Review Article
iDarwin Volume 1, pp. 3-36
Published on February 12, AS 0021 (2021 AD)

Conserved noncoding sequences: Evolving puzzles

Isaac Adeyemi BABARINDE
Department of Biological Sciences
Southern University of Science and Technology
1088 Xueyuan Avenue, Shenzhen, P. R. China, 518055
Email: babarindeia@sustech.edu.cn; iababarinde@gmail.com

Abstract Majority of the molecular evolutionary studies in the pre-genomic era were focused on the protein-coding parts which account for less than 2% of the human genome and are mostly evolutionarily conserved. However, the abundance and the properties of noncoding functional sequences were a puzzle. With the advent of sequencing technologies and the release of vertebrate genome sequences, it became obvious that certain noncoding parts of the genomes tend to be evolutionarily conserved over hundreds of millions of years. These regions of the genome with specific properties, termed conserved noncoding sequences (CNSs), are usually computationally discovered. Unlike genes with specific transcription units, the definitions of CNSs are vague and thresholds are often set to delimit the regions under conservations and other regions that are neutrally evolving. Thought to be regulatory elements, many of these sequences have been shown to be transcribed and deletions have occasionally had no observable effects. The puzzles surrounding CNSs are evolving. In this review, I discuss the various definitions of CNSs and the computational approaches available for discovering them with the accompanied puzzles. The reported features of the CNSs and the reported functionalities are discussed. I conclude this review by discussing the current puzzles in the studies of CNSs.

Keywords: Conserved noncoding sequences; computational search; junk; regulatory elements; gene deserts

Introduction

Numerous studies reported in the pre-genomic eras have suggested that genomic functional sequences might reside outside the protein-coding regions (Dawson et al. 1995; Hezroni et al. 2017; King and Wilson 1975; Lou et al. 1995; Oeltjen et al. 1997a). If the sequences are functionally important, they would not evolve neutrally (Kimura 1983). Rather, obvious signatures of evolutionary constraints would be expected (Kimura and Ohta 1974). The regions outside the protein coding parts of the genomes that show detectable signatures of evolutionary constraints are referred to as conserved noncoding sequences, abbreviated as CNS (Hardison 2000; Loots et al. 2000) or conserved noncoding elements abbreviated as CNEs (McEwen et al. 2006; Vavouri et al. 2007). Some of the other variants of the name (**Table 1**) include ultraconserved elements or UCEs (Bejerano et al. 2004; Siepel et al. 2005), long conserved noncoding sequences or LCNSs (Sakuraba et al. 2008), highly conserved noncoding regions or HCNRs (De La Calle-Mustienes et al. 2005). Each of the names is often defined by the percent identity and/or length thresholds (**Table 1**). For some studies (Lowe et al. 2011; Siepel et al. 2005), there is no specified constant thresholds. Rather, definition of constraint is based on models that are able to distinguish functional regions from neutrally evolving ones (**Table 1**). In cases with variable thresholds, the studies used specific models to delimit conserved regions from neutrally evolving regions. Such models might have no rigid and constant identity and/or length thresholds. For Siepel et al. (2005), sliding window of 5 bp was used.

Unlike mRNA, lncRNAs or transcribed enhancers, CNSs are not necessarily transcribed but are rather computationally discovered. Using various computational strategies, conserved noncoding regions have been reported in numerous taxonomic groups such as mammals (Babarinde andSaitou 2013, 2016; Bejerano et al. 2004; Loots et al. 2000; Mahmoudi Saber and Saitou 2017; Saber et al. 2016), insects (Brody et al. 2020; Glazov et al. 2005), plants (Hettiarachchi et al. 2014; Van de Velde et al. 2016) and other species (Siepel et al. 2005; Vavouri et al. 2007).

Table 1: Representative names for the conserved noncoding part of the genomes

Name	Identity constraint (%)	Length threshold (bp)	Representative reference
Ultraconserved element (UCE)	100 (human, mouse and rat)	200	(Bejerano et al. 2004; Glazov et al. 2005)
Ultraconserved noncoding element (uCNE)	95 (human and chicken)	200	(Dimitrieva and Bucher 2013)
Conserved noncoding sequence (CNS)	70 (human, mouse)	100	(Hardison 2000; Loots et al. 2000)
	variable	100	(Babarinde and Saitou 2013, 2016)
Conserved noncoding element (CNE)	variable	variable	(Siepel et al. 2005)
Conserved nonexonic elements (CNEEs)	variable	variable	(Lowe et al. 2011)
Long conserved noncoding sequences (LCNS)	95 (mouse, human)	500	(Janes et al. 2011; Sakuraba et al. 2008)
Highly conserved noncoding sequence (HCNS)	variable	100	(Saber et al. 2016; Takahashi and Saitou 2012)
Highly conserved noncoding regions (HCNR)	75 (in vertebrates)	100	(De La Calle-Mustienes et al. 2005)
Noncoding highly conserved elements (noncoding HCE)	variable	5	(Siepel et al. 2005)

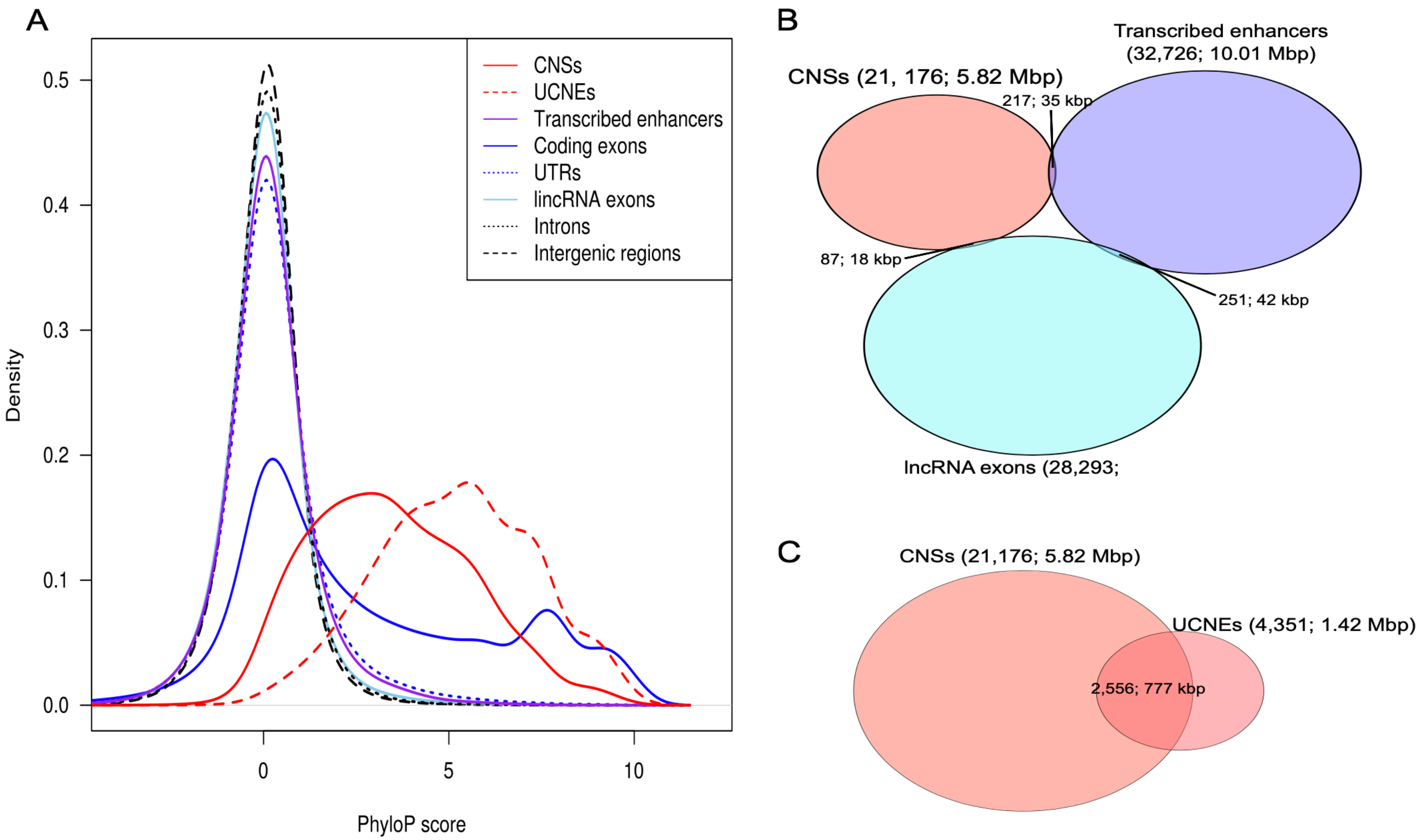
Definitions of CNSs

By definition, any region of the genome that is under evolutionary constraint but does not code for protein is defined as CNS (Hardison 2000; Kellis et al. 2014). Although the definition is grammatically simple, the actual identification of the CNSs in a genome is not such a simple task and different approaches have been made to get the task done. The first task is the definition of “coding”. In the pre-genomic era, “coding” was mostly used to represent protein-coding part of the genome (Poon et al. 1978). However, it is clear that not all the parts of protein-coding genes code for amino acids. Therefore, intron should be considered as noncoding (Hardison 2000). Furthermore, not all exons code for amino acid. Therefore, UTR might be considered noncoding, if the definition of “coding” is protein-coding part of the genome. With increasing sequencing power, many more noncoding RNAs are now being discovered (Iyer et al. 2015; Zhao et al. 2016). In fact, majority of human genome has been reported to be “biochemically active”, an expression used to describe transcribed regions or regions bound or modified by proteins or other elements (ENCODE Project Consortium 2012). Although these transcripts exist, they mostly do not code for amino acids (Iyer et al. 2015; Zhao et al. 2016), and so can be referred to as noncoding. For simplicity, this review defines “noncoding” as “genomic regions that do not code for protein”. Therefore, “noncoding” can include intergenic regions, intronic regions and untranslated regions (UTRs), irrespective of transcription. This represents a proportion of the genomic sequences that are not translated. Of course, each part of the genome is operating under different intensities of evolutionary forces (**Figure 1A**). CNSs and UCNEs show conservation level higher than those of the UTR, intronic and intergenic regions. Comparison of three noncoding sequences shown in **Figure 1B** shows very little overlaps among transcribed enhancers, lincRNAs and CNSs.

The second aspect of the definition of CNS implies that the region must be under evolutionary constraints stronger than those operating on neutrally evolving regions (Babarinde and Saitou 2013). Then the next task is defining neutrally evolving regions. The classic definition of neutrally evolving regions includes intron and intergenic regions (Kimura 1983). These are the exact regions where CNSs are located. And the presence of transposable elements in these regions can make the expectation under neutral evolution much more complicated. A number of studies (Babarinde and Saitou 2013; Hettiarachchi and Saitou 2016; Saber et al. 2016; Takahashi and Saitou 2012) masked repeats in the genome, effectively shifting the focus away from transposable elements, which might be functionally important (Hutchins and Pei 2015; Kim et al. 2012; Ramsay et al. 2017).

Figure 1. Functional constraint and genomic overlap of noncoding sequences in human genome. A. The PhyloP scores were obtained for CNSs (Babarinde and Saitou 2016), UCNEs (Dimitrieva and Bucher 2013), transcribed enhancers (Andersson et al. 2014), coding exons, UTRs, lincRNAs, introns and intergenic regions. The vertical axis represents the proportion of the distribution. **B**. CNSs, transcribed enhancers and lincRNA exons have small overlaps. **C**. Different criteria give imperfect but significant overlaps. Assuming 51.4% unique genomic regions (Hutchins

and Pei 2015), this much overlap is less likely to be obtained by chance (binomial p value $< 2.2 \times 10^{-308}$). Unless otherwise stated, the genomic coordinates were obtained from version 96 of Ensembl database. For **B** and **C**, the first values were the numbers of sequences while the second values were the numbers of base pairs covered.



In identifying genomic regions under evolutionary constraints, statistical tests are often conducted to separate regions of interest from neutrally evolving regions. Two parameters are often checked. The first parameter is sequence conservation. These are sometimes estimated as percent identity (Babarinde and Saitou 2013; Hettiarachchi and Saitou 2016; Kellis et al. 2014; Takahashi and Saitou 2012). However, conservation score per site can also be calculated by phyloP (Pollard et al. 2010), PhastCons score (Siepel et al. 2005), GERP (Pollard et al. 2010) or SiPhy (Garber et al. 2009). The second parameter is length. Evolutionary strength is often estimated as sequence constraint over certain length. For example, UCE (Bejerano et al. 2004) requires 100% sequence identity over >=200bp between human and mouse. LCNSs (Sakuraba et al. 2008) have at least 95% identity over >=500bp. The sequence identity thresholds are often arbitrary. That is why Babarinde and Saitou (2013) proposed thresholds that were informed by protein-coding gene properties. Examples of the definitions of CNSs are presented in **Table 1**. Various definitions of criteria make overlap between regions reported from different studies not perfect, but still significantly more than random expectations (binomial p value $< 2.2 \times 10^{-308}$, **Figure 1C**). The binomial test was conducted to find the probability of having at least that many nucleotide overlaps under the assumption that the sequences come from the 51.4% unique part of human genome (Hutchins and Pei 2015).

Digging into the junk: Computational tools for the discovery of CNSs

The search for CNSs starts with the acquisition of genome sequence and annotation data. The search can be restricted to certain genomic region (Bulger et al. 1999; Hardison et al. 1997; Loots et al. 2000; Oeltjen et al. 1997b; Poon et al. 1978) based on gene neighborhood (Bush and Lahn 2005; Jareborg et al. 1999) or genome wide (McEwen et al. 2006). Many studies (Babarinde and Saitou 2013; Hettiarachchi and Saitou 2016; Saber et al. 2016; Takahashi and Saitou 2012) use repeat-masked genome sequences as the repeat sequences are usually complex to analyze. The annotation data are usually used to mask the coding parts of the genome such that only the noncoding part of the genome is searched (Babarinde and Saitou 2013). The search for conserved regions can be initiated from each nucleotide position or genomic regions. Genomic position scores such as GERP (Pollard et al. 2010) and phyloP (Pollard et al. 2010) can be used to identify the conservation level of each genomic position. Other scores such as phastCons (Pollard et al. 2010) and SiPhy (Garber et al. 2009) can also be used, and are also available for short genomic regions. The scores for each genomic position can be easily obtained from UCSC database (Haeussler et al. 2019) or other sources like PhastWeb (Ramani et al. 2019). Using sliding window analyses, regions that are evolving significantly below neutral expectations can be retrieved. This approach was employed by some previous studies (e.g. Siepel et al. 2005).

Another approach is to download the alignment files from databases like UCSC (Haeussler et al. 2019). Using sliding window analyses, regions that satisfy specified percent identity and length thresholds are then extracted as done by Takahashi and Saitou (2012). Another common way is to use pairwise searches with local alignment tools such as BLAST (Altschul et al. 1997) or BLAT (Kent 2002) to identify regions conserved between two species at specified percent identity and length thresholds. In this approach, the species divergence should be considered to effectively delimit neutrally evolving sequences from functional ones. If species are phylogenetically too close, more stringent conditions should be applied (Saber et al. 2016; Takahashi and Saitou 2012). However, less stringent conditions can be applied if the species are significantly divergent (McEwen et al. 2006; Siepel et al. 2005; Van Hellemont et al. 2005; Woolfe et al. 2004). This consideration also applies to overall evolutionary rates of the species being considered (Babarinde and Saitou 2013; Takahashi and Saitou 2012) as the heterogeneity of evolutionary rates across species has been established (Babarinde and Saitou 2020). Using a reference species, regions that are conserved across multiple species can then be retrieved. This approach has been employed in multiple studies (Babarinde and Saitou 2016; Hettiarachchi and Saitou 2016; Mahmoudi Saber and Saitou 2017; Saber et al. 2016). One major challenge with pairwise search is the false discovery rates of short regions (Kamoun et al. 2013; Lai et al. 2017). To circumvent this issue, STAG-CNS (Lai et al. 2017) was published with the ability to identify CNSs as short as 9bp. CNEFinder (Ayad et al. 2018) offers the users the opportunity to input several thresholds to retrieve CNSs. Attempts have been made not only to discover CNSs, but also to prepare accessible databases of the discovered CNSs (Dousse et al. 2016; Inoue and Saitou 2020; Persampieri et al. 2008; Woolfe et al. 2007).

Identified features of CNSs

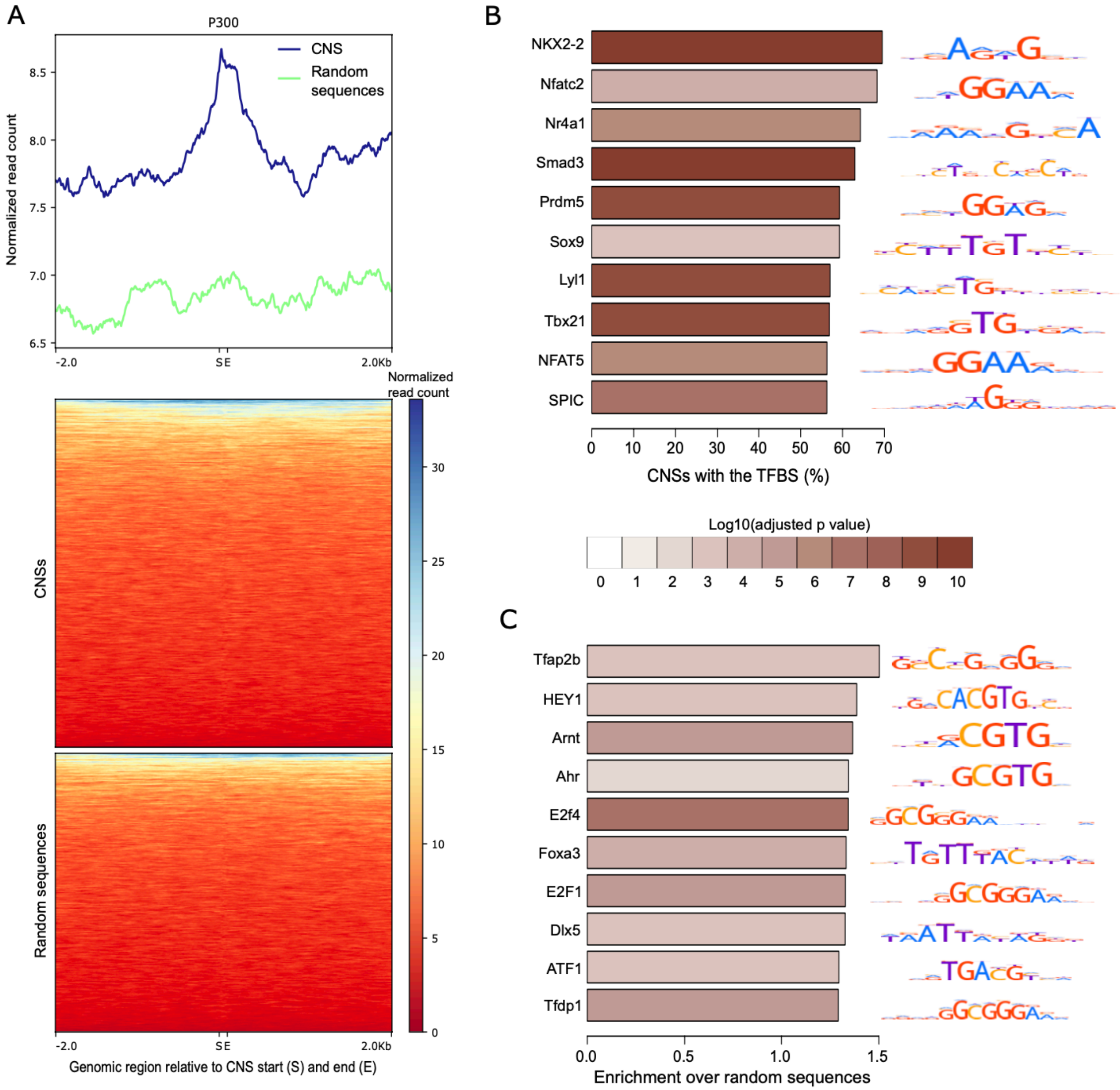
CNSs are known to have genomic distribution that are different from random expectation. First, they tend to occur in clusters close to certain types of genes (Matsunami et al. 2010; Mignone et al. 2008; Takahashi and Saitou 2012). Their closeness to genes involved in processes such as developmental, transcription and nervous systems (Babarinde and Saitou 2013) suggests that they might regulate the expression of such genes. Compared to lincRNAs and random intergenic sequences, CNSs found in intergenic regions tend to be located farther from transcription start sites or TSSs (Babarinde and Saitou 2016). This suggests that distance might not be a barrier for the functionality of CNSs.

In fact, Visel et al. (2009) reported that ChIP-Seq can accurately predict regulatory elements. **Figure 2A** shows that the enrichment pattern of p300 gene which marks enhancer activity (Raisner et al. 2018; Visel et al. 2009) is different between intergenic CNSs and randomly selected genomic sequences. Specifically, CNSs tend to be located in the peaks of p300 ChIP-Seq in mouse embryonic forebrain (Visel et al. 2009) suggesting the CNS activities at this stage. Two things stand out in **Figure 2A**. First, average ChIP-Seq outside CNSs is higher than random averages, suggesting that CNSs and ChIP-Seq do not share the same boundary.

Second, only a subset of CNSs are enriched in mouse embryonic forebrain p300 binding, highlighting CNS tissue specificity as previously reported (Babarinde and Saitou 2016). This p300 enrichment and enrichment of genes involved in transcription suggest that CNSs might be enriched in certain transcription factor binding motifs (TFBM). Indeed, the overrepresentation of certain motifs has been reported (Babarinde and Saitou 2016). For example, TFBM analyses using AME (McLeay and Bailey 2010) in MEME Suite (Bailey et al. 2009) with default settings returned 217 transcription factors from human and mouse full HOCOMOCO database (Kulakovskiy et al. 2018). The results indicated that about 70% of the intergenic CNSs from Babarinde and Saitou (2016) have NKX2-2 binding motifs. The top 10 most represented transcription factors are shown in **Figure 2B**. It is important to note that the number of CNSs with TFBMs for each transcription factor is significantly higher than the corresponding number in randomly selected genomic sequences (adjusted p value of Fisher exact test is shown in **Figure 2B**). For each TFBM, the enrichment was calculated as the number of CNSs with the motif divided by the number of randomly selected genomic sequences with the motif. The top 10 enriched TFBMs which include Tfap2b, HEY and Arnt are shown in **Figure 2C**.

Figure 2. CNS binding motif enrichments

A. CNSs are found at the peak of P300 bound regions. The short reads sequences, obtained from embryonic forebrain (Visel et al. 2009) were aligned with BWA (Li and Durbin 2010). Deeptools (Ramírez et al. 2016) was then used to plot the profile and heatmaps of intergenic CNSs with 2kbp flanking regions. Similar analyses were also conducted for random sequences of the similar number and lengths. The top panel represents the average normalized read count over 20 bp windows. The horizontal axis in the top and bottom panels are the positions, relative to CNSs. Each row in the lower panel represents each genomic region. **B**. The top 10 transcription factors with the most represented TFBM in CNSs are presented. Mouse intergenic CNSs and randomly selected intergenic sequences were analyzed using AME package (McLeay and Bailey 2010) in MEME Suite (Bailey et al. 2009). Human and mouse (HOCOMOCO v11 FULL) motif database (Kulakovskiy et al. 2018) was used. For other parameters, default values were used. **C**. The enrichment scores of the top 10 TFBMs. Enrichments were computed by dividing the number of motif-containing CNSs by the number of motif-containing random sequences. The colors of the bars in B and C correspond to the adjusted Fisher exact test p values reported in AME.



Associating a CNS to the closest TSS, **Figure 3** shows the enrichment of 1,385 genes with at least four associated CNSs. As previously reported (Elgar and Vavouri 2008; Hettiarachchi and Saitou 2016; Ishibashi et al. 2012; Matsunami et al. 2010), there is higher enrichment in genes associated with development. The topmost associated gene ontology terms include head development, embryonic morphogenesis and others (**Figure 3**). These gene ontology terms are heavily connected in gene networks (**Figure 4**) suggesting interconnections in activities.

Figure 3. Gene ontology enrichment analysis of CNS-associated genes

CNSs tend to cluster around genes associated with development. CNSs in mouse genomes reported by Babarinde and Saitou (2016) were associated to the genes with the closest TSS. The top 1385 genes with at least four associated CNSs were then analyzed using Metascape (Zhou et al. 2019).

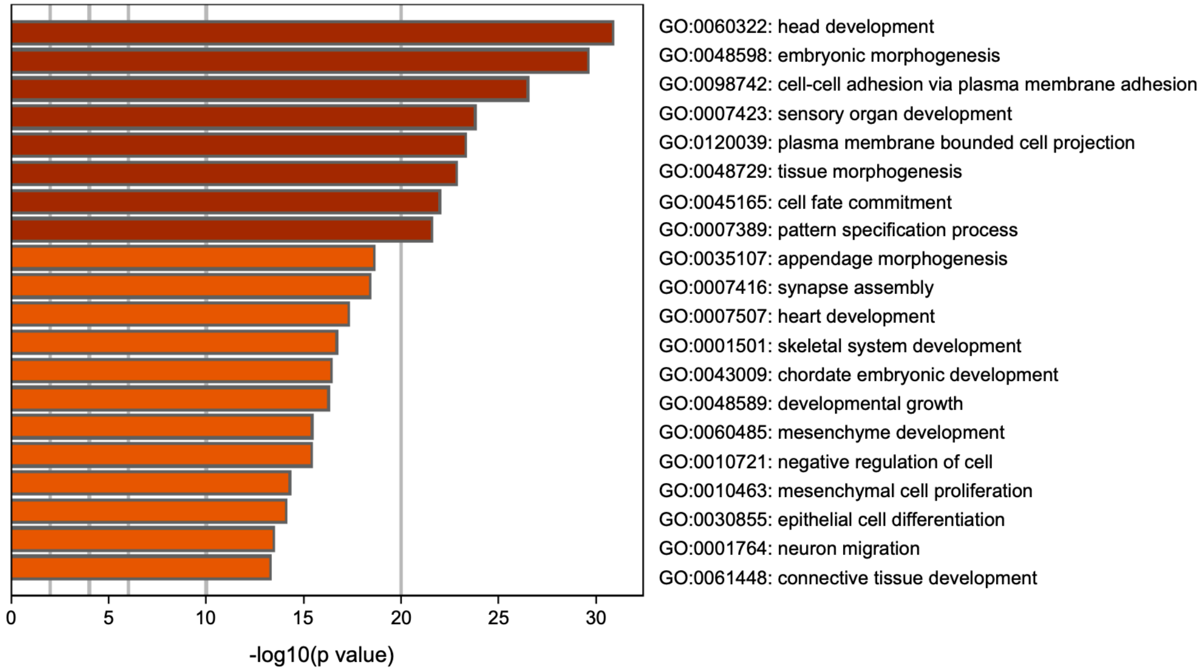
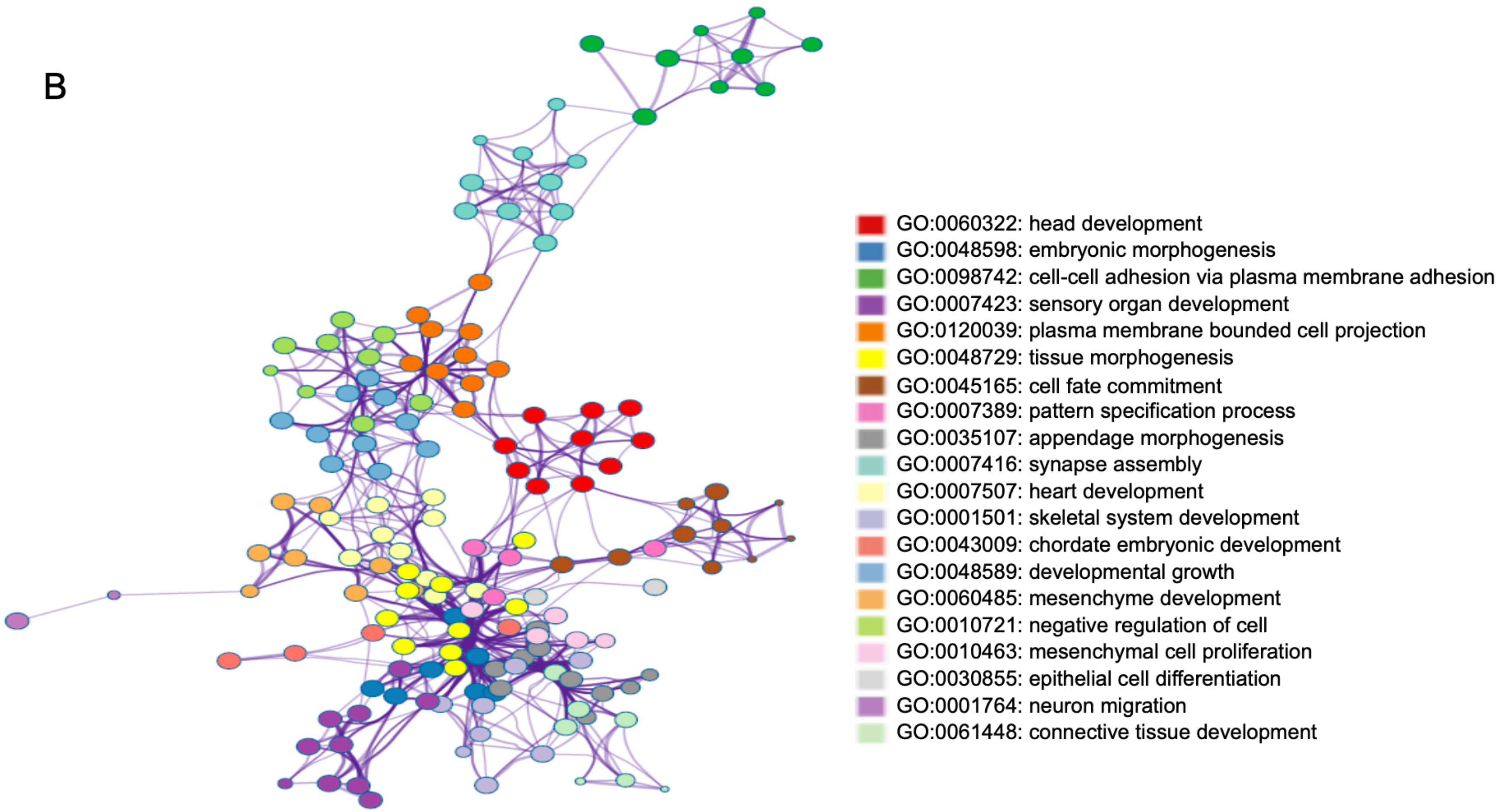


Figure 4. Network analysis of CNS-associated genes

The top 1000 CNS-enriched genes are heavily connected in functions. The network presented was produced with Metascape (Zhou et al. 2019).



Associated gene enrichment pattern suggests that CNS activities might be specific to certain stage and tissues. For example, enrichment of genes involved in development and nervous system implies that CNSs might be more active in embryonic brain (Dickel et al. 2018; Saber et al. 2016). Indeed, we previously found evidence of CNS activities in mouse embryonic brain (Babarinde and Saitou 2016). The evidence involved ChIP-Seq signal and the conservation in expression. Conversely, genes associated with defense and immunity and response to stimulus were found to be underrepresented in CNSs (Babarinde and Saitou 2016; Mahmoudi Saber and Saitou 2017). Further, genes expressed in testes or housekeeping genes with ubiquitous expressions were found to be underrepresented in CNSs (Babarinde and Saitou 2016). These observations reveal the specificity of CNS activities. Another nature of CNSs is the GC content heterogeneity (Babarinde and Saitou 2013; Hettiarachchi and Saitou 2016). We previously reported that tetrapod CNSs of different evolutionary ages, associated with recently acquired or more ancestral functions, tend to have different GC contents (Babarinde and Saitou 2013). Other studies (Hettiarachchi et al. 2014; Hettiarachchi and Saitou 2016) also showed that the heterogeneity exists also in non-tetrapod species. Interestingly, the GC contents of CNSs tend to be different from those of the flanking regions (Hettiarachchi and Saitou 2016). This GC content patterns have been associated with nucleosome occupancy (Dekker 2007; Hettiarachchi and Saitou 2016; Zhu et al. 2011).

Another interesting feature of CNSs is the conservation of CNS-TSS distance. The distance conservation was reported by (Babarinde and Saitou 2016). Comparing CNS-TSS distance in human and mouse, the study showed that genes associated with conserved CNS-TSS distance in human and mouse tend to have higher expression correlation, suggesting more stable expression across evolutionary timescales. The evolutionary analyses suggested that proper CNS function is dependent not only on the nucleotide sequences, but also on the genomic location of the CNSs. Later, Bagadia et al. (2019) further explored this possibility and found that indeed, evolutionary loss of genomic proximity of CNSs impacts expression dynamics during mammalian brain development. They reported that CNS-gene proximity interrupted by mechanisms such as chromosomal rearrangements could cause brain abnormality of germ line origin. These studies reveal another level of the functionality of CNSs.

Functionality of CNSs

It is known that different genomic regions have different evolutionary rates (Kimura 1983). Therefore, high cross-species sequence similarity found in genomes could be due to mutation cold spot or actual evolutionary constraint. However, derived allele frequency analyses showed that CNSs are under evolutionary constraint (Asthana et al. 2007; Drake et al. 2006; Ishibashi et al. 2012; J. Xie et al. 2018). Therefore, high sequence conservation in CNSs indicates functionality but does not necessarily indicates what type of functions they perform. However, since the discovery of abundant CNSs in vertebrate genomes, many studies have attributed CNSs to regulatory functions (Fishilevich et al. 2017; Frankel et al. 2010; Ishibashi et al. 2012; Osterwalder et al. 2018; Sumiyama et al. 2012; Visel et al. 2007). CNSs tend to cluster around specific genes (**Figures 3 and 4**), and there is an overrepresentation of certain TFBSs (**Figures 2B and C**). Furthermore, ChIP-Seq analyses show that many CNSs function as, or at least overlap, regulatory sequences (Visel et al. 2009). In a previous study (Babarinde and Saitou 2016), we demonstrated that CNSs have substantial signatures of regulatory functions and genes associated with more CNSs tend to have more stable expression patterns as indicated by expression correlations between mouse and human. Although many of these studies are computationally conducted, a good number of studies have experimentally validated regulatory functions of CNSs through *in vitro* enhancer assay. Furthermore, phenotypic effects of the genomic deletions of some CNSs have been reported. For example, a number of studies (Furniss et al. 2008; Lettice et al. 2002; Sagai et al. 2005) have attributed abnormal limb deformity and polydactyly to loss of certain CNSs. Vista (Visel et al. 2007), CONDOR (Woolfe et al. 2007) and Genehancer (Fishilevich et al. 2017) are databases of CNSs, including those with experimentally confirmed enhancer activities. CNSs have also been reported to have silencing functions (Hermann and Heckert 2005; Mahmoudi Saber and Saitou 2017). Consequently, many CNSs are believed to be involved in regulating proximal gene expression (Babarinde and Saitou 2016; Nelson et al. 2013).

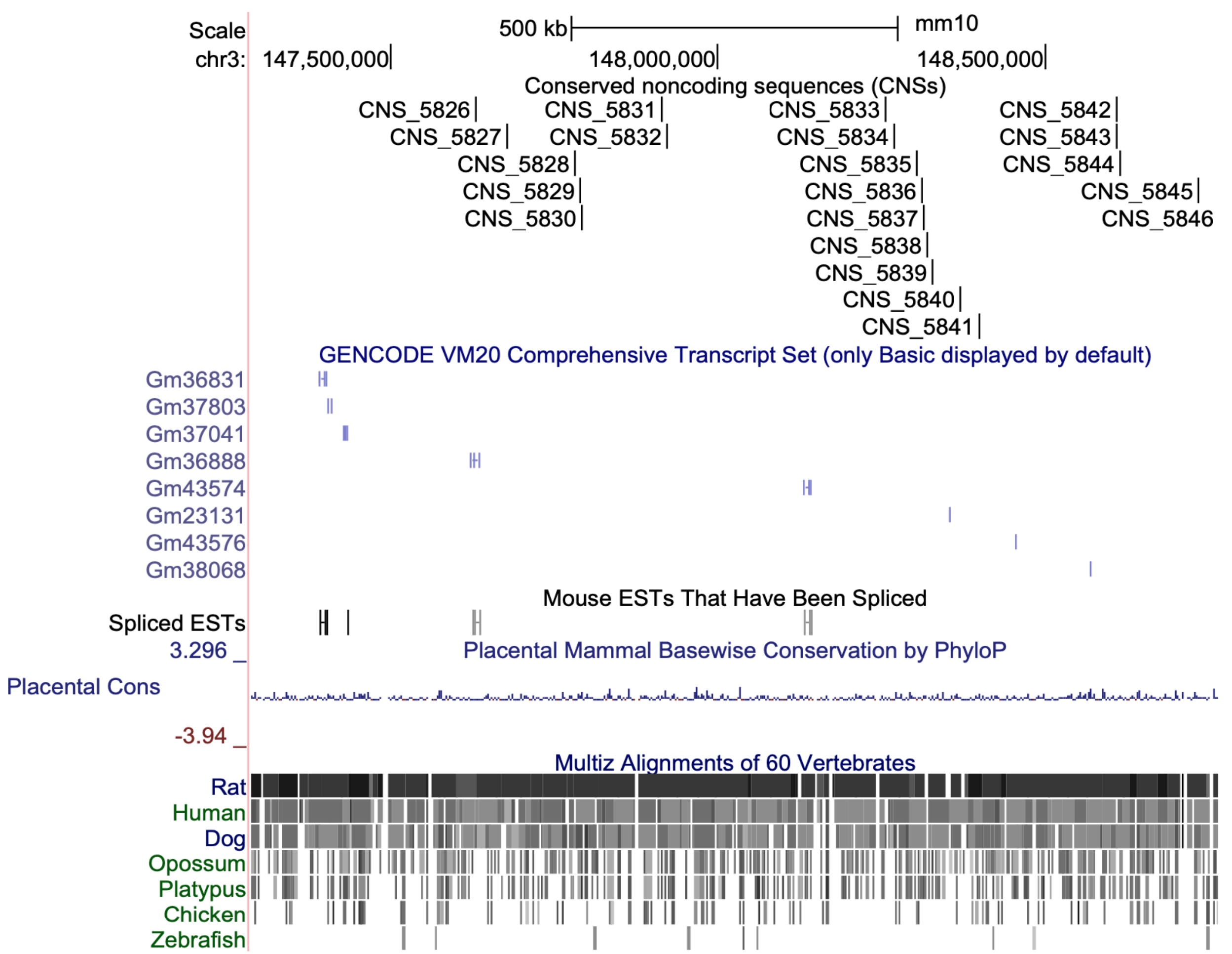
There are other reported possibilities of CNS functions. For example, Hezroni et al. (2017) reported that some pseudogenes are conserved across species. The conservation of promoters of these pseudogenes is particularly reported to be relatively high. Whether the conserved sequences of these pseudogenes acquired new functions after pseudogenization or they had another function before pseudogenization is often not very clear. Also, recent improvement in sequencing technologies has made it possible to detect even lowly expressed transcripts. This has led to the annotation of tens of thousands of lncRNAs (Harrow et al. 2012; Lagarde et al. 2017; Uszczynska-Ratajczak et al. 2018; C. Xie et al. 2014). Low expression level and extremely high tissue and stage specificity of lncRNAs (Mattioli et al. 2019; Necsulea et al. 2014; Talyan et al. 2018) make their detection difficult. Cheaper and better sequencing technologies however have made it possible to detect more lncRNAs. Therefore, there is a possibility that some of the CNSs might correspond to lncRNAs or other classes of RNAs. Indeed some ultraconserved elements have been reported to be transcribed in human cancers (Peng et al. 2013). However, there is a little overlap between GENCODE lincRNA exons and CNSs (**Figure 1B**). Also, ENCODE project (ENCODE Project Consortium 2012) ascribed some "biochemical functions" to about 80% of the human genome. Some of these functions were linked to the ChIP-Seq results. Specifically, data on DNA binding, transcription, DNA accessibility and DNA methylation would give relevant insights about the functions of specific genomic regions. Thus, overlap with ENCODE data might give insights into other likely functions of CNSs. Finally, roadmap epigenome data (Bernstein et al. 2010) might be valuable as some histone modification marks have been associated with enhancer activities (Babarinde and Saitou 2016; Creyghton et al. 2010).

Unexpected switch: No obvious effect of CNS deletion on mutated individuals

With the extremity of CNS conservation at nucleotide level, the evolutionary constraint is believed to be extremely high. Therefore, deletion of such sequences is expected to result in seriously affected phenotype. In a surprising turn, no obvious phenotypic change was found when 1.5 Mbp CNS-rich gene desert in chromosome 3 (**Figure 5**) and another 845 kbp gene desert in chromosome 19 were deleted in mouse (Nóbrega et al. 2004). The 1.5 Mbp region contains 21 CNSs identified in (Babarinde and Saitou 2016). Similarly, deletion of ultraconserved elements have also been reported to yield viable mice (Ahituv et al. 2007). These reports were puzzling, considering the level of constraints on CNSs. A number of studies have shed some lights into the puzzle. First, the laboratory environment under which the mice were raised might be very different from real environment in which the animals live. This possibility is reiterated by a *Drosophila* experiment in which the deletion of an enhancer element led to an observable phenotype only at a specific temperature range (Frankel et al. 2010).

Figure 5. UCSC genome view of the mouse region deleted with no obvious phenotype

The genome view, captured from https://www.genome.ucsc.edu/ on 12th September, 2019, shows MMU3 of Nóbrega et al. (2004) which corresponds to chr3:147287066-148764954 of mm10 build of mouse genome. This region contains 21 CNSs identified in Babarinde and Saitou (2016).



From the population genetics point of view, the controlled environment and the smaller effective population size could greatly reduce the impact of evolutionary force especially under small selection coefficient (Kousathanas et al. 2011; Mueller et al. 2013). Also, it is possible that there is redundancy in CNS function. Such enhancer redundancy has been reported to provide phenotypic robustness in mammalian development (Osterwalder et al. 2018). Another thing to consider is the ability to measure phenotypic changes. This greatly impacts the observable phenotypic changes. For example, Dickel et al. (2018) reported that mice with deletions in each or pairs of the studied ultraconserved elements were viable and fertile as previously reported (Ahituv et al. 2007; Nóbrega et al. 2004). However, further analyses revealed that in nearly all cases, there were neurological or growth abnormalities (Dickel et al. 2018). These abnormalities included alterations of neuron populations and structural brain defects. They concluded that some of the phenotypic changes induced by deletion of ultraconserved elements are real but might be too subtle to be discovered in normal laboratory settings (Dickel et al. 2018). Therefore, inability to discover the phenotypic effect of a CNS might not imply lack of function; it might just reflect the limited experimental ability to discover such phenotypic effects.

Current puzzles: unanswered questions

While some CNSs have strong phenotypic effects, others have rather subtle effects that are often difficult to observe. However, not much is known about the characteristic differences of these classes of CNSs. As numerous properties of CNSs have been enumerated, one of these properties might be different depending on the phenotypic effect. Another puzzle that has attracted less attention is the reason for the length of CNSs. As an enhancer element which binds a transcription factor, the typical TFBS can be shorter than 10 bp. However, many CNSs spans more than 100bp (Table 1). What differentiates longer CNSs from shorter CNSs at the functional level is not yet fully understood despite the understanding of super enhancers (Hnisz et al. 2013; Pott and Lieb 2015; Whyte et al. 2013). With the rapidly falling sequencing cost, it is now possible to detect new transcripts with tissue-specific expression. It is important to understand how overlapping the transcripts and CNSs are, in a hypothetical scenario of monitoring expressions across all stages and tissues as well as cell types. Finally, there are other epigenomic signatures (Tsankov et al. 2015; W. Xie et al. 2013) of regulatory elements including transcribed enhancers (Andersson et al. 2014), protein-DNA interaction (Landt et al. 2012; Visel et al. 2009) and histone methylation or acetylation (Barski et al. 2007; Creyghton et al. 2010) investigated by ChIP-Seq technologies as well as DNA methylation data (Sharifi-Zarchi et al. 2017). It would be interesting to dissect the activities of these sequences to unravel the functional dynamics of the sequences as they relate to gene expression regulation.

Conclusion

The studies of CNSs have undergone dramatic evolution of puzzles. In the pre-genomic eras, abundance and the nature of CNSs were a puzzle. After the release of multiple genomes, the puzzles have morphed into those of functions. The inability to observe a phenotypic change when CNSs were deleted was puzzling. However, new studies are illuminating the puzzle of seemingly intact phenotypes in mutants despite the extreme conservation of CNSs. Therefore, the puzzle is gradually evolving to another form. Now, the puzzle is how to extensively link the properties of these sequences to their functional importance. Although a lot of puzzles have been solved about CNSs, it appears that the puzzles are not static. The puzzles are evolving and as they are evolving, so is the biological knowledge of properties and functions of these sequences.

List of abbreviations

BLAST: Basic Local Alignment Search Tool
BLAT: Blast-Like Alignment Tool
ChIP-Seq: Chromatin ImmunoPrecipitation Sequencing
CNEE: Conserved NonExonic Element
CNS: Conserved Noncoding Sequence
CONDOR: COnserved Non-coDing Orthologous Region
ENCODE: ENCyclopedia Of DNA Element
GERP: Genomic Evolutionary Rate Profiling
HCE: Highly Conserved Element
HCNR: Highly Conserved Noncoding Region
HCNS: Highly Conserved Noncoding Sequence
LCNS: Long Conserved Noncoding Sequence
lncRNA, lincRNA: long noncoding RNA, long intergenic noncoding RNA
RNA, mRNA: RiboNucleic Acid, messenger RNA

STAG-CNS: Suffix Tree Arbitrary Gene number: Conserved Noncoding Sequence
TFBM: Transcription Factor Binding Motif
TSS: Transcription Start Site
UCE: UltraConserved Element
UCNE: UltraConserved Noncoding Element
UTR: UnTranslated Region

References

AHITUV Nadav, ZHU Yiwen, VISEL Axel, HOLT Amy, AFZAL Veena, PENNACCHIO Len A., and RUBIN Edward M. (2007) Deletion of ultraconserved elements yields viable mice. PLoS Biology, vol. 5, pp. e234.

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, GEBHARD Claudia, MIGUEL-ESCALADA Irene, HOOF Ilka, BORNHOLDT Jette, BOYD Mette, CHEN Yun, ZHAO Xiaobei, SCHMIDL Christian, SUZUKI Takahiro, NTINI Evgenia, ARNER Erik, VALEN Eivind, LI Kang, SCHWARZFISCHER Lucia, GLATZ Dagmar, RAITHEL Johanna, LILJE Berit, RAPIN Nicolas, … , and SANDELIN Albin (2014) An atlas of active enhancers across human cell types and tissues. Nature, vol. 507, pp. 455–461.

ASTHANA Saurabh, NOBLE William S., KRYUKOV Gregory, GRANT Charles E., SUNYAEV Shamil, and STAMATOYANNOPOULOS John A. (2007) Widely distributed noncoding purifying selection in the human genome. Proceedings of the National Academy of Sciences of the United States of America, vol. 104, pp. 12410–12415.

AYAD Lorraine A. K., PISSIS Solon P., and POLYCHRONOPOULOS Dimitris (2018) CNEFinder : finding conserved non-coding elements in genomes pp. 2003–2007.

BABARINDE Isaac Adeyemi and SAITOU Naruya (2013) Heterogeneous Tempo and Mode of Conserved Noncoding Sequence Evolution among Four Mammalian Orders. Genome Biology and Evolution, vol. 5, pp. 2330–2343.

BABARINDE Isaac Adeyemi and SAITOU Naruya (2016) Genomic Locations of Conserved Noncoding Sequences and Their Proximal Protein-Coding Genes in Mammalian Expression Dynamics. Molecular Biology and Evolution, vol. 33, pp. 1807–1817.

BABARINDE Isaac Adeyemi and SAITOU Naruya (2020) The Dynamics, causes, and impacts of mammalian evolutionary rates revealed by the analyses of capybara draft genome sequences. Genome Biology and Evolution, vol. 12, pp. 1444–1458.

BAGADIA Meenakshi, CHANDRADOSS Keerthivasan Raanin, JAIN Yachna, SINGH Harpreet, LAL Mohan, and SANDHU Kuljeet Singh (2019) Evolutionary Loss of Genomic Proximity to Conserved Noncoding Elements Impacted the Gene Expression Dynamics During Mammalian Brain Development. Genetics, vol. 211, pp. 1239–1254.

BAILEY Timothy L., BODEN Mikael, BUSKE Fabian A., FRITH Martin, GRANT Charles E., CLEMENTI Luca, REN Jingyuan, LI Wilfred W., and NOBLE William S. (2009) MEME Suite: Tools for motif discovery and searching. Nucleic Acids Research, vol. 37, pp. W202–W208.

BARSKI Artem, CUDDAPAH Suresh, CUI Kairong, ROH Tae-Young, SCHONES Dustin E., WANG Zhibin, WEI Gang, CHEPELEV Iouri, and ZHAO Keji (2007) High-Resolution Profiling of Histone Methylations in the Human Genome. Cell, vol. 129, pp. 823–837.

BEJERANO Gill, PHEASANT Michael, MAKUNIN Igor, STEPHEN Stuart, KENT W. James, MATTICK John S., and HAUSSLER David (2004) Ultraconserved elements in the human genome. Science (New York, N.Y.), vol. 304, pp. 1321–1325.

BERNSTEIN Bradley E., STAMATOYANNOPOULOS John A., COSTELLO Joseph F., REN Bing, MILOSAVLJEVIC Aleksandar, MEISSNER Alexander, KELLIS Manolis, MARRA Marco A., BEAUDET Arthur L., ECKER Joseph R., FARNHAM Peggy J., HIRST Martin, LANDER Eric S., MIKKELSEN Tarjei S., and THOMSON James A. (2010) The NIH Roadmap Epigenomics Mapping Consortium. Nature Biotechnology, vol. 28, pp. 1045–1048.

BRODY Thomas, YAVATKAR Amarendra, KUZIN Alexander, and ODENWALD Ward F. (2020) Ultraconserved non-coding DNA within diptera and hymenoptera. G3: Genes, Genomes, Genetics, vol. 10, pp. 3015–3024.

BULGER Michael, VON Hikke Doorninck J., SAITOH Noriko, TELLING Agnes, FARRELL Catherine, BENDER M. A., FELSENFELD Gary, AXEL Richard, and GROUDINE Mark (1999) Conservation of sequence and structure flanking the mouse and human ß-globin loci: The ß-globin genes are embedded within an array of odorant receptor genes. Proceedings of the National Academy of Sciences of the United States of America, vol. 96, pp. 5129–5134.

BUSH Eliot C. and LAHN Bruce T. (2005) Selective constraint on noncoding regions of hominid genomes. PLoS Computational Biology, vol. 1, pp. e73.

CREYGHTON Menno P., CHENG Albert W., WELSTEAD G. Grant, KOOISTRA Tristan, CAREY Bryce W., STEINE Eveline J., HANNA Jacob, LODATO Michael A., FRAMPTON Garrett M., SHARP Phillip A., BOYER Laurie A., YOUNG Richard A., 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.

DAWSON S. R., TURNER D. L., WEINTRAUB H., and PARKHURST S. M. (1995) Specificity for the hairy/enhancer of split basic helix-loop-helix (bHLH) proteins maps outside the bHLH domain and suggests two separable modes of transcriptional repression. Molecular and Cellular Biology, vol. 15, pp. 6923–6931.

DE LA CALLE-MUSTIENES Elisa, FEIJÓO Cármen Gloria, MANZANARES Miguel, TENA Juan J., RODRÍGUEZ-SEGUEL Elisa, LETIZIA Annalisa, ALLENDE Miguel L., and GÓMEZ-SKARMETA José Luis (2005) A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Research, vol. 15, pp. 1061–1072.

DEKKER Job (2007) GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biology, vol. 8, pp. R116.
DICKEL Diane E., YPSILANTI Athena R., PLA Ramón, ZHU Yiwen, BAROZZI Iros, MANNION Brandon J., KHIN Yupar S., FUKUDA-YUZAWA Yoko, PLAJZER-FRICK Ingrid, PICKLE Catherine S., LEE Elizabeth A., HARRINGTON Anne N., PHAM Quan T., GARVIN Tyler H., KATO Momoe, OSTERWALDER Marco, AKIYAMA Jennifer A., AFZAL Veena, RUBENSTEIN John L. R., … VISEL Axel (2018) Ultraconserved enhancers are required for normal development. Cell, vol. 172, pp. 491- 499.e15.
DIMITRIEVA Slavica and BUCHER Philipp (2013) UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Research, vol. 41, pp. D101-9.
DOUSSE Aline, JUNIER Thomas, and ZDOBNOV Evgeny M. (2016) CEGA — a catalog of conserved elements from genomic alignments vol. 44, pp. 96–100.
DRAKE Jared A., BIRD Christine, NEMESH James, THOMAS Daryl J., NEWTON-CHEH Christopher, REYMOND Alexandre, EXCOFFIER Laurent, ATTAR Homa, ANTONARAKIS Stylianos E., DERMITZAKIS Emmanouil T., and HIRSCHHORN Joel N. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nature Genetics, vol. 38, pp. 223–227.
ELGAR Greg and VAVOURI Tanya (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends in Genetics, vol. 24, pp. 344–352.
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, vol. 489, pp. 57–74.
FISHILEVICH Simon, NUDEL Ron, RAPPAPORT Noa, HADAR Rotem, PLASCHKES Inbar, INY Stein Tsippi, ROSEN Naomi, KOHN Asher, TWIK Michal, SAFRAN Marilyn, LANCET Doron, and COHEN Dana (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database, vol. 2017, pp. 1-17.
FRANKEL Nicolás, DAVIS Gregory K., VARGAS Diego, WANG Shu, PAYRE François, and STERN David L. (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature, vol. 466, pp. 490–493.
FURNISS Dominic, LETTICE Laura A., TAYLOR Indira B., CRITCHLEY Paul S., GIELE Henk, HILL Robert E., and WILKIE Andrew O. M. (2008) A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. Human Molecular Genetics, vol. 17, pp. 2417–2423.
GARBER Manuel, GUTTMAN Mitchell, CLAMP Michele, ZODY Michael C., FRIEDMAN Nir, and XIE Xiaohui (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics, vol. 25, pp. i54–i62.
GLAZOV Evgeny A., PHEASANT Michael, MCGRAW Elizabeth A., BEJERANO Gill, and MATTICK John S. (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. Genome Research, vol. 15, pp. 800–808.
HAEUSSLER Maximilian, ZWEIG Ann S., TYNER Cath, SPEIR Matthew L., ROSENBLOOM Kate R., RANEY Brian J., LEE Christopher M., LEE Brian T., HINRICHS Angie S., GONZALEZ Jairo Navarro, GIBSON David, DIEKHANS Mark, CLAWSON Hiram, CASPER

Jonathan, BARBER Galt P., HAUSSLER David, KUHN Robert M., and KENT W. James (2019) The UCSC Genome Browser database: 2019 update. Nucleic Acids Research, vol. 47, pp. D853–D858.

HARDISON Ross C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends in Genetics, vol. 16, pp. 369–372.

HARDISON Ross C., OELTJEN John, and MILLER Webb (1997) Long human mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. Genome Research, vol. 7, pp. 959–966.

HARROW Jennifer, FRANKISH Adam, GONZALEZ Jose M., TAPANARI Electra, DIEKHANS Mark, KOKOCINSKI Felix, AKEN Bronwen L., BARRELL Daniel, ZADISSA Amonida, SEARLE Stephen, BARNES If, BIGNELL Alexandra, BOYCHENKO Veronika, HUNT Toby, KAY Mike, MUKHERJEE Gaurab, RAJAN Jeena, DESPACIO-REYES Gloria, SAUNDERS Gary, …, and HUBBARD Tim J. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research, vol. 22, pp. 1760–1774.

HERMANN Brian P. and HECKERT Leslie L. (2005) Silencing of Fshr Occurs through a Conserved, Hypersensitive Site in the First Intron. Molecular Endocrinology, vol. 19, pp. 2112–2131.

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.

HEZRONI Hadas, BEN-TOV Perry Rotem, MEIR Zohar, HOUSMAN Gali, LUBELSKY Yoav, and ULITSKY Igor (2017) A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. Genome Biology, vol. 18, pp. 162.

HNISZ Denes, ABRAHAM Brian J., LEE Tong Ihn, LAU Ashley, SAINT-ANDRÉ Violaine, SIGOVA Alla A., HOKE Heather A., and YOUNG Richard A. (2013) Super-enhancers in the control of cell identity and disease. Cell, vol. 155, pp. 934–947.

HUTCHINS Andrew Paul and PEI Duanqing (2015) Transposable elements at the center of the crossroads between embryogenesis, embryonic stem cells, reprogramming, and long non-coding RNAs. Science Bulletin, vol. 60, pp. 1722–1733.

INOUE Jun and SAITOU Naruya (2020) dbCNS : a new database for conserved noncoding sequences. Molecular Biology and Evolution, Advance Access publication November 16, 2020 (doi:10.1093/molbev/msaa296).

ISHIBASHI Minaka, NODA Akiko Ogura, SAKATE Ryuichi, and IMANISHI Tadashi (2012) Evolutionary growth process of highly conserved sequences in vertebrate genomes. Gene, vol. 504, pp. 1–5.

IYER Matthew K., NIKNAFS Yashar S., MALIK Rohit, SINGHAL Udit, SAHU Anirban, HOSONO Yasuyuki, BARRETTE Terrence R., PRENSNER John R., EVANS Joseph R., ZHAO Shuang, POLIAKOV Anton, CAO Xuhong, DHANASEKARAN Saravana M., WU Yi-Mi, ROBINSON Dan R., BEER David G., FENG Felix Y., IYER Hariharan K., and CHINNAIYAN Arul M. (2015) The landscape of long noncoding RNAs in the human transcriptome. Nature Genetics, vol. 47, pp. 199–208.

JANES D. E., CHAPUS C., GONDO Y., CLAYTON D. F., SINHA S., BLATTI C. A., ORGAN C. L., FUJITA M. K., BALAKRISHNAN C. N., and EDWARDS S. V. (2011) Reptiles and mammals have differentially retained long conserved noncoding sequences from the amniote ancestor. Genome Biology and Evolution, vol. 3, pp. 102–113.

JAREBORG Niclas, BIRNEY Ewan, and DURBIN Richard (1999) Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. Genome Research, vol. 9, pp. 815–824.

KAMOUN Choumouss, PAYEN Thibaut, HUA-VAN Aurélie, and FILÉE Jonathan (2013) Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. BMC Genomics, vol. 14, article number 700.

KELLIS Manolis, WOLD Barbara, SNYDER Michael P., BERNSTEIN Bradley E., KUNDAJE Anshul, MARINOV Georgi K., WARD Lucas D., BIRNEY Ewan, CRAWFORD Gregory E., DEKKER Job, DUNHAM Ian, ELNITSKI Laura L., FARNHAM Peggy J., FEINGOLD Elise A., GERSTEIN Mark, GIDDINGS Morgan C., GILBERT David M., GINGERAS Thomas R., GREEN Eric D., … Hardison Ross C. (2014) Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences of the United States of America, vol. 111, pp. 6131–6138.

KENT W. James (2002) BLAT--the BLAST-like alignment tool. Genome Research, vol. 12, pp. 656–664.

KIM Yun-Ji, LEE Jungnam, and HAN Kyudong (2012) Transposable Elements: No More "Junk DNA". Genomics & Informatics, vol. 10, pp. 226–233.

KIMURA Motoo (1983) The neutral theory of molecular evolution Cambridge University Press.

KIMURA Motoo and OHTA Tomoko (1974) On some principles governing molecular evolution. Proceedings of the National Academy of Sciences of the United States of America, vol. 71, pp. 2848–2852.

KING Mary Claire and WILSON A. C. (1975) Evolution at two levels in humans and chimpanzees. Science, vol. 188, pp. 107–116.

KOUSATHANAS Athanasios, OLIVER Fiona, HALLIGAN Daniel L., and KEIGHTLEY Peter D. (2011) Positive and Negative Selection on Noncoding DNA Close to Protein-Coding Genes in Wild House Mice. Molecular Biology and Evolution, vol. 28, pp. 1183–1191.

KULAKOVSKIY Ivan V., VORONTSOV Ilya E., YEVSHIN Ivan S., SHARIPOV Ruslan N., FEDOROVA Alla D., RUMYNSKIY Eugene I., MEDVEDEVA Yulia A., MAGANA-MORA Arturo, BAJIC Vladimir B., PAPATSENKO Dmitry A., KOLPAKOV Fedor A., and MAKEEV Vsevolod J. (2018) HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Research, vol. 46, pp. D252–D259.

LAGARDE Julien, USZCZYNSKA-RATAJCZAK Barbara, CARBONELL Silvia, PÉREZ-LLUCH Sílvia, ABAD Amaya, DAVIS Carrie, GINGERAS Thomas R., FRANKISH Adam, HARROW Jennifer, GUIGO Roderic, and JOHNSON Rory (2017) High-throughput annotation of full-length long noncoding RNAs with Capture Long-Read Sequencing. Nature Genetics, vol. 49, pp. 1731-1740.

LAI Xianjun, BEHERA Sairam, LIANG Zhikai, LU Yanli, DEOGUN Jitender S., and SCHNABLE James C. (2017) STAG-CNS: An Order-Aware Conserved Noncoding Sequences Discovery Tool for Arbitrary Numbers of Species. Molecular Plant, vol. 10, pp. 990–999.

LANDT Stephen G., MARINOV Georgi K., KUNDAJE Anshul, KHERADPOUR Pouya, PAULI Florencia, BATZOGLOU Serafim, BERNSTEIN Bradley E., BICKEL Peter, BROWN James B., CAYTING Philip, CHEN Yiwen, DESALVO Gilberto, EPSTEIN Charles, FISHER-AYLOR Katherine I., EUSKIRCHEN Ghia, GERSTEIN Mark, GERTZ Jason, HARTEMINK Alexander J., HOFFMAN Michael M., …, and SNYDER Michael (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research, vol. 22, pp. 1813–1831.

LETTICE Laura A., HORIKOSHI Taizo, HEANEY Simon J. H., VAN Baren Marijke J., VAN der Linde Herma C., BREEDVELD Guido J., JOOSSE Marijke, AKARSU Nurten, OOSTRA Ben A., ENDO Naoto, SHIBATA Minoru, SUZUKI Mikio, TAKAHASHI Eiichi, SHINKA Toshikatsu, NAKAHORI Yutaka, AYUSAWA Dai, NAKABAYASHI Kazuhiko, SCHERER Stephen W., HEUTINK Peter, …, and NOJI Sumihare (2002) Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proceedings of the National Academy of Sciences of the United States of America, vol. 99, pp. 7548–7553.

LI Heng and DURBIN Richard (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), vol. 26, pp. 589–595.

LOOTS G. G., LOCKSLEY R. M., BLANKESPOOR C. M., WANG Z. E., MILLER W., RUBIN E. M., and FRAZER K. A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science, vol. 288, pp. 136–140.

LOU H., YANG Y., COTE G. J., BERGET S. M., and GAGEL R. F. (1995) An intron enhancer containing a 5' splice site sequence in the human calcitonin/calcitonin gene-related peptide gene. Molecular and Cellular Biology, vol. 15, pp. 7135–7142.

LOWE Craig B., KELLIS Manolis, SIEPEL Adam, RANEY Brian J., CLAMP Michele, SALAMA Sofie R., KINGSLEY David M., LINDBLAD-TOH Kerstin, and HAUSSLER David (2011) Three periods of regulatory innovation during vertebrate evolution. Science (New York, N.Y.), vol. 333, pp. 1019–1024.

MAHMOUDI Saber Morteza and SAITOU Naruya (2017) Silencing Effect of Hominoid Highly Conserved Noncoding Sequences on Embryonic Brain Development. Genome Biology and Evolution, vol. 9, pp. 2122–2133.

MATSUNAMI Masatoshi, SUMIYAMA Kenta, and SAITOU Naruya (2010) Evolution of Conserved Non-Coding Sequences Within the Vertebrate Hox Clusters Through the Two-Round Whole Genome Duplications Revealed by Phylogenetic Footprinting Analysis. Journal of Molecular Evolution, vol. 71, pp. 427–436.

MATTIOLI Kaia, VOLDERS Pieter-Jan, GERHARDINGER Chiara, LEE James C., MAASS Philipp G., MELÉ Marta, and RINN John L. (2019) High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. Genome Research, vol. 29, pp. 344–355.

MCEWEN Gayle K., WOOLFE Adam, GOODE Debbie, VAVOURI Tanya, CALLAWAY Heather, and ELGAR Greg (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Research, vol. 16, pp. 451–465.

MCLEAY Robert C. and BAILEY Timothy L. (2010) Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. BMC Bioinformatics, vol. 11, article number 165.

MIGNONE Flavio, ANSELMO Anna, DONVITO Giacinto, MAGGI Giorgio P., GRILLO Giorgio, and PESOLE Graziano (2008) Genome-wide identification of coding and non-coding conserved sequence tags in human and mouse genomes. BMC Genomics, vol. 9, article number 277.

MUELLER Laurence D., JOSHI Amitabh, SANTOS Marta, and ROSE Michael R. (2013) Effective population size and evolutionary dynamics in outbred laboratory populations of Drosophila. Journal of Genetics, vol. 92, pp. 349–361.

NECSULEA Anamaria, SOUMILLON Magali, WARNEFORS Maria, LIECHTI Angélica, DAISH Tasman, ZELLER Ulrich, BAKER Julie C., GRÜTZNER Frank, and KAESSMANN Henrik (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, vol. 505, pp. 635–640.

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

NÓBREGA Marcelo A., ZHU Yiwen, PLAJZER-FRICK Ingrid, AFZAL Veena, and RUBIN Edward M. (2004) Megabase deletions of gene deserts result in viable mice. Nature, vol. 431, pp. 988–993.

OELTJEN John C., MALLEY Tracy M., MUZNY Donna M., MILLER Webb, GIBBS Richard A., and BELMONT John W. (1997a) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. Genome Research, vol. 7, pp. 315–329.

OELTJEN John C., MALLEY Tracy M., MUZNY Donna M., MILLER Webb, GIBBS Richard A., and BELMONT John W. (1997b) Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. Genome Research, vol. 7, pp. 315–329.

OSTERWALDER Marco, BAROZZI Iros, TISSIÈRES Virginie, FUKUDA-YUZAWA Yoko, MANNION Brandon J., AFZAL Sarah Y., LEE Elizabeth A., ZHU Yiwen, PLAJZER-FRICK Ingrid, PICKLE Catherine S., KATO Momoe, GARVIN Tyler H., PHAM Quan T., HARRINGTON Anne N., AKIYAMA Jennifer A., AFZAL Veena, LOPEZ-RIOS Javier, DICKEL Diane E., VISEL Axel, and PENNACCHIO Len A. (2018) Enhancer redundancy provides phenotypic robustness in mammalian development. Nature, vol. 554, pp. 239–243.

PENG Jiang Chen, SHEN Jun, and RAN Zhi Hua (2013) Transcribed ultraconserved region in human cancers. RNA Biology, vol. 10, pp. 1771–1777.

PERSAMPIERI Jason, RITTER Deborah I., LEES Daniel, LEHOCZKY Jessica, LI Qiang, GUO Su, and CHUANG Jeffrey H. (2008) cneViewer: A database of conserved non-coding elements for studies of tissue-specific gene regulation. Bioinformatics, vol. 24, pp. 2418–2419.

POLLARD Katherine S., HUBISZ Melissa J., ROSENBLOOM Kate R., and SIEPEL Adam (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Research, vol. 20, pp. 110–121.

POON R., WAI Kan Y., and BOYER H. W. (1978) Sequence of the 3' noncoding and adjacent coding regions of human ? globin mRNA. Nucleic Acids Research, vol. 5, pp. 4625–4630. http://www.ncbi.nlm.nih.gov/pubmed/318163

POTT Sebastian and LIEB Jason D. (2015) What are super-enhancers? Nature Genetics, vol. 47, pp. 8–12.

RAISNER Ryan, KHARBANDA Samir, JIN Lingyan, JENG Edwin, CHAN Emily, MERCHANT Mark, HAVERTY Peter M., BAINER Russell, CHEUNG Tommy, ARNOTT David, FLYNN E. Megan, ROMERO F. Anthony, MAGNUSON Steven, and GASCOIGNE Karen E. (2018) Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. Cell Reports, vol. 24, pp. 1722–1729.

RAMANI Ritika, KRUMHOLZ Katie, HUANG Yi-Fei, and SIEPEL Adam (2019) PhastWeb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastCons and phyloP. Bioinformatics, vol. 35, pp. 2320–2322.
RAMÍREZ Fidel, RYAN Devon P., GRÜNING Björn, BHARDWAJ Vivek, KILPERT Fabian, RICHTER Andreas S., HEYNE Steffen, DÜNDAR Friederike, and MANKE Thomas (2016) deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research, vol. 44, pp. W160–W165.
RAMSAY LeeAnn, MARCHETTO Maria C., CARON Maxime, CHEN Shu-Huang, BUSCHE Stephan, KWAN Tony, PASTINEN Tomi, GAGE Fred H., and BOURQUE Guillaume (2017) Conserved expression of transposon-derived non-coding transcripts in primate stem cells. BMC Genomics, vol. 18, article number 214.
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.
SAGAI Tomoko, HOSOYA Masaki, MIZUSHINA Youichi, TAMURA Masaru, SHIROISHI Toshihiko, and BALLING R. (2005) Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development (Cambridge, England), vol. 132, pp. 797–803.
SAKURABA Yoshiyuki, KIMURA Toru, MASUYA Hiroshi, NOGUCHI Hideki, SEZUTSU Hideki, TAKAHASI K. Ryo, TOYODA Atsushi, FUKUMURA Ryutaro, MURATA Takuya, SAKAKI Yoshiyuki, YAMAMURA Masayuki, WAKANA Shigeharu, NODA Tetsuo, SHIROISHI Toshihiko, and GONDO Yoichi (2008) Identification and characterization of new long conserved noncoding sequences in vertebrates. Mammalian Genome, vol. 19, pp. 703–712.
SHARIFI-ZARCHI Ali, GEROVSKA Daniela, ADACHI Kenjiro, TOTONCHI Mehdi, PEZESHK Hamid, TAFT Ryan J., SCHÖLER Hans R., CHITSAZ Hamidreza, SADEGHI Mehdi, BAHARVAND Hossein, and ARAÚZO-BRAVO Marcos J. (2017) DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics, vol. 18, pp. 964.
SIEPEL Adam, BEJERANO Gill, PEDERSEN Jakob S., HINRICHS Angie S., HOU Minmei, ROSENBLOOM Kate, CLAWSON Hiram, SPIETH John, HILLIER Ladeana W., RICHARDS Stephen, WEINSTOCK George M., WILSON Richard K., GIBBS Richard A., KENT W. James, MILLER Webb, and HAUSSLER David (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research, vol. 15, pp. 1034–1050.
SUMIYAMA Kenta, MIYAKE Tsutomu, GRIMWOOD Jane, STUART Andrew, DICKSON Mark, SCHMUTZ Jeremy, RUDDLE Frank H., MYERS Richard M., and AMEMIYA Chris T. (2012) Theria-specific homeodomain and cis-regulatory element evolution of the Dlx3-4 bigene cluster in 12 different mammalian species. Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution, vol. 318, pp. 639–650.
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.
TALYAN Sweta, ANDRADE-NAVARRO Miguel A., and MURO Enrique M. (2018) Identification of transcribed protein coding sequence remnants within lincRNAs. Nucleic Acids Research, vol. 46, pp. 8720–8729.

TSANKOV Alexander M., GU Hongcang, AKOPIAN Veronika, ZILLER Michael J., DONAGHEY Julie, AMIT Ido, GNIRKE Andreas, and MEISSNER Alexander (2015) Transcription factor binding dynamics during human ES cell differentiation. Nature, vol. 518, pp. 344–349.

USZCZYNSKA-RATAJCZAK Barbara, LAGARDE Julien, FRANKISH Adam, GUIGÓ Roderic, and JOHNSON Rory (2018) Towards a complete map of the human long non-coding RNA transcriptome. Nature Reviews Genetics, vol. 19, pp. 535–548.

VAN DE VELDE Jan, VAN Bel Michiel, VANEECHOUTTE Dries, and VANDEPOELE Klaas (2016) A collection of conserved noncoding sequences to study gene regulation in flowering plants. Plant Physiology, vol. 171, pp. 2586–2598.

VAN Hellemont Ruth, MONSIEURS Pieter, THIJS Gert, DE Moor Bart, VAN de Peer Yves, and MARCHAL Kathleen (2005) A novel approach to identifying regulatory motifs in distantly related genomes. Genome Biology, vol. 6, article number R113.

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 number R15.

VISEL Axel, BLOW Matthew J., LI Zirong, ZHANG Tao, AKIYAMA Jennifer A., HOLT Amy, PLAJZER-FRICK Ingrid, SHOUKRY Malak, WRIGHT Crystal, CHEN Feng, AFZAL Veena, REN Bing, RUBIN Edward M., and PENNACCHIO Len A. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature, vol. 457, pp. 854–858.

VISEL Axel, MINOVITSKY Simon, DUBCHAK Inna, and PENNACCHIO Len A. (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Research, vol. 35, pp. D88-92.

WHYTE Warren A., ORLANDO David A., HNISZ Denes, ABRAHAM Brian J., LIN Charles Y., KAGEY Michael H., RAHL Peter B., LEE Tong Ihn, and YOUNG Richard A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell, vol. 153, pp. 307–319.

WOOLFE Adam, GOODE Debbie K., COOKE Julie, CALLAWAY Heather, SMITH Sarah, SNELL Phil, MCEWEN Gayle K., and ELGAR Greg (2007) CONDOR: A database resource of developmentally associated conserved non-coding elements. BMC Developmental Biology.

WOOLFE Adam, GOODSON Martin, GOODE Debbie K., SNELL Phil, MCEWEN Gayle K., VAVOURI Tanya, SMITH Sarah F., NORTH Phil, CALLAWAY Heather, KELLY Krys, WALTER Klaudia, ABNIZOVA Irina, GILKS Walter, EDWARDS Yvonne J. K., COOKE Julie E., and ELGAR Greg (2004) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biology, vol. 3, article number e7.

XIE Chaoyong, YUAN Jiao, LI Hui, LI Ming, ZHAO Guoguang, BU Dechao, ZHU Weimin, WU Wei, CHEN Runsheng, and ZHAO Yi (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Research, vol. 42, pp. D98–D103.

XIE Jianbo, QIAN Kecheng, SI Jingna, XIAO Liang, CI Dong, and ZHANG Deqiang (2018) Conserved noncoding sequences conserve biological networks and in fl uence genome evolution. Heredity, pp. 437–451.

XIE Wei, SCHULTZ Matthew D., LISTER Ryan, HOU Zhonggang, RAJAGOPAL Nisha, RAY Pradipta, WHITAKER John W., TIAN Shulan, HAWKINS R. David, LEUNG Danny, YANG Hongbo, WANG Tao, LEE Ah Young, SWANSON Scott A., ZHANG Jiuchun, ZHU Yun, KIM Audrey, NERY Joseph R., URICH Mark A., … Ren Bing (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell, vol. 153, pp. 1134–1148.

ZHAO Yi, LI Hui, FANG Shuangsang, KANG Yue, WU Wei, HAO Yajing, LI Ziyang, BU Dechao, SUN Ninghui, ZHANG Michael Q., and CHEN Runsheng (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Research, vol. 44, pp. D203–D208.

ZHOU Yingyao, ZHOU Bin, PACHE Lars, CHANG Max, KHODABAKHSHI Alireza Hadj, TANASEICHUK Olga, BENNER Christopher, and CHANDA Sumit K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications, vol. 10, article number 1523.

ZHU Shijia, JIANG Qinghua, WANG Guohua, LIU Bo, TENG Mingxiang, and WANG Yadong (2011) Chromatin structure characteristics of pre-miRNA genomic sequences. BMC Genomics, vol. 12, article number 329.

==

Publication history of this article

November 18, AS 0020: review article was submitted from Dr. Isaac BABARINDE to iDarwin (handled by SAITOU Naruya)

December 11, AS 0020: review result (one associate editor and SAITOU reviewed this manuscript) was sent to author

December 31, AS 0020: revision was sent from author

January 5, AS 0021: this revision was accepted for iDarwin

January 11, AS 0021: first proof of this review article was sent to author

January 15, AS 0021: reply to first proof was sent from author

January 17, AS 0021: second proof of this review article was sent to author

January 31, AS 0021: third proof of this review article was sent to author

February 12, AS 0021: this review article is published in iDarwin

Dr. Isaac Adeyemi BABARINDE