading to, great apes having the largest neocortex (Stephen and Andy. 1970) in the animal kingdom. The neocortex is known as

the centre for higher brain functions, such as decision making, reasoning, language and consciousness. The expansion of this region and functions related to it in higher level primates such as great apes may have required new regulatory mechanisms thus newly originated regulatory elements. Even though we cannot directly narrow down all the elements we have identified in this study are related to neocortex functions, the tissue expression analyses imply that the conserved elements here are predominantly related to regions of the brain.

Also hominoidae specific sequences in this study being related to brain activity further solidify the reasoning above. Nevertheless it is important to note that mammalian common sequences showed much higher functional fold enrichment for central nervous system related activity compared to primate common and hominoidae specific elements. Could this mean that basic building blocks for proper functionality and structure of central nervous system occurred way before primates? This is something worth looking at in more detail.

Our findings in this study have shed light upon primate common, mammalian common and hominoidae specific CNSs and ncRNA-gene-CEs from numerous functional perspectives and have provided evidence that both groups are important for proper functionality of the genomes while signifying ncRNA-gene conservation may also be indispensable to genomes than we give credit for.

References

ALTSCHUL Stephen F., MADDEN Thomas L., SCHÄFFER Alejandro A., ZHANG Jinghui, ZHANG Zheng, MILLER Webb, and LIPMAN David. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, pp. 3389-3402.

ANDERSSON Robin, …, FORREST Alistair R. R., CARNINCI Piero, REHLI Michael, and SANDELIN Albin (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, vol. 507, pp. 455-461.

AWAN Hassaan Mehboob, SHAH Abdullah, RASHID Farooq, and SHAN Ge (2017) Primate-specific long non-coding RNAs and microRNAs. *Genomics, Proteomics & Bioinformatics*, vol. 15, pp. 187-195.

BABARINDE Isaac Adeyemi (2021) Conserved noncoding sequences: Evolving puzzles. *iDarwin*, vol. 1, pp. 3-36.

BAO Jianqiang, WU Jingwen, SCHUSTER Andrew S., HENNIG Grant W., and YAN Wei (2013) Expression profiling reveals developmentally regulated lncRNA repertoire in the mouse male germline. *Biology of Reproduction*, vol. 89, Article 107.

BARAKAT Tahsin Stefan, HALBRITTER Florian, ZHANG Man, RENDEIRO André F., PERENTHALER Elina, BOCK Christoph, and CHAMBERS Ian (2018) Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell*, vol. 23, pp. 276-288.

BERNAT John A., aCRAWFORD Gregory E., OGURTSOV Aleksey Y., COLLINS Francis S., GINSBURG David, and KONDRASHOV Alexy S. (2015) Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Human Molecular Genetics*, vol. 15, pp. 2098-2105.

BHATIA Shipra, MONAHAN Jack, RAVI Vydianathan, GAUTIER Phillippe, MURDOCH Emma, BRENNER Sydney, van HEYNINGEN Veronica, VENKATESH Byrappa, and KLEINJAN DIRK A. (2014) A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. *Developmental Biology,* vol. 387, pp.214-228.

BLUM Roy, VETHANTHAM Vasupradha, BOWMAN Christopher, RUDNICKI Michael, and DYNLACHT Brian D. (2012) Genome-wide identification of enhancers in skeletal muscle: the role of MyoD1. *Genes & Development,* vol. 26, pp. 2763-2779.

CHEN Grace Y., SATPATHY Ansuman T., and CHANG Howard Y. (2017) Gene regulation in the immune system by long noncoding RNAs. *Nature Immunology*, vol. 18, pp. 962-972.

CLARKE Shoa L., VANDERMEER Julia E., WENGER Aaron M., SCHAAR Bruce T., AHITUV Nadav, and BEJERANO Gill (2012) Human developmental enhancers conserved between deuterostomes and protostomes. *PLoS Genetics*, vol. 8, e1002852.

CREYGTON Menno P., …, and JAENISCH Rudolf (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*, vol. 107, pp.21931-21936.

D'ALESSANDRO Giuseppina and di FAGAGNA Fabrizio d'Adda (2018) Long non-coding RNAs in the control of genome stability and cancer phenotypes. *Non-coding RNA Investigation,* vol. 2, article 13.

DRAKE Jared A., …, DERMITZAKIS Emmanuel T., and HIRSCHHORN Joel N. (2005) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature Genetics*, vol. 38, pp. 223-227.

ELGAR Greg (2009) Pan-vertebrate conserved non-coding sequences associated with developmental regulation. ***Briefings in Functional Genomics***, vol. 8, pp. 256-265.

The ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, vol. 489, pp. 57–74.

GE Xin Quang and LIN Haifan (2014) Noncoding RNAs in the regulation of DNA replication. *Trends in Biochemical Sciences*, vol. 39, pp. 341-343.

GLAZKO Galina V. and NEI Masatoshi (2003) Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution*, vol. 20, pp. 424-434.

GOODMAN Morris, PORTER Calvin A.., CZELUSNIAK John, PAGE Scott L., SCHNEIDER Horacio., SHOSHANI Jeheskel, GUNNELL Gregg, and GROVES Colin P. (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. ***Molecular Phylogenetics and Evolution***, vol. 9, pp. 585-598.

GRAUR Dan, ZHENG Yichen, PRICE Nicholas, AZEVEDO Ricardo B. R., ZUFALL Rebecca A., and ELHAIK Ran (2013) On the immortality of television sets: "function" in the human genome according to evolution-free gospel of ENCODE. *Genome Biology and Evolution,* vol 5, pp. 578–590.

HAWLEY Ben R., LU Wei-Ting, WILCZYNSKA Ania, and BUSHELL Martin (2017) The emerging role of RNAs in DNA damage repair. *Cell Death and Differentiation,* Vol. 24, pp. 580-587.

HEINTZMAN Nathaniel, …, and REN Bing (2009) Histone modifications at human enhancers reflect global cell type-specific gene expression. *Nature*, Vol. 459, pp. 108-112.

HETTIARACHCHI Nilmini, KRYUKOV Kirill, SUMIYAMA Kenta, and SAITOU Naruya (2014) Lineage-specific conserved noncoding sequences of plant genomes: their possible role in nucleosome positioning. *Genome Biology and Evolution,* vol 6, pp. 2527–2542.

HETTIARACHCHI Nilmini and SAITOU Naruya (2016) GC content heterogeneity transition of conserved noncoding sequences occurred at the emergence of vertebrates. *Genome Biology and Evolution*, vol. 8, pp. 3377–3392.

INADA Dan Choffnes, BASHIR Ali, LEE Chunghau, THOMAS Brian C., KO Cynthia, GOFF Stephen A, and FREELING Michael (2003) Conserved noncoding sequences in the grasses. *Genome Research*, vol. 13, pp. 2030-2041.

KATZMAN Sol, KERN Andrew W., BEJERANO Gill, FEWELL Ginger, FULTON Lucinda, WILSON Richard K., SALAMA Sofie R., and HAUSSLER David (2007) Human genome ultra-conserved elements are ultra-selected. *Science*, vol. 317, p. 915.

KIM Tae-Kyung, …, and GREENBERG Michael E. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, Vol. 465, pp. 182-187.

KIM Tae-Kyung and SHIEKHATTAR Ramin (2015) Architectural and functional commonalities between enhancers and promoters. *Cell*, vol. 162, pp. 948-959.

KLEIN Eric and ASSOIAN Richard K. (2008) Transcriptional regulation of the cyclin D1 gene at a glance. *Journal of Cell Science*, vol. 121, pp. 3853-3857.

KOENECKE Nina, JOHNSTON Jeff, HE Qiye, MEIER Samuel, and ZEITLINGER Julia (2017) Drosophila poised enhancers are generated during tissue patterning with the help of repression. *Genome Research*, vol. 27, pp. 64-74.

KRITSAS Konstantinos, WUEST Samuel E., HUPALO Daniel, KERN Andrew D., WICKER Thomas, and GROSSNIKLAUS Ueli (2012) Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Research*, vol. 22, pp. 2455-2466.

LEE Alison P., KERK Sze Yen, TAN Yue Ying, BRENNER Sydney, and VENKATESH Byrappa (2011) Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Molecular Biology and Evolution*, vol. 28, pp. 1205-1215.

LI Ren and CUI Xia (2018) Genome-wide mapping of DNaseI hypersensitive sites in tomato. Pp. 367-379 in YAMAGUCHI N. Ed., "Plant Transcription Factors –*Methods and Protocols*", Springer.

MADRIGAL Pedro and KRAJEWSKI Paweł (2012) Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Frontiers in Genetics*, vol. 3, article 230.

MATSUNAMI Masatoshi and SAITOU Naruya (2013) Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. *Genome Biology and Evolution,* vol. 5, pp. 140-150.

MELAMED Philippa, YOSEFZON Yahav, RUDNISKY Sergei, and PNUELI Lilach (2016) Transcriptional enhancers: Transcription, function and flexibility. *Transcription*, vol. 7, pp. 26-31.

MEYER Kyle A., MARQUES-BONET Tomas, and SESTAN Nenad (2017) Differential gene expression in the human brain is associated with conserved, but not accelerated, noncoding sequences. *Molecular Biology and Evolution*, vol. 34, pp. 1217-1229.

MICHELINI Flavia, …, and di FAGAGNA Fabrizio d'Adda (2017) Damage-induced lncRNAs control the DNA damage response through interaction with DDRNAs at individual double-strand breaks. *Nature Cell Biology,* vol. 19, pp.1400-1411.

NATOLI Gioacchino and ANDRAU Jean-Christophe (2012) Noncoding transcription at enhancers: general principles and functional models. *Annual Review of Genetics*, vol. 46, pp.1-19.

NELSON Andrew C. and WARDLE Fiona (2013) Conserved non-coding elements and cis regulation: actions speak louder than words. *Development*, vol. 140, pp. 1385-1395.

POLYCHRONOPOULOS Dimitris, SELLIS Diamantis, and ALMIRANTIS Yannis (2014) Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics. *PLoS One*, vol. 9, e95437.

PRABHAKAR Shyam, NOONAN James P., PÄÄBO Svante, and RUBIN Edward M. (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science*, vol. 314, pp. 786-796.

SABER Morteza Mahmoudi, BABARINDE Isaac Adeyemi, HETTIARACHCHI Nilmini, and SAITOU Naruya (2016) Emergence and evolution of Hominidae-specific coding and noncoding genomic sequences. *Genome Biology and Evolution,* vol. 8, pp. 2076–2092.

SAITOU Naruya and NEI Masatoshi (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* vol. 4, pp. 406-425.

SAKURABA Yoshiyuki, …, and Gondo Yoichi (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. *Mammalian Genome*, vol. 19, pp. 703-712.

SANDELIN Albin, BAILEY Peter, BRUCE Sara, ENGSTRÖM Pär G., KLOS Joanna M., WASSERMAN Wyeth W., ERICSON Johan, and LENHARD Boris (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, vol. 5, article 99.

SHARMA Vivek and MISTELI Tom (2013) Non-coding RNAs in DNA damage and repair. *FEBS Letters*, vol. 587, pp. 1832-1839.

SHEN Yong and others (2018) Identification of a novel enhancer/chromatin opening element associated with high-level γ-globin gene expression. *Molecular and Cellular Biology*, vol. 38, e00197-18.

SONG Lingyun, …, and BUNGERT Jörg (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, vol. 21, pp. 1757-1767.

SUMIYAMA Kenta, IRVINE Steven Q., STOCK David W., WEISS Kenneth M., KAWASAKI Kazuhiko, SHIMIZU Nobuyuki, SHASHIKANT Cooduvalli, MILLER Webb, and RUDDLE Frank H. (2002) Genomic structure and functional control of the Dlx3-7 bigene cluster. *Proceedings of the United States of Americas of the National Academy of Sciences*, vol. 99(2), pp. 780-785.

TAKAHASHI Mahoko and SAITOU Naruya (2012) Identification and characterization of lineage-specific highly conserved noncoding sequences in mammalian genomes. *Genome Biology and Evolution*, vol. 4, pp. 641-657.

TAMURA Koichiro, PETERSON Daniel, PETERSON Nicholas, STECHER Glen, NEI Masatoshi, and KUMAR Sudhir (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution,* vol. 28, pp. 2731-2739.

THOMAS Paul, CAMPBELL Michael J., KEJARIWAL Anish, MI Huaiyu, KARLAK Brian, DAVERMAN Robin, DIEMER Karen, MURUGANUJAN Anushya, and NARECHANIA Apurva (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research,* vol. 13, pp. 2129-2141.

TIPPENS Nathaniel D., LIANG Jin, LEUNG Alden King-Yung, WIERBOWSKI Shayne D., OZER Abdullah, BOOTH James G., LIS John T., and YU Haiyuan (2020) Transcription imparts architecture, function and logic to enhancer units. *Nature Genetics*, vol. 52, pp. 1067-1075.

VAVOURI Tanya, WALTER Klaudia, GILKS Walter R., LEHNER Ben, and ELGAR Greg (2007) Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biology*, vol. 8, article R15.

WANG Xianting, ARAI Shigeki, SONG Xiaoyuan, REICHART Donna, DU Kun, PASCAL Gabriel, TEMPST Paul, ROSENFELS Michael G., GLASS Christopher K., and KUROKAWA Riki (2008) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature*, vol. 454, pp. 126-130.

WEI Xiaoyan, MA Tengfei, CHENG Yifeng, HUANG Cathay C., WANG Xuehua, LU Jiayi, and WANG Jun (2017) Dopamine D1 or D2 receptor-expressing neurons in the central nervous system. *Addiction Biology*, vol. 2, pp. 569-584.

WOOLFE Adam, …, and ELGAR Greg (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biology*, vol. 3, e7.

YATES Andrew, …, and FLICEK Paul (2020) Ensembl 2020. *Nucleic Acids Research*, vol. 48, pp. D682-D688.

ZHANG Ali, ZHOU Nanjing, HUANG Jianquo, LIU Qian, FUKUDA Koji, MA Ding, BAI Cunxue, WATABE Kounosuke, and MO Yin-Yuan (2013) The human long non-coding RNA-RoR is a p53 repressor in response to DNA damage. *Cell Research*, vol. 23, pp. 340–350.