MBE Advance Access published March 26, 2016

Article

Genomic locations of conserved noncoding sequences and their proximal protein-coding genes in mammalian expression dynamics

Isaac Adeyemi Babarinde[1,2] and Naruya Saitou[2,1]

[1]Department of Genetics, Graduate University for Advanced Studies, Mishima, Japan;

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

Correspondence to:

Naruya Saitou

Division of Population Genetics, National Institute of Genetics

Mishima 411-8540, Japan

Email: saitounr@nig.ac.jp

Phone/Fax: +81-55-981-6790/6789

Abstract

Experimental studies have found the involvement of certain conserved noncoding sequences (CNSs) in the regulation of the proximal protein-coding genes in mammals. However, reported cases of long range enhancer activities and inter-chromosomal regulation suggest that proximity of CNSs to protein-coding genes might not be important for regulation. To test the importance of the CNS genomic location, we extracted the CNSs conserved between chicken and four mammalian species (human, mouse, dog and cattle). These CNSs were confirmed to be under purifying selection. The intergenic CNSs are often found in clusters in gene deserts, where protein-coding genes are in paucity. The distribution pattern, ChIP-Seq and RNA-Seq data suggested that the CNSs are more likely to be regulatory elements and not corresponding to long intergenic noncoding RNAs (lincRNAs). Physical distances between CNS and their nearest protein coding genes were well conserved between human and mouse genomes, and CNS-flanking genes were often found in evolutionarily conserved genomic neighborhoods. ChIP-Seq signal and gene expression patterns also suggested that CNSs regulate nearby genes. Interestingly, genes with more CNSs have more evolutionarily conserved expression than those with fewer CNSs. These computationally obtained results suggest that the genomic locations of CNSs are important for their regulatory functions. In fact, various kinds of evolutionary constraints may be acting to maintain the genomic locations of CNSs and protein-coding genes in mammals to ensure proper regulation.

Key words: Conserved noncoding sequence, gene expression, regulation, mammalian genome, physical distance.

Introduction

Conserved noncoding sequences (CNSs) are the noncoding parts of the genome that are under sequence constraint probably due to functional importance. Generally identified by computational searches, the exact numbers of CNSs are difficult to estimate. This is because the number of CNSs retrieved from a computational search depends on the threshold used (e.g., Bejerano et al. 2004; Takahashi and Saitou 2012; Babarinde and Saitou 2013). However, some consistent properties of CNSs were found despite the

difference in thresholds used. One such property is the general tendency to cluster around certain types of genes (Woolfe et al. 2005; Takahashi and Saitou 2012; Babarinde and Saitou 2013; Bhatia et al. 2014). Specifically, CNSs are found to be overrepresented around genes involved in transcription, development and the nervous system. On the contrary, mammalian genes associated with defense, immunity and response to stimulus have been shown to have a lower number of CNSs around them (Babarinde and Saitou 2013). Based on their proximity to protein-coding genes, mammalian lineage-specific CNSs were suggested to be important for the regulation of the genes around which they are found (Takahashi and Saitou 2012; Babarinde and Saitou 2013). In fact, computational analyses and experimental analyses have shown the involvement of certain CNSs in the regulation of the closest genes (e.g., Sumiyama et al. 2002; Bhatia et al. 2014). This understanding has been classically employed in identifying the potential regulatory elements of genes of interest (e.g., Göttgens et al. 1999; Sumiyama et al. 2002). Basically, certain lengths of the flanking regions of genes of interest are searched to identify the conserved regions. The identified potential regulatory regions are then tested experimentally (e.g., Göttgens et al. 1999; Antoniv et al. 2001; Sumiyama et al. 2002, 2003; Visel et al. 2009). This approach fundamentally assumes that the regulatory elements are found within a reasonable distance from the gene of interest.

Location of genes in syntenic blocks has also been suggested to be important for gene regulation (Irimia et al. 2012). In this case, the regulatory element of a gene resides in the intron of a proximal gene (see Lettice et al. 2003; Sagai et al. 2004, 2009 for specific examples). Housekeeping genes tend to have shorter introns (Eisenberg and Levanon 2003; Rao et al. 2010), while some genes may have larger introns so that they can “house” their regulatory elements or those of their neighboring genes (Calle-Mustienes et al. 2005). These observations imply that the function of one gene may be affected if the gene and its regulatory elements are not located at the specified location.

Some observations, however, challenge the importance of the coexistence of the regulatory elements and the target genes. One example is the mega base pair long range regulatory activities (Lettice et al. 2003; Sagai et al. 2004, 2009). The distal regulatory element of the *shh* gene is not only located far away from the gene, but it is also found inside intron 5 of another gene, *Lmbr1* (Lettice at al. 2003). This long-range interaction can be brought about by the DNA looping structure (Ong and Corces 2009). In

addition, recent chromosome conformation capture, also called 3C (Dekker et al. 2002), circularized chromosome conformation capture, also called 4C (Zhao et al. 2006) and carbon copy chromosome conformation capture, also called 5C (Dostie and Dekker 2007) technologies have revealed inter-chromosomal DNA interaction suggesting that the regulatory element and the target genes may be located on different chromosomes. This observation is not entirely new, as one promoter on a chromosome was reported to initiate the transcription on a separate chromosome by Morris et al. (1998).

We previously reported that CNSs are not always in homologous positions with respect to genes (Babarinde and Saitou 2013). For example, we found cases in which intergenic CNSs in one species is intronic in another species. These observations suggest that regulatory elements can interact with their target genes from any part of the genome. They only have to be brought in contact during their activity. We can thus hypothesize that the clustering of regulatory elements is not because they regulate nearby gene expression, but because they are more stable and/or active in that region.

Regarding the importance of the genomic location of CNSs with respect to the protein-coding genes, two hypotheses can be tested. The first hypothesis is that the genomic location of CNSs is not related to their regulatory function. The second hypothesis is that the genomic location is important, and that CNSs function best at a specific location. To investigate these hypotheses, we extracted the CNSs conserved among chicken, human, mouse, dog and cattle. Employing a combination of evolutionary and statistical approaches, we found a series of evidence supporting the second hypothesis that the genomic location is important for the regulatory activities of most mammalian CNSs.

Results

Acquisition of mammalian CNSs and their basic characteristics

We conducted a homology search using BLASTN (Altschul et al. 1997) on repeat- and coding sequence-masked genomes of chicken, human, mouse, dog and cattle (see Material and Methods) to identify mammalian CNSs. The chicken genome was used as the query. Since the nucleotide divergence between mammals and chicken is sufficiently large (synonymous substitution rate > 1), we did not have to set percent identity threshold. Therefore, we defined “CNS” in this study as a noncoding region

conserved between chicken and the four mammalian species with a minimum length of 100bp (see Materials and Methods). The searches gave 21,584 chicken CNSs that are conserved in all four mammalian species.

When the coordinates were mapped to human and mouse genomes, 21,191 and 21,026 CNSs were found, respectively (figure S1 and Table S1). The slight difference in the numbers of CNSs is due to lineage-specific duplications. In fact, a closer evaluation of the CNSs with more than one copy shows that duplications are often lineage specific. Of the total 549 CNSs with duplicates in the four species, 508 or 91% were found to have been duplicated only in a single species (see Supplementary material). The duplicates were both tandem and distal. Interestingly, duplicated CNSs tend to be associated with paralogous genes, even if the duplicates are not on the same chromosome (Binomial p value < 0.001). A number of studies have reported more detailed analyses of the functional importance of duplicated CNSs (e.g. Matsunami and Saitou 2013; Vavouri et al. 2006).

We focused on human and mouse genomes because of the quality of the genomes and the availability of data. Out of the human CNSs, 10,120 were found to overlap with protein-coding genes (mostly intronic), hereafter referred to as intragenic CNSs. The remaining 11,071 CNSs which do not overlap with any protein-coding gene are referred to as intergenic CNSs. We found 9,482 intragenic and 11,544 intergenic CNSs from the mouse genome. The conservation levels of the retrieved CNSs, random sequences, lincRNAs and various codon positions of protein-coding genes were tested using the phastcons (figures 1A and S2A) and phyloP (figures 1B and S2B) conservation scores. The average phastcons scores for CNSs are more than 7-fold higher than random sequences. Interestingly, conservation scores clearly distinguished lincRNAs from CNSs (figures 1A, 1B, S2A and S2B). The phastcons scores for CNSs are higher than all the codon positions of the protein coding genes. The distribution patterns of phyloP conservation scores in figure 1B show that the distribution of CNSs divides the distribution of second codon position into three parts. In the first part (scores ≤ 2), there is a higher proportion of second codon datasets. In the second part (scores 2-5), CNSs dominate. In the third part (scores ≥ 5), second codon dominates. This wide distribution pattern of second codon positions probably indicates the wide spectrum of selection forces acting on protein-coding gene evolution. Some

protein-coding regions are poorly conserved while other regions are highly conserved. CNSs on the other hands have fewer poorly conserved regions. In fact, they are more conserved than random sequences, lincRNAs and third codon positions.

However, higher conservation scores do not necessarily imply functional importance. The higher similarity revealed by the conservation scores might just be the result of the low substitution rate in the region (mutation cold-spot hypothesis). To confirm if the regions are actually under purifying selection, we performed derived allele frequency (DAF) spectrum analyses (Drake et al. 2006). For regions that are under purifying selection, derived alleles would be quickly removed and would not be able to spread in the population. Therefore, there would be an excess of low-frequency alleles in the regions. Using genome sequences of chimpanzee, gorilla and orangutan to determine the ancestral states and SNP data from the Asian population of 1000 genome projects (McVean et al. 2012), we compared the DAF spectra in random sequences, lincRNAs, protein-coding genes and CNSs (see Material and Methods). We found an excess of low-frequency alleles in protein-coding genes and CNSs, compared to the random expectation (figure 1C). Like in protein-coding regions, DAF in CNSs is significantly lower (Mann-Whitney U p-value <0.05) than in random sequences and lincRNA exons (Tables S2, S3). However, there is no significant difference in DAF between lincRNA exons and random sequences. This clearly demonstrates that CNSs, like the protein-coding genes, are under purifying selection. The strength of purifying selection is, however, not that high in lincRNAs.

Examination of CNS locations

Having established that the CNSs are under purifying selection, we then analyzed the genomic distribution of intergenic CNSs. We asked whether the CNSs often occur in clusters or in isolation. To answer this, we made 2Mbp sliding windows with a step size of 500kbp, and counted the numbers of coordinates in each window. Figures 2A and S3A show that, compared to the random intergenic sequences, CNSs often exist in clusters in human and mouse genomes, respectively (Chi square p value < 0.001). This shows that the distribution of CNSs is not random.

Are the CNSs preferentially located around protein-coding genes? We associated intergenic

CNSs to the gene with the closest TSS and compared the distribution of the distance to the closest genes. Figures 2B and S4A show that the intergenic CNSs tend to be located far away from the TSS when compared to the random intergenic sequences both for human and mouse genomes. On the contrary, lincRNAs are closer to the protein-coding genes than random expectation. Direct comparison of lincRNAs and intergenic CNSs might be biased because of length difference. We therefore randomly selected intergenic coordinates with the same length and number and found on the same chromosome as lincRNAs (see Supplementary methods for details). The evaluation of the distances to the closest TSS in human (figure S4B) and mouse (figure S4C) shows that lincRNAs tend to be closer to TSS than random sequences. This implies that the results in figures 2B and S4A are not because of the length differences between CNSs and lincRNAs. The location of intergenic CNSs far away from genes suggests that proximity may not be important. However, it does not say much about the importance of the actual genomic location.

To probe the importance of the genomic location between genes and CNSs, we investigated the evolutionary stability of the CNS-gene distance. Conservation of the distance would suggest that their genomic physical distance is important. If the CNS-gene distance is evolutionarily conserved during the mammalian evolution, the distance between human CNS and the closest human gene would be similar to the distance between the orthologous CNS and orthologous gene in the mouse genome. The normalized difference (distance relative difference) between human distance and the orthologous mouse distance would be close to zero. We thus devised a new measure, the relative distance difference (RDD) for this purpose (see Materials and Methods for definition of RDD). We computed RDDs for both intergenic and intragenic CNSs and the corresponding closest protein-coding genes. As control, we also computed RDDs for closest gene pairs (gene-gene). Figure 3A shows that CNS-gene RDDs are lower than gene-gene RDDs (Chi square p value < 0.001). The mean of gene-gene RDDs was 0.55. This value is more than twice the mean of CNS-gene RDD both for intergenic and intragenic CNSs (0.22 and 0.26, respectively). Similar results were found when the RDD values between human and dog (figure S6A) and between human and cow (figure S6B) were computed. This suggests that an evolutionary force may be acting to stabilize CNS-gene distances over time, and that the genomic physical distance may be important for the

integrity of the regulatory function. Interestingly, the corresponding mean RDD value (0.23) for experimentally confirmed vista enhancer elements (Viesel et al. 2007) is comparable to those of CNS-gene. If the distance conservation can be used as an important feature of functionally active enhancer elements, our results suggest that most of the CNSs in our dataset have regulatory activity.

If the conservation of CNS-gene distance is important, we would expect to find more stable expressions in genes with more conserved CNS-gene distance. For this analysis, spearman's expression correlation was computed from human and mouse expression data of 12 tissues with one-to-one orthology between human and mouse. The genes were grouped into four classes based on the RDD values. For genes associated with more than one CNS, the median values of the RDD values were used. Figure 3B shows that genes with lower RDD values tend to have higher expression correlation. This suggests that genes with more conserved CNS distance (lower RDD value) tend to have more stable expression across the evolutionary timescale. This highlights the importance of CNS-gene distance conservation.

Features of CNS-associated genes

Having established the nonrandom distribution of CNSs, we then asked if genes flanked by CNSs have unique features. It has been previously reported that CNSs tend to cluster around genes involved in the nervous system, transcription regulation and development (McEwen et al. 2009; Takahashi and Saitou 2012; Babarinde and Saitou 2013), while they tend to be underrepresented around genes involved in response to stimuli as well as defense and immunity (Babarinde and Saitou 2013). We first established that these gene ontology enrichment patterns also apply to our CNS dataset (figure 4A). The biased genomic location around certain gene functional categories suggests that CNSs may be preferentially located for realizing specific tissue expression patterns. The enrichment patterns of gene functional categories suggest that genes expressed in embryonic brain would have more CNSs because genes involved with development, nervous system and transcription regulation would be expressed at that stage. On the contrary, testis, which is functional at the adult stage, might not express many CNS-associated genes.

To probe these hypotheses that testis and embryonic brain expression patterns follow the gene

ontology enrichment prediction, we used the RNA-Seq data of embryonic brain and testis (Necsulea et al. 2014). We set tissue expression cutoff at 5RPKM and performed an enrichment test. As expected, genes expressed in embryonic brain tended to be associated with more CNSs while genes highly expressed in testis as well as housekeeping genes were associated with fewer CNSs (figure 4B). At different expression levels (0.5RPKM and 1RPKM), the results are similar (figure S6). Focusing on gene expression patterns of embryonic brain, we asked whether genes with more CNSs have higher expression than those with fewer CNSs. Our analyses indeed revealed that genes with more CNSs have higher expression. Genes with no CNS close to them have lower expression levels (figures S7A, S7B). The reverse is found in testis-related expression; genes with no associated CNSs tend to have higher expression in testis (figures S7A, S7B).

In terms of the gene structure and genomic background, are there any features that distinguish CNS-associated genes from others? For genes with more intragenic CNSs, we would expect them to have larger noncoding proportions. To keep the CNSs in one gene, some evolutionary force should act to prevent loss of noncoding regions (e.g. intron). Therefore, they would be expected to have larger noncoding proportion than genes with no CNSs residing in them. Indeed, that is what we observed (figure 5A). Human genes flanked with no CNSs have significantly lower noncoding proportion (86.95%) compared to 96.68%, 98.44% and 99.18% for genes with 1-3 CNSs, 4-9 CNSs and genes with 10 or more CNSs, respectively. The result is similar for the mouse genome (figure 5A). Previous studies have suggested that the evolutionary force may be acting on housekeeping genes such that they would always have shorter intron size (Castillo-Davis et al. 2002; Eisenberg and Levanon 2003; Rao et al. 2010). The prediction of this model is that housekeeping genes would have a more stable intron size over evolutionary time. Since the intron size is related to the proportion of noncoding regions, the noncoding region proportion of housekeeping genes should be stable over evolutionary time. Hence, the correlation between human and mouse noncoding region proportions for housekeeping genes should be higher than for tissue-specific genes. As shown in figure 4B, housekeeping genes are underrepresented in CNSs. Therefore, we would expect to see higher noncoding correlation in genes with no CNSs, which are enriched in housekeeping genes. Although the correlation of noncoding proportion between human and

mouse genomes is high, this feature is not unique to the genes with no CNSs (figure S8A). Therefore, the evolutionary force acting on short-intron genes may be similar in magnitude to the force acting on large-intron genes.

To understand the nature of the intergenic regions of CNS-associated genes, we analyzed the distance of the CNSs to the nearest genes. The distance to the nearest gene tells whether the genes exist in clusters, or in isolation. As would be predicted from figure 2B, genes with no CNSs are significantly closer to the next gene than for genes with many CNSs (figure 5B). To understand the effect of evolution on the distance between closest genes, we computed the correlation of the distances for the genes with one-to-one correspondence between human and mouse genomes. Interestingly, the correlation coefficient (0.59) for the genes with no CNSs is lower than that (>0.68 depending on the number of CNSs) for genes with CNSs (figure S8B). This higher correlation of the distance to the nearest genes in CNS-associated genes suggests that the flanking regions of CNS-associated genes might be conserved in structure. To further probe this, we computed the modified conservation of genomic neighborhood (modCGN) score for the genes. Conservation of gene neighborhood (CGN), originally proposed by De et al. (2009), is the proportion of the number of genes in a window that are found in the homologous window in another species (see Material and Methods for more details). We computed the modCGN value for genes with one-to-one correspondence between human and mouse. Figure 5C shows that genes associated with no CNS have lower values than CNS-associated genes. This result further shows that the genomic neighborhood of genes associated with CNSs is conserved.

CNS association with gene expression dynamics

So far, we have shown that the distribution of CNSs is nonrandom and that genes with CNSs have unique features. However, we have not examined direct involvement of CNSs in the gene expression dynamics. To address this issue, we analyzed the ChIP-Seq and RNA-Seq data. As shown in figures 4B and S7, and as would be expected from the GO enrichment test (figure 4A), CNS-associated genes are more expressed in embryonic brain, and less expressed in testis (figure S7). If CNSs are associated with enhancer activity, higher signal of H3k4me1 and H3k27ac, which are marks of enhancer elements (Akhtar-Zaidi 2005; Kim

et al. 2010; Creyghton et al. 2010), would be expected in brain than in testis. For this analysis, H3k4me1 and H3k4me3 data for 46 human tissues were downloaded from Roadmap Epigenomics Project. In addition, mouse H3k4me1, H3k4me3 and H3k27ac data previously reported by Shen et al. (2012) were used. Human tissues were grouped into four classes, namely fetal brain, other fetal tissues, adult brain and other adult tissues. As would be expected, the highest signal of H3k4me1 (mark of enhancer elements) was found in fetal brain while the lowest was found in non-brain adult tissues (figure 6A). CNS signal in fetal brain is even higher than the signal of lincRNAs. The pattern of signal in H3k4me3, which is the mark of active promoter (e.g. Cain et al. 2010), is different (figure S9). Similar patterns were found in mouse tissues. The embryonic brain has a higher intensity of H3k4me1 than random sequences. The difference between the CNSs and random sequences is not as obvious in testis and liver (figure S10A. The result of H3k27ac, another enhancer mark, is similar to that of H3k4me1 (figure S10C). The higher signal of CNSs is not observed in H3k4me3, as shown in figures 10A and S10B. As would be expected, lincRNAs have higher signals of the three examined marks (figures S9, S10A, B, C). This is because of lincRNA transcription and proximity to protein-coding genes. One exception to this pattern was in human fetal brain tissues in which the CNS signal was higher than that of lincRNAs. Another exception was in adult brain in which the difference was not obvious (figure 6A). These exceptions might be due to gene association with CNSs in these tissues (figure 4B). H3k4me3 specifically differentiates lincRNAs from other coordinates (figures S9, S10B). These analyses show that CNSs are more associated with gene regulation in fetal or embryonic brain.

If the CNSs are actually associated with gene expression regulation, they should be enriched in transcription factor binding sites (TFBS). TFBS enrichment analysis was done using Analysis of Motif Enrichment, AME (McLeay and Bailey, 2010) implemented in MEME suite (Bailey et al. 2009). Using random sequences as control, we ran TFBS enrichment independently for human and mouse in their respective genomes. We found that CNSs are enriched (corrected Fisher Exact p-value <0.05) in 344/426 (>80%) of the total human TFBS in the HOCOMOCO database (Kulakovskiy et al. 2012). Similar results were found in mouse with CNSs found to be statistically enriched in 348/386 (>90%) of the total mouse TFBS. Interestingly, CNSs are not enriched in CTCF binding sites. CTCF is important for insulator

activity and enhancer blocking (Herold et al. 2012, Hou et al. 2008). CTCF protein has been reported to be functionally important and highly conserved across species (Filippova 1996). Also, some of its binding sites have also been previously reported to be evolutionarily conserved (Farrell et al. 2002). That CTCF binding sites are not enriched in CNSs suggest that the CNSs do not function as insulators.

Finally, we probed the expression conservation of protein-coding genes with respect to the number of flanking CNSs. Genes with more conserved expression would have a higher correlation coefficient. To do this, we calculated the correlation coefficient for each gene using RNA-Seq data of 12 human and mouse tissues from Necsulea et al. (2014). As shown in figure 6B, genes with more CNSs tend to have a higher expression correlation coefficient. Specifically, genes with no CNSs have the lowest expression correlation coefficient, while genes with at least 10 CNSs have the highest expression correlation coefficient. This shows the involvement of CNS in conserved gene expression of the neighboring genes.

Discussion

Using computational searches, we have identified ~20,000 CNSs that are conserved among chicken and four mammalian species. The conservation levels of the CNSs are significantly higher than those of random sequences and lincRNA exons. Purifying selection on CNSs is clearly stronger than observed in random sequences or lincRNAs. Interestingly, intragenic CNSs tend to have stronger constraint than their intergenic counterparts. In fact, there is overrepresentation of intragenic CNSs. For example, while the human genome gene percent is 41.54%, intragenic CNSs represent 47.76% of the total CNSs. For mouse, genome gene percent and intragenic CNS percent are 36.55% and 45.10%, respectively (these proportions are based on values given in table S1). This suggests that intragenic CNSs are more stable than intergenic ones.

There is an overrepresentation of intragenic CNSs, while intergenic CNSs tend to be located in clusters, far away from protein-coding genes. This clustering of CNSs in gene deserts suggests that the proximity to genes is not very important. In addition, among CNS-associated genes there is overrepresentation of certain gene categories. Specifically, genes associated with development,

transcription and nervous system, and/or genes expressed in embryonic brain tend to have more CNSs. Focusing on protein-coding genes, we found that genes associated with more CNSs tend to be located far away from other genes, demonstrating the bias in CNS-gene genomic location. In fact, we showed that the distance between CNSs and closest genes tends to be conserved between human and mouse genomes in terms of RDD measures. This suggests that the evolutionary constraints are acting on the genomic location of the CNSs with regard to the closest target genes. Specifically, genes that require more strict regulation may have to be located in such a genomic location as to allow controlled access to the regulatory regions.

The previously reported differences in gene expression patterns across tissues and stages might be explained, at least in part, by our results. For example, Khaitovich et al. (2005) analyzed the expression patterns of several tissues and reported that the expression divergence is lowest in brain and highest in liver. This pattern is also found in our analyses (figures S6A, B). Therefore, the observed differences in the expression divergences across tissues may be related to gene association with CNSs. As we have demonstrated, genes associated with more CNSs tend to have lower expression divergence (higher expression correlation) whereas genes associated with fewer CNSs tend to have higher divergence or lower expression correlation (figure 6B). Because genes expressed in brains tend to be enriched in CNSs, the observed lower divergence in brain expression could be attributable to CNS enrichment. The opposite explanations are true for testis and liver with CNS underrepresentation.

Our results highlight some differences between lincRNAs and CNSs. First, the sequence constraint on lincRNAs is much weaker than those on CNSs (figures 1 and S2, Tables S2 and S3). Second, while lincRNAs tend to be located close to the TSS of the next protein-coding genes, intergenic CNSs tend to be located far away from the TSS (figures. 2b and S3b). Third, the histone modification signals are different (figure S10B). Specifically, lincRNA H3k4me3 (promoter mark) is more than four-fold that of CNSs (figure S10B). This may reflect the transcriptional activity of lincRNAs or the overlap of many lincRNAs with protein-coding promoter regions. These differences suggest that CNSs do not function as lincRNAs. We then investigated the regulatory activity of CNSs. We compared RDDs of the functionally verified vista enhancer elements and that of the identified CNS. We discovered that distance conservation

of vista enhancer elements is comparable to those of our CNSs (figure 3). If the distance conservation is due to the regulatory activity, the comparable strength of distance conservation between functionally verified vista enhancer elements and our CNS dataset suggests that the majority of our the identified CNSs have regulatory function. For example, we checked the previously reported enhancers in *Pax6* locus (Bhatia et al. 2014) and found that three of our CNSs overlapped with Id855, agCNE1 and agCNE4. These three elements drive conserved expression in forebrain, trigeminal ganglia and hindbrain, respectively. In addition, we found that genes associated with CNSs tend to be under stronger purifying selection, as revealed by dN/dS ratio (figure S11). Comparing the distance conservation across selected GO terms, we found that CNS-gene RDDs of genes involved in nervous system, development and transcription tend be higher than CNS-gene RDDs without the ontology terms (figure S12). The level of significance of the difference is not as high in genes involved in response to stimulus, defense and immunity (see the p values in figure S12).

The importance of CNS genomic location is highlighted by the association between RDD values and gene expression dynamics. Specifically, figure 3A shows that genes with lower RDD value (higher CNS-gene distance evolutionary constraint) tend to have higher correlation coefficients than genes with higher RDD values. This result suggests that expression is more efficiently regulated across evolutionary timescales if the distance between CNSs and flanking genes is conserved. It would be interesting to experimentally demonstrate the significance of the genomic location in gene expression dynamics. Indeed, previous studies have hinted on the importance of the genomic location of regulatory elements in certain genes. For example, Webber et al. (1998) reported that the genomic location of enhancers is important in *Igf2* genes. In a recent report (Seruggia et al. 2015) using CRISPR–Cas9-mediated mutagenesis, the change of direction of enhancer was shown to result in the distortion of expression patterns with obvious coat color phenotype. The result provided indirect support for the importance of genomic location of CNSs in gene regulatory activity. Our evolutionary analyses suggest that the importance of CNS genomic location seems to be widespread.

Furthermore, using the ChIP-Seq data, we have shown that CNSs have higher enhancer signals than random coordinates, especially in embryonic brain. In fact, TFBS enrichment shows that CNSs are

enriched in many TFBSs. On the contrary, no enrichment was found for insulator element as revealed by CTCF binding sites. The expression patterns of CNS-proximal protein-coding genes showed unique properties, demonstrating that CNSs are preferentially located close to certain protein-coding genes. Our results suggest that even for long-range enhancer elements, the physically closest gene might be a target. For a case like that of *shh* enhancer in which the target gene of a CNS is not the closest gene, it is possible that such enhancers have multiple targets. Cases in which the proximal genes are not the target genes seem to be rare. Our results therefore suggest that the genomic location of a CNS is important for its regulatory function. This further implies that phenotypic changes could be observed from genomic deletion experiments even if the regulatory element is intact. If such deletion is large enough, the genomic location of the regulatory element with respect to the target gene may be affected, and that may produce certain phenotype changes.

Our genomic and evolutionary analyses have highlighted some important features of CNSs. Additionally, unique properties of CNS-proximal genes were revealed. We thus have demonstrated that the genomic location of CNSs with respect to genes is important for the proper gene regulation, and that evolutionary forces act to maintain the genomic location. The previously reported non-homologous location of CNSs with respect to genes (Babarinde and Saitou 2013) may be due to the change in gene structure (see figure S10 for one example case). While the actual genomic locations of CNSs are relatively fixed, changes in gene structure, such as exon loss or gain or *de novo* gene evolution may lead to such observed differential location. Because all CNSs are homologous, the numbers of intragenic CNSs should be similar across species. However, there is variation in the number (Table S1), implying that some intragenic CNSs in one species are intergenic in another species. This suggests that inter- or intragenic location of CNSs might not affect their function. Also, we found that intergenic CNSs are located far away from genes (figures 2B and S3B), suggesting that proximity may also not be important. However, we found evidence for conservation of CNS-gene distance. The conservation of CNS-gene distance may be due to the looping structure of DNA. In the loop, only certain regions could be brought into contact with the promoter regions. A regulatory element should therefore sit on such a genomic region that could be easily brought into contact with its gene promoter for effective regulation. While

too-close regions may be difficult to bend, the location of the CNS inside or outside the gene might not affect the looping very adversely. In conclusion, we have shown the importance of CNS genomic location and demonstrated that the CNSs are likely regulatory elements associated with conserved expression of the proximal genes.

Materials and Methods

Dataset used and CNS retrieval. The sources of the datasets are presented in the supplementary methods. For the retrieval of CNSs, repeats and coding regions of chicken, human, mouse, dog and cattle genomes were first masked. Pairwise homology searches with chicken as query against the four mammalian species were performed using BLASTN with E-value threshold of 10^{-5}. The resulting CNSs with at least 100bp length were 21,584 in chicken. The numbers of CNSs in human, mouse, dog and cow were 21,191, 21,026, 21,385 and 21,155, respectively. Because of the annotation quality and the availability of data, human and mouse CNSs were used for further analyses. Any CNS which overlaps any protein-coding gene (intron or UTR) was classified "intragenic", while others with no protein-coding gene overlap were classified "intergenic". The coordinates of the CNSs are in the Supplemental Information.

Derived allele frequency spectrum. The ancestral states of SNPs overlapping our datasets were parsimoniously determined (see Supplementary methods). The frequency of the derived alleles for each SNP position was extracted from the VCF file of Asian Population of 1000 Genome Projects. The distribution of the derived allele frequencies was computed for each category of coordinates.

CNS-gene association. For intragenic CNS, the CNS is assumed to be associated with the gene inside which it is found. For intergenic CNS, the CNS is assumed to be associated with the gene with the closest transcription start site (TSS).

Distance conservation. For this analysis, we used CNSs and protein-coding genes with only one copy in

human and mouse genomes. Cases in which the closest gene overlaps with another gene were discarded. We determined the distance between the CNS and the closest gene in human for examining the CNS-gene distance conservation. We also determined the distance between the orthologous CNS and the orthologous closest genes in the mouse genome. The difference between human and mouse difference of CNS-gene distance was normalized by the average of the distances to give the "relative distance difference" (RDD). As control, we also computed the distance conservation of similar numbers of closest pairs of protein-coding genes and vista enhancer elements (Visel et al. 2007). Because of the limited number, intragenic and intergenic vista enhancer elements were not separated. The RDD between human and mouse was computed with the following equation:

$$\mathrm{RDD} = \frac{|Xh - Xm|}{mean},$$

where *Xh* and *Xm* are the CNS-gene physical distance in the human genome and the mouse genome, respectively, and *mean* = (*Xh*+*Xm*)/2. Normalized difference of the TSS genomic coordinates of the two adjacent protein-coding genes is defined as the gene-gene distance. Computation of RDD was also done for human-dog and human-cow pairs.

To investigate the relationship between RDD and expression correlation, CNS-associated genes were ranked based on human-mouse RDD values. For genes associated with multiple CNSs, the median RDD value was used. For each gene, Spearman's expression was calculated from twelve tissues (see Table S4) with expression data in human and mouse. The average expression coefficient for each RDD-ranked quartile group was computed.

Retrieval of housekeeping genes. We followed the definition of housekeeping genes reported by Eisenberg and Levanon (2013). For this analysis, we used the expression data of Necsulea et al. (2014). The criteria used to define housekeeping genes as reported by Eisenberg and Levanon (2013) include; (i) expression observed in 15 tissues; (ii) low variance over tissue expression values: standard-deviation [log2(RPKM)]<1; and (iii) no exceptional expression in any single tissue; that is, no log-expression value differed from the averaged log2(RPKM) by two (fourfold) or more. This returned 4,161 genes in human

and 3,902 genes in mouse.

Gene enrichment test. We first made a list of all genes in GO terms with A_{total} elements. We then made a list of CNS-associated genes with GO terms with A_{CNS} elements. Genes associated with multiple CNSs were represented multiple times in the CNS-associated gene list. For each tested GO term, we counted the number (T_{CNS}) of CNS-associated genes. We also counted the number (T_{total}) of the genes with the term in the all gene list. The expression for the enrichment is given as:

$$\text{Enrichment (fold change)} = (T_{CNS} \times A_{total}) / (A_{CNS} \times T_{total}).$$

Statistical significance was calculated with a binomial test with Bonferroni correction as implemented in Panther (Thomas et al. 2003).

Modified CGN (modCGN) computation. ModCGN was modified from CGN originally proposed by De et al. (2009). The equation of modCGN in this study is given as;

$$\text{modCGN} = \frac{Number\ of\ human\ genes\ conserved\ in\ a\ window}{Average\ number\ of\ human\ and\ mouse\ genes\ in\ the\ window}.$$

The detail of the computation is provided in Supplementary methods.

Expression conservation. We used twelve tissues (see Table S2) with expression data in human and mouse. For every gene with one-to-one orthology in human and mouse, we computed the Spearman's correlation coefficient from the expression data of the twelve tissues. For each category (0CNS, 1-3CNSs, 4-9CNSs, >9CNSs) of genes, we calculated the average Spearman's correlation coefficient.

Acknowledgements

This study was partially supported by a grant-in-aid from Japan Society for the Promotion of Science to N.S. I.A.B. received MEXT scholarship during this study. The authors appreciate comments from Dr. Tim Jinam, Ms. Nilmini Hettiaracchi, Mr. Mamoudisaber Morteza, Ms. Shayire Shokat, Dr. Shiroishi Toshihiko, Dr. Fujiyama Asao, Dr. Kitano Jun and Dr. Nozawa Masafumi. Some of the analyses were

performed on NIG supercomputer.

References

Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, Myeroff L, Lutterbaugh J, Jarrar A, Kalady MF, et al. 2005. Epigenomic enhancer profiling defines a signature of colon cancer. Science 336: 736-739.

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

Antoniv TT, De VS, Wells D, Denton CP, Rabe C, de Crombrugghe B, Ramirez F, Bou-Gharios G, et al. 2001. Characterization of an evolutionarily conserved far-upstream enhancer in the human alpha 2(I) collagen (COL1A2) gene. J Biol Chem. 276:21754-21764.

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

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37:W202-8.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved Elements in the Human Genome. Science 304:1321-1325.

Bhatia S, Monahan J, Ravi V, Gautier P, Murdoch E, Brenner S, van Heyningen V, Venkatesh B, Kleinjan DA. 2014. A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. Dev Biol. 387:214-228.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene expression differences among primates are associated with changes in a histone epigenetic modification. Genetics 187:1225-1234.

Calle-Mustienes E, Feijóo CG, Manzanares M, Tena JJ, Rodríguez-Seguel E, Letizia A, Allende ML, Gómez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Res. 15:1061-1072.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short

introns in highly expressed genes. Nat Genet. 31:415-418.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA 107:21931-21936.

De S, Teichmann SA, Babu MM. 2009. The impact of genomic neighborhood on the evolution of human and chimpanzee transcriptome. Genome Res. 19: 785-794.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing Chromosome Conformation. Science 295:1306-1311.

Dostie J, Dekker J. 2007. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc. 2:988-1002.

Drake JA, Bird C, NemeshJ, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn1 JN. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 38:223-227.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. Trends Genet. 19:362-365.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. Trends Genet. 29:569–574.

Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV. 1996. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol Cell Biol. 16:2802-13.

Göttgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M,, et al. 2000. Analysis of vertebrate SCL loci identifies conserved enhancers. Nat Biotechnol. 18:181-186.

Herold M, Bartkuhn M, Renkawitz R. 2012. CTCF: insights into insulator function during development. Development. 139:1045-57.

Hou C, Zhao H, Tanimoto K, Dean A. 2008. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc Natl Acad Sci U S A. 105:20398-403.

Irimia M, Tena JJ, Alexis MS, Fernandez-Miñan A, Maeso I, Bogdanovic O, de la Calle-Mustienes E,

Roy SW, Gómez-Skarmeta JL, Fraser HB. 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Res. 22:2356-2367.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S. 2005. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science 309:1850-4.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465:182-187.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12: 1725-1735.

Matsunami M, Saitou N. 2013. Vertebrate paralogous conserved noncoding sequences may be related to gene expressions in brain. Genome Biol Evol. 5:140-50.

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Greg Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. PLoS Genet. 5(12).

McLeay RC, Bailey TL. 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics. 11:165.

Morris JR, Geyer PK, Wu C. 1999. Core promoter elements can regulate transcription on a separate chromosome in trans. Genes Dev. 13:253-258.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505:635-640.

Ong CT, Corces VG. 2009. Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. J. Biol 8:73.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res. 15:137-145.

Rao YS, Wang ZF, Chai XW, Wu GZ, Zhou M, Nie QH, Zhang XQ. 2010. Selection for the compactness

of highly expressed genes in Gallus gallus. Biol Direct 5:35.

Sagai T, Amano T, Tamura M, Mizushina Y, Sumiyama K, Shiroishi T. 2009. A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. Development 136, 1665-1674.

Sagai T, Masuya H, Tamura M, Shimizu K, Yada Y, Wakana S, Gondo Y, Noda T, Shiroishi T. 2004. Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh). Mamm Genome 15:23-34.

Seruggia D, Fernández A, Cantero M, Pelczar P, Montoliu L. 2015. Functional validation of mouse tyrosinase non-coding regulatory DNA elements by CRISPR-Cas9-mediated mutagenesis. Nucleic Acids Res. 43:4855-67.

Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature 488:116-20.

Sumiyama K, Irvine SQ, Stock DW, Weiss KM, Kawasaki K, Shimizu N, Shashikant CS, Miller W, Ruddle FH. 2002. Genomic structure and functional control of the Dlx3-7 bigene cluster. Proc Natl Acad Sci USA 99:780-785.

Sumiyama K, Ruddle FH. 2003. Regulation of Dlx3 gene expression in visceral arches by evolutionarily conserved enhancer elements. Proc Natl Acad Sci U S A. 100:4030-4034.

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

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

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 13:2129-2141.

Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. 2006. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet. 22:5-10.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 19:327-335.

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

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854-858.

Webber AL, Ingram RS, Levorse JM, Tilghman SM. 1998. Location of enhancers is essential for the imprinting of H19 and Igf2 genes. Nature 391:711-5.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3: e7.

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al. 2006. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 38, 1341-1347.

Figure Legends

Fig. 1. Retrieved CNSs have strong constraints in the human genome. (A) CNSs have the highest phastcons score. (B) CNSs have higher phylop conservation score than random sequences. (C) CNSs are under purifying selection, and not mutational cold spots. (Chi square p value < 0.001). Error bars are 99.99% CI from ten independent random samplings. The differences of the median values are statistically significant (Mann-Whitney-U p value < 0.001).

Fig. 2. Nonrandom genomic location of CNSs in the human genome. (A) Compared to random sequences and lincRNAs, CNSs tend to exist in genomic clusters. (B) Intergenic CNSs tend to be located far away from protein-coding genes. Error bars are 99.999% CI from ten independent random samplings. For the two charts, chi square p value < 0.001.

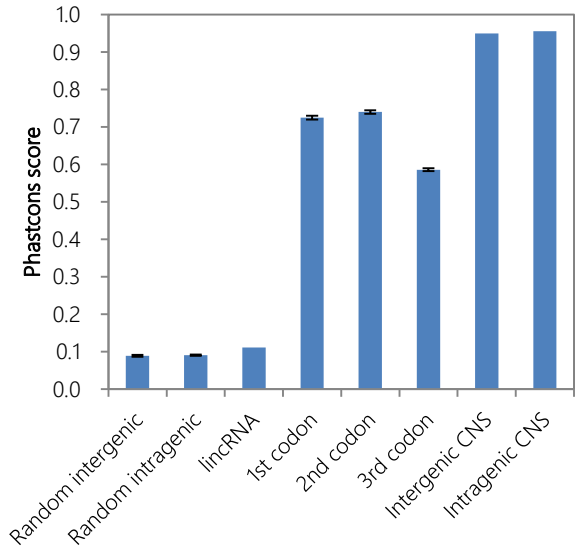
Fig. 3. The importance of CNS-gene distance. (A) The CNS-gene distance is evolutionarily more conserved than the gene-gene distance (Chi square p value < 0.001). RDD measures the relative difference between human CNS-gene or gene-gene physical distance and mouse orthologous distance. (B) Genes with lower RDD values tend to have higher expression correlation. The genes with CNSs (n = 2,847) were ordered by RDD values, and were divided into quartiles. * p value < 0.05; ** p value < 0.01 (Mann-Whitney U test).

Fig. 4. Enrichment of CNS-associated genes. (A) Enrichment test of selected gene ontology terms. (B) Enrichment test of genes expressed in certain tissues. The threshold expression level was 5RPKM. All the enrichment directions were statistically significant (Bonferroni corrected binomial p value < 0.001).

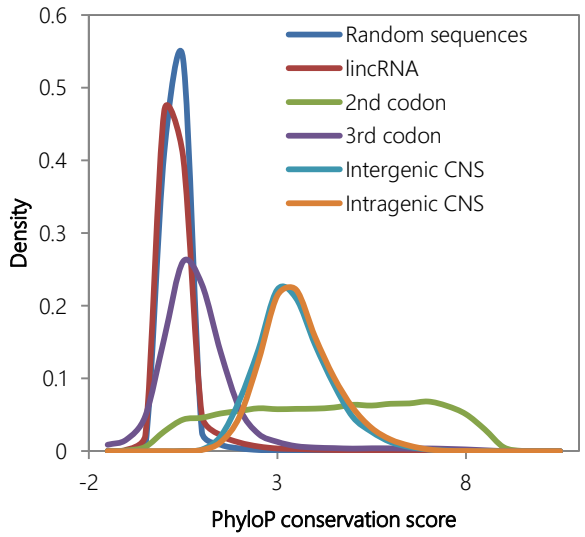
Fig. 5. Unique features of genes associated with CNSs. (A) Genes with no associated CNSs have lower noncoding percent (*** T-test p value < 0.001). (B) Genes with more CNSs tend to be located far away from other genes (*** Mann-Whitney U test p value < 0.001). (C) Genes with more CNSs are located in more conserved genomic neighborhoods (*** Mann-Whitney U test p value < 0.001).

Fig. 6. CNSs are associated with more conserved expression. (A) CNSs have stronger signals for H3k4me1 (enhancer) than random sequences. The difference is more obvious in fetal brain tissues. Tissues were classified into four groups based on stage and association with brain. The number of tissues in each group is shown in brackets. The error bars show the 95% confidence interval. (B) Genes with more CNSs tend to have more conserved expression. Error bars are 99.999% CI from the genes for which correlation coefficient was calculated.

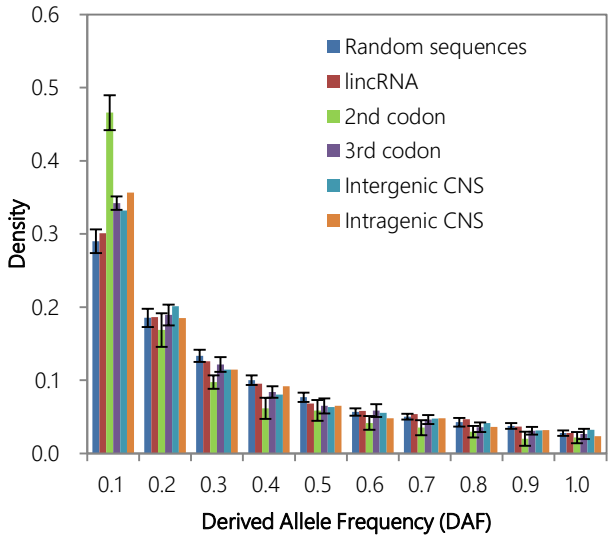
A


Phastcons score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Random intergenic
Random intragenic
lincRNA
1st codon
2nd codon
3rd codon
Intergenic CNS
Intragenic CNS

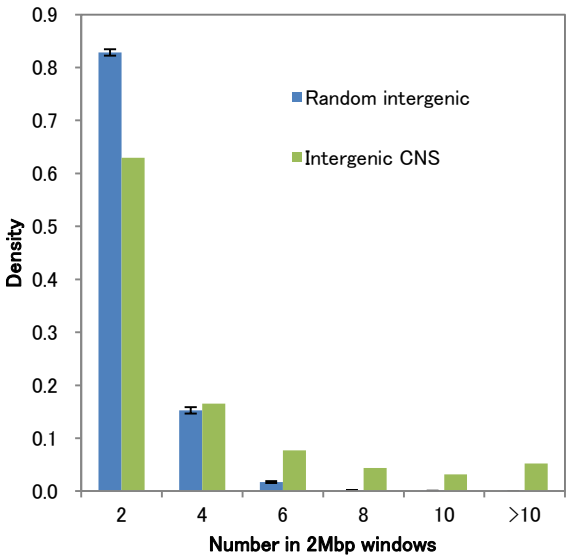
B


Density
0.6
0.5
0.4
0.3
0.2
0.1
0
-2
3
8
PhyloP conservation score
Random sequences
lincRNA
2nd codon
3rd codon
Intergenic CNS
Intragenic CNS

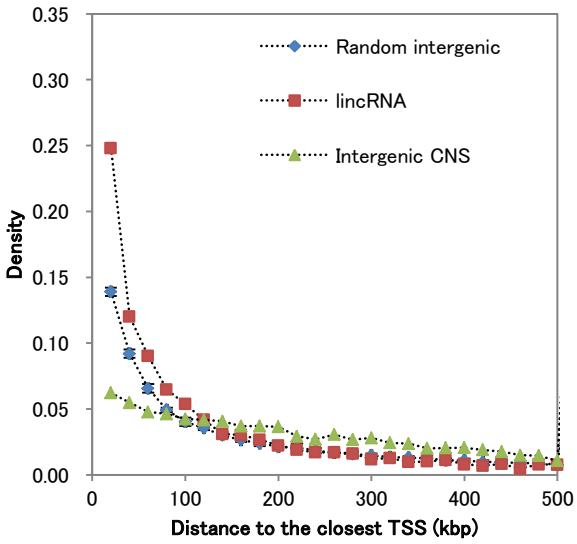
C


Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Derived Allele Frequency (DAF)
Random sequences
lincRNA
2nd codon
3rd codon
Intergenic CNS
Intragenic CNS

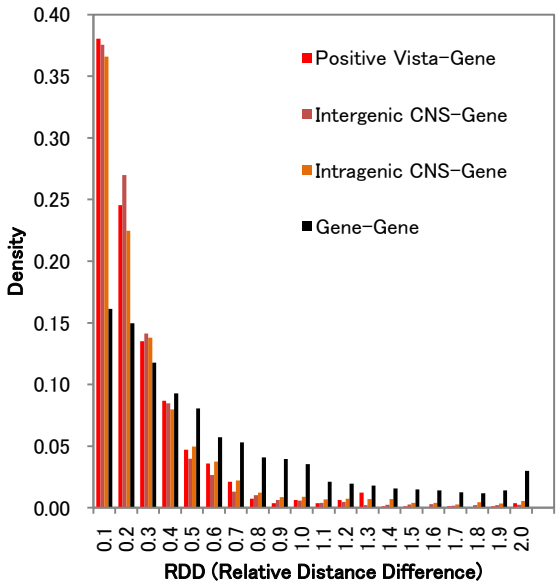
A


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
Random intergenic
Intergenic CNS
2
4
6
8
10
>10
Number in 2Mbp windows

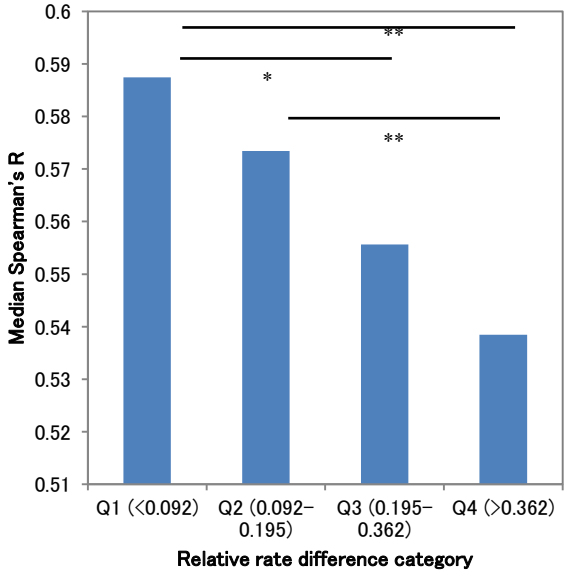
B


0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
Random intergenic
lincRNA
Intergenic CNS
0
100
200
300
400
500
Distance to the closest TSS (kbp)

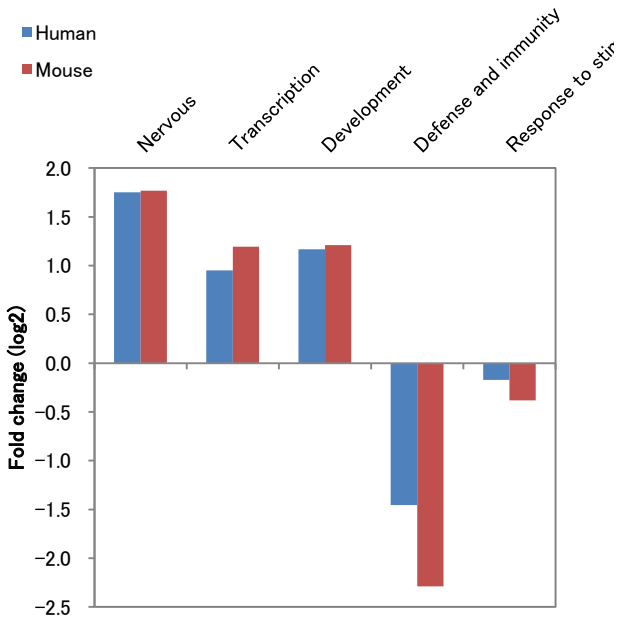
A


0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Density
Positive Vista-Gene
Intergenic CNS-Gene
Intragenic CNS-Gene
Gene-Gene
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
RDD (Relative Distance Difference)

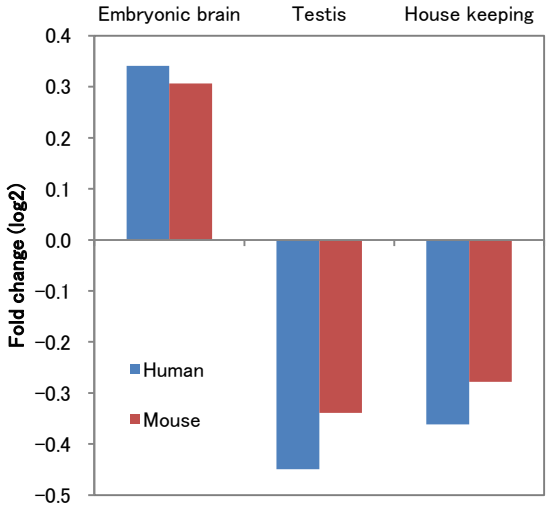
B


0.6
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
Median Spearman's R
**
*
**
Q1 (<0.092)
Q2 (0.092-0.195)
Q3 (0.195-0.362)
Q4 (>0.362)
Relative rate difference category

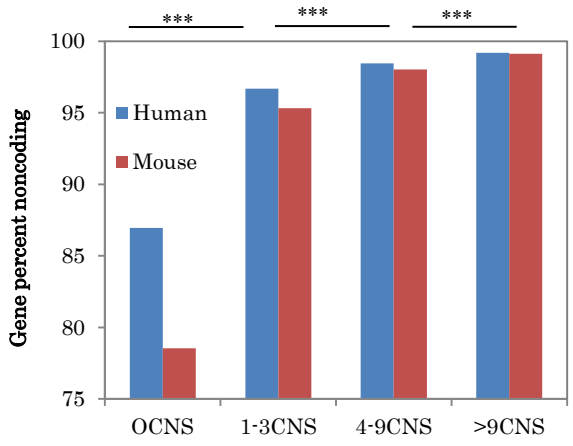
A


Human
Mouse
Nervous
Transcription
Development
Defense and immunity
Response to stir
Fold change (log2)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5

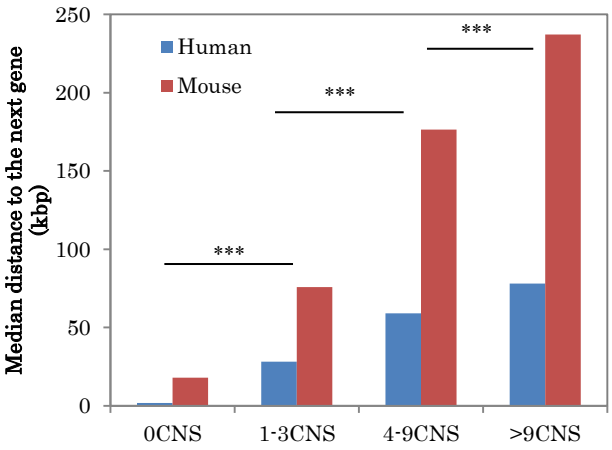
B


Embryonic brain
Testis
House keeping
Fold change (log2)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
Human
Mouse

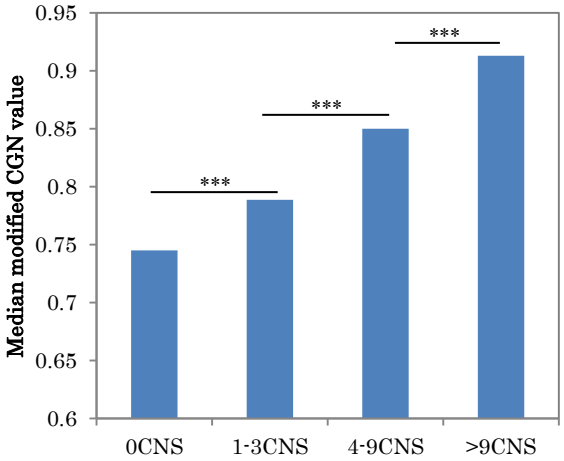
A


***
***
***
100
95
90
85
80
75
Human
Mouse
Gene percent noncoding
OCNS
1-3CNS
4-9CNS
>9CNS

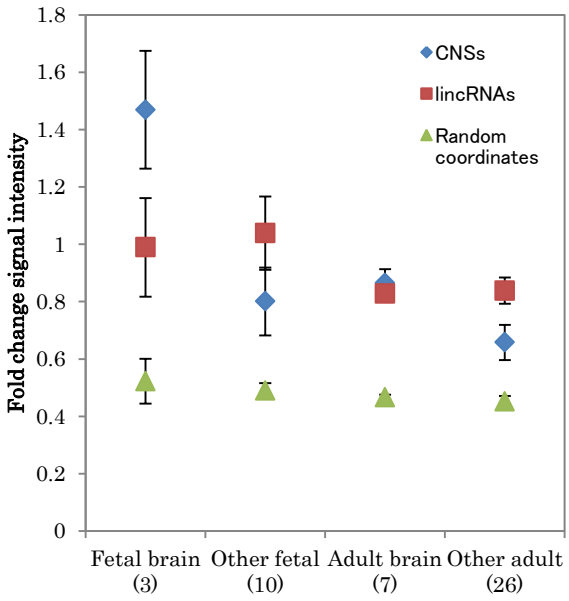
B


250
200
150
100
50
0
Human
Mouse
***
***
***
Median distance to the next gene (kbp)
0CNS
1-3CNS
4-9CNS
>9CNS

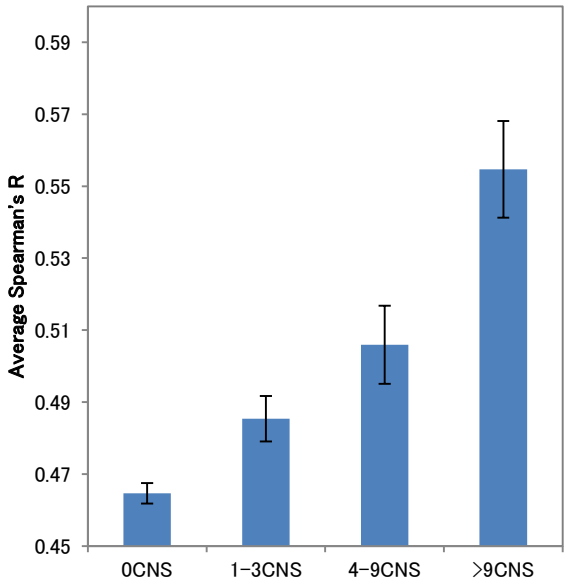
C


0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
***
***
***
Median modified CGN value
0CNS
1-3CNS
4-9CNS
>9CNS

A


1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Fold change signal intensity
CNSs
lincRNAs
Random coordinates
Fetal brain (3)
Other fetal (10)
Adult brain (7)
Other adult (26)

B


0.59
0.57
0.55
0.53
0.51
0.49
0.47
0.45
Average Spearman's R
0CNS
1–3CNS
4–9CNS
>9CNS