Functional Implications of RNA Splicing for Human Long Intergenic Noncoding RNAs

Feng-Chi Chen[1–3], Chia-Lin Pan[1] and Hsuan-Yu Lin[1]

[1]Institute of Population Health Sciences, National Health Research Institutes, Taiwan. [2]Department of Biological Science and Technology, National Chiao-Tung University, Taiwan. [3]Department of Dentistry, China Medical University, Taiwan.

ABSTRACT: Long intergenic noncoding RNAs (lincRNAs) have been suggested as playing important roles in human gene regulation. The majority of annotated human lincRNAs include multiple exons and are alternatively spliced. However, the connections between alternative RNA splicing (AS) and the functions/regulations of lincRNAs have remained elusive. In this study, we compared the sequence evolution and biological features between single-exonic lincRNAs and multi-exonic lincRNAs (SELs and MELs, respectively) that were present only in the hominoids (hominoid-specific) or conserved in primates (primate-conserved). The MEL exons were further classified into alternatively spliced exons (ASEs) and constitutively spliced exons (CSEs) for evolutionary analyses. Our results indicate that SELs and MELs differed significantly from each other. Firstly, in hominoid-specific lincRNAs, MELs (both CSEs and ASEs) evolved slightly more rapidly than SELs, which evolved approximately at the neutral rate. In primate-conserved lincRNAs, SELs and ASEs evolved slightly more slowly than CSEs and neutral sequences. The evolutionary path of hominid-specific lincRNAs thus seemed to have diverged from that of their more ancestral counterparts. Secondly, both of the exons and transcripts of SELs were significantly longer than those of MELs, and this was probably because SEL transcripts were more resistant to RNA splicing than MELs. Thirdly, SELs were physically closer to coding genes than MELs. Fourthly, SELs were more widely expressed in human tissues than MELs. These results suggested that SELs and MELs represented two biologically distinct groups of genes. In addition, the SEL–MEL and ASE–CSE differences implied that splicing might be important for the functionality or regulations of lincRNAs in primates.

KEYWORDS: long intergenic noncoding RNA, alternative splicing, sequence evolution, gene regulation

Introduction

Alternative RNA splicing (AS) is an essential post-transcriptional regulatory mechanism in practically all of the fundamental biological processes in complex organisms, such as gene regulation, metabolism, development, cell cycle control, immune response, signal transduction, and human diseases.[1–7] By definition, AS targets only multi-exonic genes, which are present in multicellular organisms but rarely observed in unicellular organisms. The vast majority of the genes in unicellular genomes contain only one single exon.[8,9] AS is thus arguably a characteristic of multicellular organisms, and an important source for functional diversity and evolutionary novelty.[6,10–12] However, previous AS-related studies have been focused on coding genes. The implications of AS in the functions/regulations of noncoding genes have been underexplored. By analyzing the gene structures of long intergenic noncoding RNAs (lincRNAs) and comparing multi-exonic lincRNAs (MELs) and single-exonic lincRNAs (SELs), we might be able to gain some insights into this important issue.

In complex organisms, single-exonic coding genes and multi-exonic coding genes ("SECs" and "MECs," respectively)

differ significantly from each other in evolutionary history, gene regulation, and molecular function. Evolutionarily, SECs in mammals were reported to have relatively recent origins as compared with MECs. Most of the SECs were found only in Chordata,[13] and some found only in mammals.[14] This observation stands in interesting contrast to the notion that SECs are characteristic of the unicellular genomes.[13] In mammals, SECs were found to evolve faster than MECs,[13] and to have undergone lineage-specific expansions.[14] These two groups of genes also differ in terms of gene regulation. SECs tend to be lowly and tissue-specifically expressed as compared with MECs.[13] In addition, SECs were reported to be resistant to mis-transcription because they tended to avoid "fragile codons," which when mutated could lead to nonsense-mediated decay.[15] Although the majority of human genes are multi-exonic, a number of functionally important genes, such as histones and the G protein-coupled receptors, are mostly composed of only one exon.[16] SECs and MECs also differ from each other in terms of biological function. SECs were reported to be enriched in functional categories such as transport and binding, cell envelope, protein translation, energy metabolism, amino acid biosynthesis, and other regulatory functions.[14]

For noncoding genes, however, the above-mentioned differences may not apply. This is because noncoding RNAs are free from the selective constraints imposed on coding genes, such as the conservation of reading frame and functional protein domains. Indeed, noncoding genes have been reported to be evolving rapidly.[17] Furthermore, the exon boundaries of lincRNAs have been recently suggested to turn over rapidly, and be unimportant for the transcriptional regulations of these genes in mammals.[18] Interestingly, the tissue specificities of lincRNA expression were found to be well conserved in mammals[18] and tetrapods[19] despite the rapid turnover of lincRNA exon boundaries. These observations imply that splicing might play a less important role in generating the functional forms of lincRNA genes than in the case of coding genes. Notably, nevertheless, the majority of the annotated lincRNA genes include multiple exons, and a considerable proportion of the multi-exonic lincRNA genes are alternatively spliced.[20] The notion that splicing is unimportant for lincRNA functions implies that MELs and SELs might be biologically similar to each other. Furthermore, if splicing is biologically meaningless for lincRNAs, splicing of a lincRNA should occur stochastically (rather than actively regulated). In other words, the probability that an SEL is spliced should be similar to that of an MEL given similar lengths and sequence compositions. In addition, if splicing is functionally irrelevant, the sequences of alternatively spliced exons and constitutively spliced exons ("ASEs" and "CSEs," respectively) of MELs should have been subject to similar levels of selective constraint, and evolved at similar rates. These propositions, however, have not been examined.

lincRNAs can regulate the transcriptions of coding genes through a number of different mechanisms, including transcriptional interference, activation of transcription factors/repressor proteins, recruitment of epigenetic modifiers, and induction of chromosomal looping.[21,22] These are fairly different regulatory mechanisms, yet most of them require the interactions between lincRNAs and other biological molecules – DNA, RNA, and/or proteins. These interactions are associated with the biological features of lincRNAs. One good example is secondary structure. Properly folded lincRNAs may serve as scaffolds to orient heterogeneous biological molecules for interactions.[23] Secondary structure is in turn affected by other sequence features such as the length and G+C content of a lincRNA. To be sure, the primary sequence per se may also be important for the functions of lincRNAs.[23] Another biological feature of lincRNAs is proximity to coding genes. This feature is functionally relevant because the transcription of a noncoding RNA could interfere with the transcription of its nearby coding gene.[22] Meanwhile, the breadth of lincRNA gene expression is also functionally informative. Expression breadth is a traditional measurement of functional importance. Widely expressed genes are regarded as biologically more important, and evolve more slowly than narrowly expressed genes.[24,25] To investigate the functional differences between SELs and MELs, it is necessary to compare these two groups of lincRNAs in view of these biological features.

In this study, we systematically compared Encyclopedia of DNA Elements (ENCODE)-annotated human SELs and MELs. Our results suggest that SELs and MELs diverge from each other in primary sequence conservation, exon/transcript length, proximity to coding genes, and expression breadth. These results imply that SELs and MELs represent two functionally distinct gene groups. Furthermore, SELs were found to be resistant to RNA splicing, and primate-conserved ASEs evolved more slowly than CSEs. We thus suggest that splicing plays an unknown yet influential role in the functions of lincRNA.

Materials and Methods

LincRNA sequences and identification of homologous sequences. A total of 13,249 human lincRNA genes, which included 22,531 transcripts and 71,864 exons, were retrieved from the ENCODE data portal at the University of California, Santa Cruz (UCSC) Genome Browser (https://genome.ucsc.edu/ENCODE). The retrieved sequences corresponded to the ENSEMBL Version 70 annotations. The human lincRNAs that were shorter than 80 nts were excluded to minimize the variations of the subsequent calculations of genetic distance. The potential nonhuman homologous exonic sequences were identified with reference to the human–nonhuman pairwise genome alignments from the UCSC Genome Browser. In cases where one human exon was aligned to multiple nonhuman genomic regions, only the exons that were in the correct exonic synteny were retained. The nonhuman sequences that were of low quality (ie, those contained "Ns") or located in uncertain chromosomal locations (eg, ChrUn) were discarded. The exonic

sequences that overlapped with either human or nonhuman coding genic regions were also removed. The nonhuman homologous sequences were examined for the presence of canonical exon boundaries (GU/AG or GC/AG), which were required to be located not farther than 10% of the human exon length from the human exon boundary according to the pairwise alignment. The homologous sequences lacking one or both exon boundaries were discarded. The exons of multi-exon genes were classified into CSEs and ASEs in cases of multiple transcript isoforms. CSEs were the exons that were always present in all the transcript isoforms of a gene. ASEs were those that were included in only some of the transcript isoforms of a gene.

We started our analysis with 13,249 ENSEMBL-annotated human lincRNA genes (Version 70), which corresponded to 22,531 transcripts and 71,864 exons. We then examined whether the orthologous exonic regions and the corresponding exon boundaries could be found separately in the genome of orangutan or rhesus macaque. We did not analyze the lincRNAs conserved in mouse because the number of human–mouse conserved lincRNAs was fairly small (only tens of genes), and the alignments were of low quality. If the ortholog of a human exon was absent in the compared genome, the corresponding transcript was discarded. (From the human–orangutan (H-O) orthologs, we removed the lincRNAs whose orthologs were present in the rhesus macaque to retain only hominoid-specific lincRNAs. The two datasets (H-O and human–macaque or "H-Ma") therefore represent the lincRNAs that were hominoid-specific and primate-conserved, respectively. The H-O and H-Ma datasets included 1,664 and 4,332 genes, which in turn included 2,235 (2,025 multi-exonic and 210 single-exonic) and 6,198 (5,763 multi-exonic and 435 single-exonic) transcripts, respectively (Table 1). Similar procedures were applied to the mouse–rat comparison, yielding 1,069 MELs (including 2,418 CSEs and 821 ASEs) and 21 SELs.

Estimates of sequence conservation. The H-O and H-Ma genetic distances of orthologous exonic sequences were calculated by using the baseml module of PAML4 (with the HKY 85 substitution model) on the basis of the UCSC pairwise sequence alignments subject to the filters mentioned in the last section. We also retrieved 300-bp intergenic sequences 5 kb upstream and downstream of the studied lincRNAs, and calculated their H-O/H-Ma distances for comparison. The genetic distances of pure introns (the sequences that were annotated as introns in all of the transcript isoforms of a lincRNA according to the ENSEMBL annotations) were also computed. The lengths of the pure introns were also limited to 300 bp, which was close to the average exon length of 345 bp in the H-Ma dataset. The genetic distances of these nonexonic regions served as the "neutral reference."

Estimates of the distance between lincRNAs and coding genes/enhancers. The nearest coding gene (or enhancer) from a lincRNA was identified with reference to the ENSEMBL human gene annotations. The distance between a lincRNA and a coding gene was defined as the smallest number of nucleotides among the four possible combinations of gene boundaries (5' coding–3' lincRNA, 5' coding–5' lincRNA, 3' coding–3' lincRNAs, and 3' coding–5' lincRNA). Similarly, the distance between a lincRNA and an enhancer was defined as the smallest number of nucleotides among the four possible combinations of lincRNA gene boundaries and the terminals of the enhancer of interest.

Exonic expression level and breadth. The RNA-sequencing (RNA-seq) data of 16 human tissues (E-MTAB-513) were retrieved from the ArrayExpress website at https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/. In this dataset, each tissue was retrieved from a human individual. The transcriptome of each tissue was subject to both single-end and paired-end RNA-seq. The short reads were mapped to the human genome using STAR,[26] and the fragments per kilobase transcript per million mapped reads for the studied lincRNAs were derived using Cufflinks.[27] We then took the average of single- and paired-end results for each tissue as the expression level of a lincRNA. The expression breadth of a lincRNA was defined as the proportion of the 16 tissues where the lincRNA was expressed.

Prediction of the secondary structure of lincRNAs. The secondary structure of each lincRNA transcript was predicted by using RNAfold.[28] To evaluate the influences of transitional mutations on the folding energy of lincRNAs, each individual nucleotide of a lincRNA transcript was subject to transitional mutation *in silico*, and the resulting change in folding energy (ΔE) was calculated based on the predictions of RNAfold. The average ΔE of each exonic region was calculated for the evaluation of structural constraint on the exonic sequence of interest.

Table 1. The H-O and H-Ma conserved lincRNAs identified in this study.

	HUMAN-ORANGUTAN		HUMAN-MACAQUE	
	MULTI-EXONIC	**SINGLE-EXONIC**	**MULTI-EXONIC**	**SINGLE-EXONIC**
# Genes	1,664	210	4,332	435
# Transcripts	2,025	210	5,763	435
# CSEs	3,569	NA[a]	9,568	NA
# ASEs	1,835	NA	5,231	NA

Note: [a]Not Applicable.

Prediction of splicing sites. The human lincRNA transcript sequences were submitted to the SplicePort web server[29] for identification of putative donor and acceptor sites with default parameters. The numbers of putative donor/acceptor sites were then normalized by the length of the transcript.

Results

Differential sequence evolution between SELs and MELs in primates. We first examined whether the primary sequences of SELs and MELs were subject to different levels of selective constraint. A recent report indicated that mammal-conserved lincRNAs tended to be SELs.[18] We were thus interested to know whether SELs were more conserved than MELs for evolutionarily recent (H-O and H-Ma) lincRNAs. We also classified the exons of MELs into CSEs and ASEs to examine whether splicing was associated with the evolution of lincRNAs. For H-O lincRNAs, unexpectedly, CSEs and ASEs had slightly larger genetic distances than pure introns and intergenic regions. Meanwhile, the H-O genetic distances of SELs were approximately the same as those of the non-exonic regions (Fig. 1A). This observation seemed to suggest that SELs were selectively neutral, whereas the evolution of CSEs and ASEs was slightly accelerated.

In comparison, for H-Ma lincRNAs, ASEs and SELs were slightly more conserved than pure introns and intergenic regions, but CSEs were similar to the non-exonic regions in this regard (Fig. 1B). Of note, for H-O lincRNAs, SELs were more conserved than both CSEs and ASEs, although not more conserved than non-exonic regions. Yet for H-Ma lincRNAs, SELs were approximately as conserved as ASEs, and both of the exon groups were more conserved than CSEs and non-exonic regions. The results based on the H-Ma dataset appeared to suggest that ASEs and SELs were subject to weak selective constraint, while CSEs were selectively neutral.

These results were different from what were derived from the H-O dataset, implying that the exons of hominoid-specific and primate-conserved lincRNAs had been evolving along different paths since the emergence of the hominoid lineage. Despite the differences between the results of H-O and H-Ma lincRNAs, one common theme was that SELs were more conserved than CSEs but not than ASEs (Fig. 1). The difference between Figures 1A and B appears to result mainly from the relative level of sequence conservation between lincRNAs and the reference neutral sequences. In Figure 1B (as compared with Fig. 1A), the genetic distances of lincRNA exons (ASEs, CSEs, and SELs) seem to have shifted downwards relative to those of intronic and intergenic regions. The possible reasons for this difference will be discussed later.

To examine whether the differences between SELs and MELs also apply to other mammalian species, we calculated the corresponding genetic distances between mouse and rat. As shown in Supplementary Figure 1, in the mouse–rat comparison, all of ASEs, CSEs, and SELs were more conserved than the neutral regions (introns and intergenic regions). ASEs were slightly more conserved than CSEs. Meanwhile, although SELs appeared to be slightly more conserved than ASEs and CSEs, the differences were statistically insignificant, probably because of the small sample size (there were only 21 SELs in this analysis). Overall, the results derived from the mouse–rat comparison (Supplementary Fig. 1) were similar to those obtained in the H-Ma comparison (Fig. 1B).

We also calculated the percentage of canonical splice sites (GT/GC-AG) that were conserved between the compared species for lincRNA exons. Since exon boundaries might have been shifted during evolution, we searched a sequence space of 10% of the exon length upstream and downstream from each exon boundary for potential homologous splice sites in
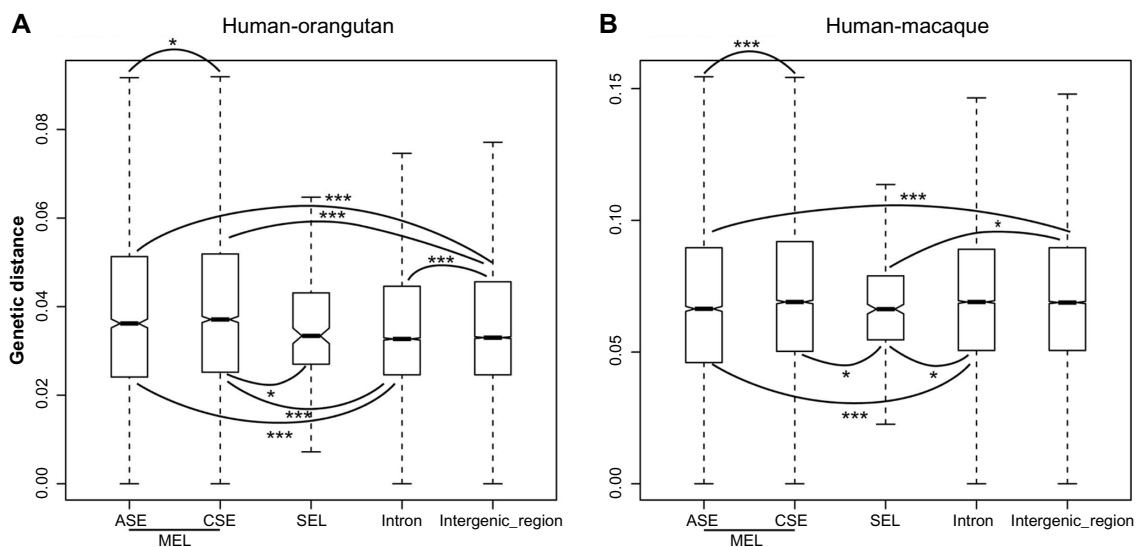

Figure 1. The genetic distances in lincRNA exons between (**A**) human and orangutan; and (**B**) human and rhesus macaque.
Notes: Statistical significance: *$P < 0.05$; ***$P < 0.001$ by Wilcoxon Rank Sum test.

the non-human genome with reference to pairwise alignments. Our results indicated that the homologous splice sites of nearly all (99.94% in H-O and 99.94% in H-Ma) of the human exon boundaries could be found in the search space in the compared genomes. Therefore, the exon boundaries of lincRNAs were highly conserved across primates in view of genomic sequence. This result was consistent with what had been previously reported.[18]

SEL transcripts were significantly longer and more resistant to splicing than MEL transcripts. Next, we compared the lengths of CSEs, ASEs, and SELs. As shown in Figure 2A, SELs were seven to eight times and five to six times longer than MEL exons in the H-O and H-Ma dataset, respectively. In terms of "transcript length," SELs were about twice (1.9–2.3) longer than MELs in both of the datasets (Fig. 2B). In other words, one single exon of an SEL was in general twice longer than the combined length of multiple exons of an MEL.

We then asked why SELs were significantly longer than the MEL transcripts. One possibility was that SEL transcripts contained fewer putative splice sites than the primary transcripts of MELs. To examine this possibility, we used an integrated splice site prediction tool, SplicePort,[29] to identify putative donor and acceptor sites in the primary transcripts (exonic plus intronic regions) of MELs and SELs. Indeed, the length-normalized numbers of putative donor and acceptor sites were significantly smaller in SELs in both of the H-O and H-Ma datasets (Fig. 3). Even if we considered the absolute numbers of predicted splicing sites (ie, regardless of transcript length), SELs still had fewer donor and acceptor sites than MELs (Supplementary Fig. 2). This difference was remarkable considering that SELs were approximately twice longer than MELs (Fig. 2). Furthermore, this difference did

not result from the difference in G+C content, for the G+C contents were very similar between SELs and MELs (median G+C%: H-O MEL – 44.0%; H-O SEL – 43.5%; H-Ma MEL – 43.0%; H-Ma SEL – 43.5%). This observation seemed to suggest that splicing sites were disfavored by selection in SELs as compared with in MELs.

SELs were physically closer to the nearest coding genes than MELs. LincRNAs have been found to be capable of affecting the transcription of their neighboring coding genes via transcriptional interference.[30–33] The distance of a lincRNA from its neighboring coding gene therefore is an important feature. Interestingly, as shown in Figure 4A, SELs were significantly closer to coding genes than MELs. The median distance between an SEL and its closest coding gene was 3,503 bp for the H-O dataset, and 4,213 bp for the H-Ma dataset. In comparison, the corresponding distances for MELs were 10,772 bp and 16,339 bp, respectively. Of note, the SELs and MELs in the H-O dataset were closer to coding genes than their H-Ma counterparts. It appeared that more recent lincRNAs tended to be located in the neighborhood of coding genes. We thus conducted Spearman's correlation analyses between H-O (or H-Ma) genetic distance and the physical distance to nearest coding gene separately for SELs and MELs. As shown in Supplementary Figure 3, the correlation was statistically significant only for MELs in the H-O dataset. However, the effect size was fairly small (rho = 0.085). The distance from a lincRNA gene to the nearest coding gene thus did not seem to be significantly correlated with evolutionary age in general.

We also examined whether SELs and MELs differed from each other in their distance to the nearest enhancer. Interestingly, SELs again were significantly closer to enhancers than MELs in both of the H-O and the H-Ma datasets
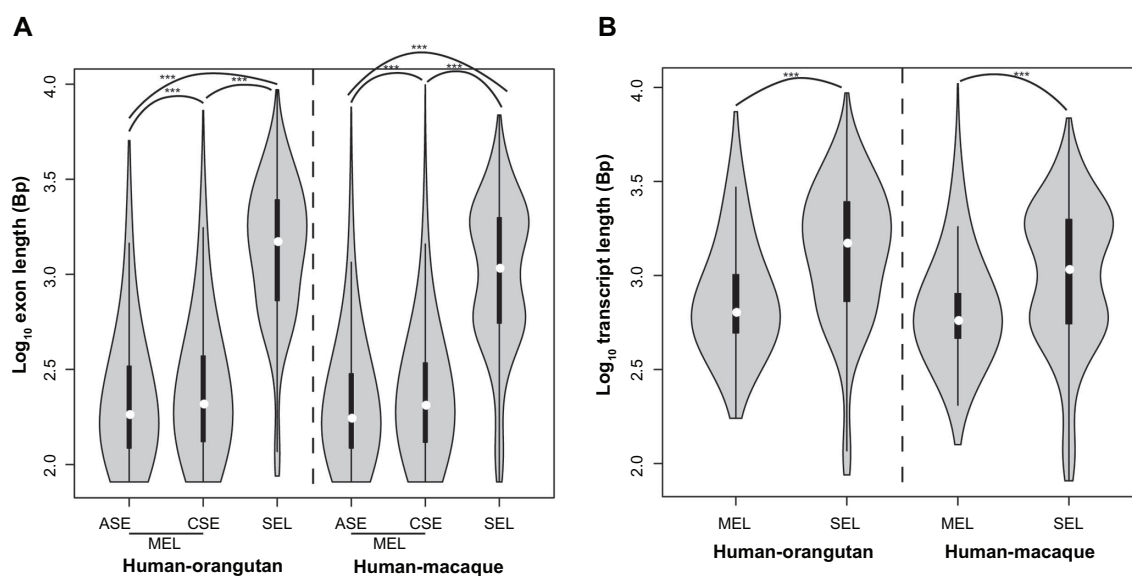

Figure 2. The Violin Plot of (**A**) exon length; and (**B**) transcript length of human lincRNAs.
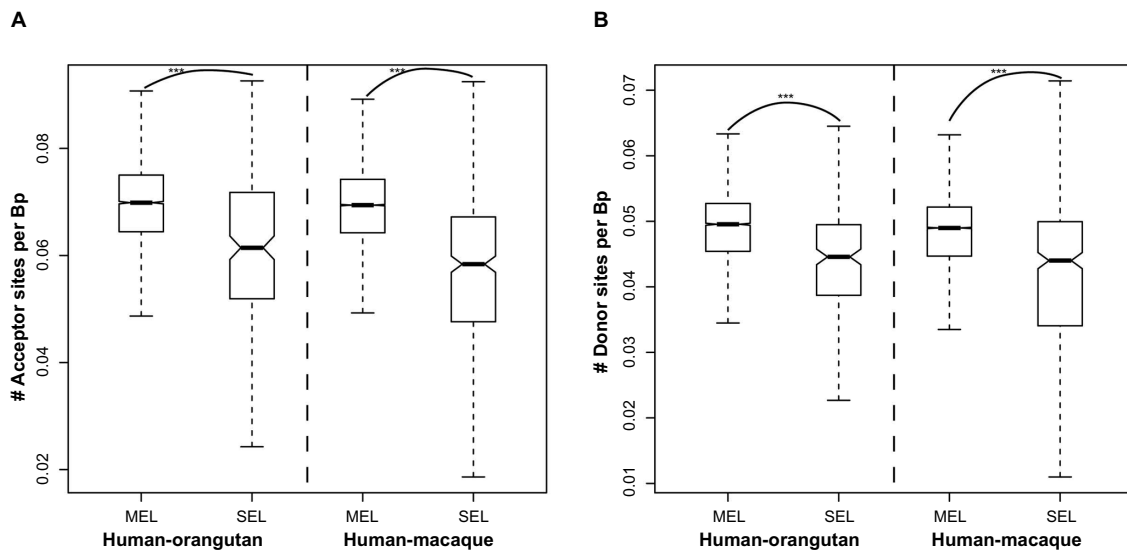**Note:** Statistical significance: ***$P < 0.001$ by Wilcoxon Rank Sum test.

Figure 3. The length normalized numbers of (**A**) putative acceptor sites; and (**B**) putative donor sites as predicted by the SplicePort program.
Note: Statistical significance: ***P < 0.001 by Wilcoxon Rank Sum test.

(Fig. 4B). This observation further supports our hypothesis that SELs were functionally different from MELs.

SELs were more widely expressed than MELs. One indicator of lincRNA functionality is the breadth of gene expression across different tissues. We thus retrieved RNA-seq data from 16 human tissues, and examined the expression breadths of SELs and MELs separately. Interestingly, as shown in Figure 5, SELs were more widely expressed than MELs in both of the datasets. Of note, hominoid-specific SELs were significantly more widely expressed in human tissues than primate-conserved SELs. Similar comments also apply to MELs. Note that the "hominoid-specific" and "primate-conserved" SELs/MELs were defined according to sequence alignability and the presence of homologous exon boundaries in the examined hominoid or primate genomes. Therefore, the "lineage specificity" of lincRNAs discussed here was supported by "genomic conservation" rather than RNA-seq data from multiple species. This is considered as a relaxed criterion because conservation of genomic sequences does not guarantee conservation of expression patterns. Interestingly, nevertheless, it has been recently reported that the tissue specificity of lincRNA expression could be conserved in mammals even without conserved exon boundaries.[34,35] In brief, the results presented in Figure 5 should be interpreted as "the level of genomic sequence conservation of a lincRNA is associated with its expression breadth in human tissues."

One potential concern of this analysis is that the expression of SELs might have been more likely to be detected by RNA-seq than that of MELs because the former were significantly longer.[36] Since lincRNAs are usually lowly expressed, short transcripts could be undetectable given insufficient sequencing depths. To address this issue, we retrieved SELs and MELs of lengths between 500 bp and 1,000 bp, and examined their expression breadths again. The

results remained similar (Supplementary Fig. 4). Therefore, the observation that SELs were more widely expressed than MELs might not have resulted from the difference in length between the two gene groups.

Discussion

In this study, we compared several biological features of hominoid-specific and primate-conserved SELs and MELs. We found these two groups of lincRNAs to differ from each other in primary sequence conservation, exon length, propensity to splicing, distance to the nearest coding gene, and expression breadth. These two groups of lincRNA genes thus appear to have experienced different evolutionary histories and represent distinct biological subtypes.

The primary sequences of long noncoding RNAs were reported to be weakly selected.[20] For hominoid-specific lincRNAs, we found no evidence for negative selection on the primary sequences of lincRNA exons (Fig. 1A). Unexpectedly, the CSEs and ASEs of MELs both had slightly larger H-O genetic distances than pure introns and intergenic regions. There are two possible explanations for this observation. First, the mutation rates in CSEs and ASEs could be higher than those in pure introns and intergenic regions. Since transposable elements were suggested to be an important contributor of lincRNA exons,[37] and repetitive elements were known to evolve rapidly,[38] the elevated H-O genetic distances in CSEs and ASEs could have resulted from a larger proportion of repetitive elements in these exonic regions. We thus calculated the proportion of exonic/non-exonic sequences that overlapped with repetitive elements. As shown in Supplementary Figure 5, CSEs and ASEs actually tended not to overlap with repetitive elements. Therefore, the content of repetitive elements cannot explain the higher H-O genetic distances in ASEs and CSEs. Other genomic features such
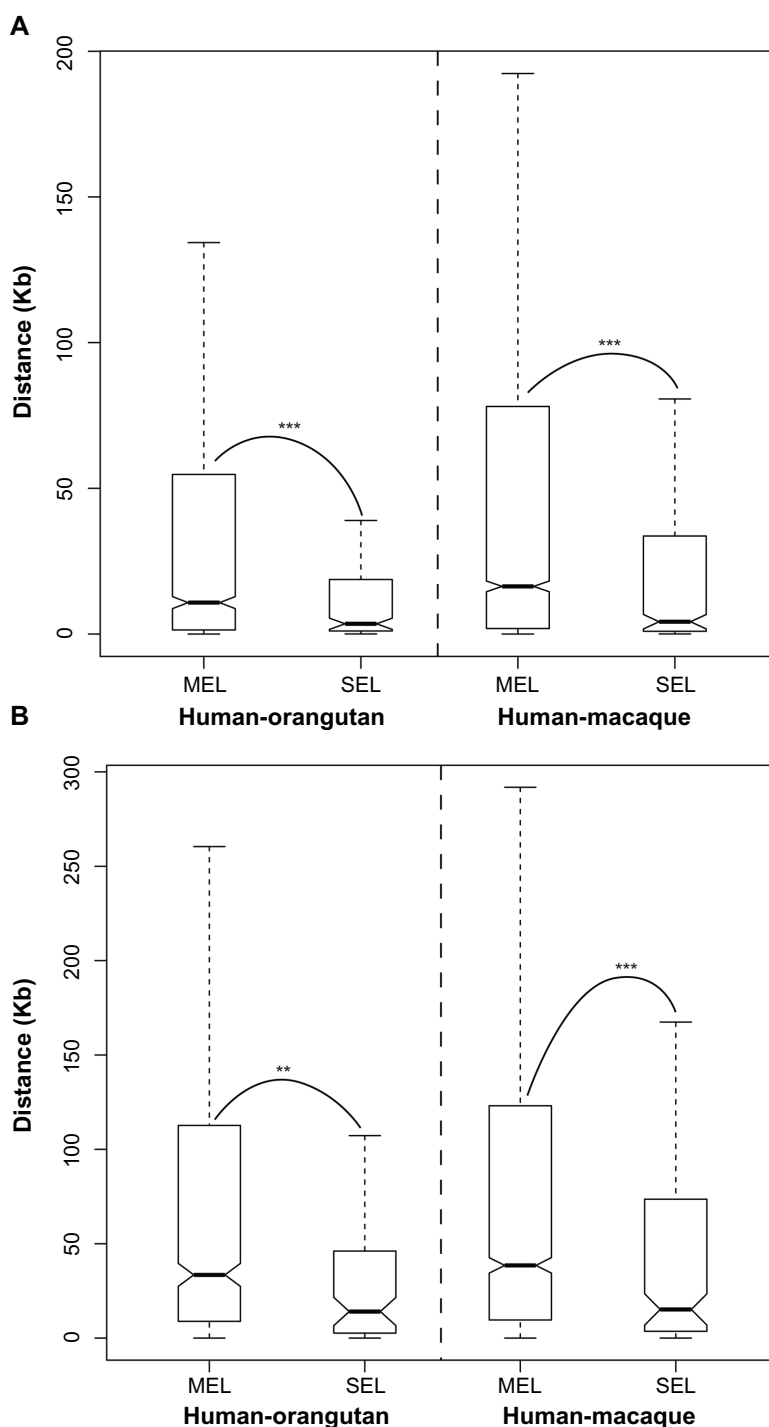
Figure 4. (**A**) The physical distances between lincRNAs and the closest coding genes. (**B**) the physical distances between lincRNAs and the closest enhancers.

Note: Statistical significance: ***$P < 0.001$ by Wilcoxon Rank Sum test.

as recombination rate and mutation hotspot may not account for the elevated genetic distances in CSEs and ASEs, for the non-exonic regions (pure introns and intergenic regions) were retrieved from the genomic regions very close to these exons (see Materials and Methods). Second, CSEs and ASEs of MELs might have been positively selected in either the human or orangutan lineage. To explore this possibility, we extracted 1000 Genome-based single-nucleotide polymorphism (SNP)

information for three representative populations (YRI, CEU, and JPT) from dbPSHP[39] as of June, 2014. If CSEs and ASEs in the H-O dataset have been subject to positive selection, the derived allele frequency should be higher in these two exonic regions than in the other regions. However, less than 0.2% of the studied exonic/non-exonic regions were found to contain SNPs. Therefore, we found no clear evidence for positive selection on the ASEs and CSEs of MELs in the H-O
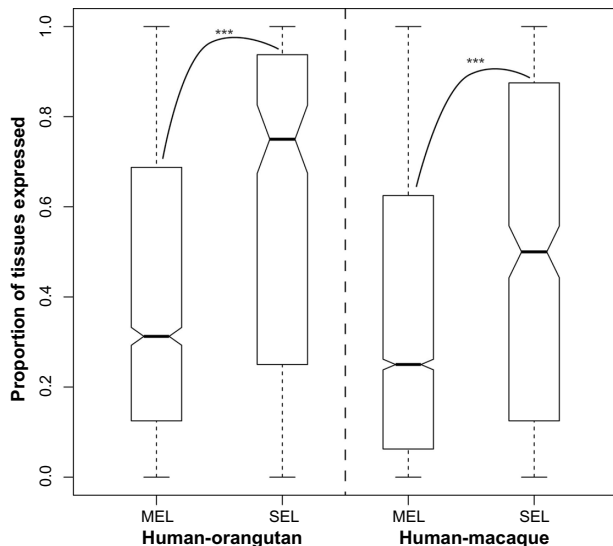
Figure 5. The proportions of human tissues (out of 16) where lincRNAs are expressed.
Note: Statistical significance: ***$P < 0.001$ by Wilcoxon Rank Sum test.

dataset. The reason for the larger H-O genetic distances in CSEs and ASEs remains unclear.

In comparison, for primate-conserved lincRNAs, the genetic distances in SELs and ASEs (but not CSEs) were slightly yet significantly smaller than those in non-exonic regions (Fig. 1B). This observation implied that the primary sequences of SELs and ASEs were weakly constrained, whereas CSEs were not. Of note, the rapid turnover of non-coding RNAs[17] suggested that the majority of the nucleotides in the exons were selectively neutral. The statistically significant deviations from neutrality of the sequences of SELs and ASEs observed in this study are thus particularly meaningful. The ~4% deviation (the %difference between the 0.069 H-Ma genetic distance in pure introns and 0.066 distance in SELs/ASEs) indicated that approximately 4% of the nucleotides in SELs and ASEs were selectively constrained. Of note, this selective constraint was not observed in the H-O dataset (Fig. 1A), which implied that hominoid-specific lincRNAs had strayed into a different evolutionary path from their more ancestral counterparts (primate-conserved lincRNAs). In view of the differences between Figures 1A and B, the hominoid–primate divergence in sequence evolution seemed to have occurred in both SELs and MELs.

Interestingly, in coding sequences, ASEs were reported to evolve more slowly at the RNA level but more rapidly at the protein level than CSEs.[40,41] We speculate that RNA splicing plays an important role in both coding and primate-conserved lincRNA genes, therefore constraining sequence evolution at splicing signals in both gene types. And this constraint in turn has led to slightly reduced evolutionary rates in ASEs. However, this may not be true for hominoid-specific lincRNA genes. It is worth noting that the exon boundaries of lincRNA genes were reported to turnover rapidly, which suggested

that exact splicing sites might be unimportant for the functions of lincRNAs.[18] However, we discovered that although the exon boundaries of human lincRNAs might not be found at the exact orthologous positions in the compared genomes, "backup" boundaries were nearby (within 10% of the exon length) in >99% of the cases. Furthermore, splicing patterns also change rapidly in coding genes,[11,42] but this regulatory mechanism remains biologically important,[1] and is involved in critical processes such as functional pleiotropy[43] and adaptation.[10] In lieu of the selective constraint on ASEs (particularly in primate-conserved lincRNAs) and the biological differences between SELs and MELs, it is likely that splicing *per se* is biologically meaningful for the functions or regulations of lincRNAs, but "exact splicing" may be unnecessary.

It was previously suggested that the primary sequences of lincRNAs evolved rapidly because the functionally important secondary structures of lincRNAs could be preserved even if the primary sequences had changed.[17] We thus conducted a simulation study by *in silico* mutating each lincRNA nucleotide transitionally, and calculated the average changes in free energy (ΔE) separately for each type of exon (see Materials and Methods). However, the differences in average ΔE were statistically insignificant among CSEs, ASEs, and SELs (Supplementary Fig. 6). Therefore, changes in secondary structure resulting from single-nucleotide transitional substitutions cannot explain the differences in sequence conservation among the three exon types. However, we cannot completely rule out the possibility that secondary structure is one of the reasons for the inter-exon type differences in H-O/H-Ma genetic distance. This is because we generated only one transitional substitution per experiment, whereas multiple substitutions (including transversional substitutions) are biologically possible. Furthermore, drastic changes in structure may occur when two or more mutations occur in different exons, yet here we considered only mutations in a single exon.

Conclusions

In this study, we provided evidence that SELs and MELs represented two biologically distinct gene groups. They differed in primary sequence evolution, exon/transcript length, proximity to nearest coding gene, and expression breadth. The differences in regulatory mechanism between the two gene types thus are worth further explorations. Notably, ASEs were found to be slightly more conserved than CSEs in primate-conserved lincRNAs. Furthermore, splicing appeared to be disfavored by selection in SELs as compared with in MELs. These results suggest that splicing might be relevant to the functionality of lincRNAs. The exact roles of splicing in lincRNA functions await future investigations.

Abbreviations

ASE: alternatively spliced exon
AS: alternative splicing

Bp: base pair
CSE: constitutively spliced exon
H-O: human-orangutan
H-Ma: human-macaque
MEC: multi-exonic coding gene
MEL: multi-exonic lincRNA
lincRNA: long intergenic noncoding RNA
RNA-seq: RNA sequencing
SEC: single-exonic coding gene
SEL: single-exonic lincRNA

Author Contributions

FCC conceived of and designed the study. CLP and HYL retrieved and analyzed the data. FCC interpreted the results and drafted the manuscript. All authors have read and approved the manuscript.

Supplementary Data

Supplementary Figure 1. The genetic distances in lincRNA exons between mouse and rat. The numbers in the parentheses indicate the sample sizes. Statistical significance: *$P < 0.05$; ***$P < 0.001$ by Wilcoxon Rank Sum test.

Supplementary Figure 2. The numbers of (**A**) putative acceptor sites; (**B**) putative donor sites as predicted by the Splice Port program. Statistical significance: ***$P < 0.001$ by Wilcoxon Rank Sum test.

Supplementary Figure 3. Spearman's correlations between human-orangutan/human-macaque genetic distances of lincRNAs and the physical distances between lincRNAs and the closest coding genes.

Supplementary Figure 4. The proportions of tissues (out of 16) where lincRNAs of lengths 500–1000 bp are expressed. Statistical significance: ***$P < 0.001$ by Wilcoxon Rank Sum test.

Supplementary Figure 5. The proportions of repetitive elements in ASEs, CSEs, SELs, pure introns, and intergenic regions. All pairwise differences are statistically significant ($P < 0.001$ by Wilcoxon Rank Sum test) except for the differences between ASEs and CSEs.

Supplementary Figure 6. The effect of single-base transitional mutations on the free energy of lincRNA secondary structure as predicted by RNAfold. ΔE: the difference in free energy between the mutated and original sequence. All of the pairwise differences are statistically insignificant.

REFERENCES

1. Kelemen O, Convertini P, Zhang Z, et al. Function of alternative splicing. *Gene*. 2013;514:1–30.
2. Singh RK, Cooper TA. Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med*. 2012;18:472–82.
3. Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int J Evol Biol*. 2012;2012:596274.
4. Xing Y, Lee C. Alternative splicing and RNA selection pressure – evolutionary consequences for eukaryotic genomes. *Nat Rev Genet*. 2006;7:499–509.
5. Wang GS, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet*. 2007;8:749–61.
6. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. 2010;11:345–55.
7. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. *Nat Rev Genet*. 2011;12:715–29.
8. Parenteau J, Durand M, Véronneau S, et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol Biol Cell*. 2008;19:1932–41.
9. Candales MA, Duong A, Hood KS, et al. Database for bacterial group II introns. *Nucleic Acids Res*. 2012;40:D187–90.
10. Parker DJ, Gardiner A, Neville MC, Ritchie MG, Goodwin SF. The evolution of novelty in conserved genes; evidence of positive selection in the drosophila fruitless gene is localised to alternatively spliced exons. *Heredity (Edinb)*. 2014;112:300–6.
11. Harr B, Turner LM. Genome-wide analysis of alternative splicing evolution among Mus subspecies. *Mol Ecol*. 2010;19(suppl 1):228–39.
12. Irimia M, Maeso I, Gunning PW, Garcia-Fernandez J, Roy SW. Internal and external paralogy in the evolution of tropomyosin genes in metazoans. *Mol Biol Evol*. 2010;27:1504–17.
13. Feng D, Xie J. Aberrant splicing in neurological diseases. *Wiley Interdiscip Rev RNA*. 2013;4:631–49.
14. Gamazon ER, Stranger BE. Genomics of alternative splicing: evolution, development and pathophysiology. *Hum Genet*. 2014;133:679–87.
15. Munaut C, Colige AC, Lambert CA. Alternative splicing: a promising target for pharmaceutical inhibition of pathological angiogenesis? *Curr Pharm Des*. 2010;16:3864–76.
16. Gentles AJ, Karlin S. Why are human G-protein-coupled receptors predominantly intronless? *Trends Genet*. 1999;15:47–9.
17. Johnsson P, Lipovich L, Grander D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta*. 2014;1840:1063–71.
18. Stranzl T, Larsen MV, Lund O, Nielsen M, Brunak S. The cancer exome generated by alternative mRNA splicing dilutes predicted HLA class I epitope density. *PLoS One*. 2012;7:e38670.
19. Agranat-Tamir L, Shomron N, Sperling J, Sperling R. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. *Nucleic Acids Res*. 2014;42:4640–51.
20. Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22:1775–89.
21. Orom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell*. 2013;154:1190–3.
22. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol*. 2013;11:59.
23. Novikova IV, Hennelly SP, Tung CS, Sanbonmatsu KY. Rise of the RNA machines: exploring the structure of long non-coding RNAs. *J Mol Biol*. 2013;425:3731–46.
24. Park SG, Choi SS. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol*. 2010;10:241.
25. Liao BY, Scott NM, Zhang J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol*. 2006;23:2072–80.
26. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
27. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;28:511–5.
28. Zhang Q, Li H, Jin H, Tan H, Zhang J, Sheng S. The global landscape of intron retentions in lung adenocarcinoma. *BMC Med Genomics*. 2014;7:15.
29. Dogan RI, Getoor L, Wilbur WJ, Mount SM. SplicePort – an interactive splice-site analysis tool. *Nucleic Acids Res*. 2007;35:W285–91.
30. Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res*. 2011;21:1916–28.
31. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk – database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*. 2011;44:839–47.
32. Mancini-Dinardo D, Steele SJ, Levorse JM, Ingram RS, Tilghman SM. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev*. 2006;20:1268–82.
33. Fang Z, Rajewsky N. The impact of miRNA target sites in coding sequences and in 3'UTRs. *PLoS One*. 2011;6:e18067.
34. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res*. 2014;24:616–28.
35. Necsulea A, Kronenberg Z, Lynch VJ, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*. 2014;505:635–40.

36. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*. 2009;4:14.

37. Kapusta A, Kronenberg Z, Lynch VJ, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long non-coding RNAs. *PLoS Genet*. 2013;9:e1003470.

38. Graur D, Li W-H. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates; 2000.

39. Li MJ, Wang LY, Xia Z, Wong MP, Sham PC, Wang J. dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res*. 2014;42:D910–6.

40. Chen FC, Wang SS, Chen CJ, Li WH, Chuang TJ. Alternatively and constitutively spliced exons are subject to different evolutionary forces. *Mol Biol Evol*. 2006;23:675–82.

41. Chen FC, Chaw SM, Tzeng YH, Wang SS, Chuang TJ. Opposite evolutionary effects between different alternative splicing patterns. *Mol Biol Evol*. 2007;24:1443–6.

42. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*. 2012;338:1593–9.

43. Reyes A, Anders S, Weatheritt RJ, Gibson TJ, Steinmetz LM, Huber W. Drift and conservation of differential exon usage across tissues in primate species. *Proc Natl Acad Sci USA*. 2013;110:15377–82.