Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci

WILFRIED HAERTY and CHRIS P. PONTING

MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3PT, United Kingdom

ABSTRACT

If sequencing was possible only for genomes, and not for RNAs or proteins, then functional protein-coding exons would be recognizable by their unusual patterns of nucleotide composition, specifically a high GC content across the body of exons, and an unusual nucleotide content near their edges. RNAs and proteins can, of course, be sequenced but the extent of functionality of intergenic long noncoding RNAs (lncRNAs) remains under question owing to their low nucleotide conservation. Inspired by the nucleotide composition patterns of protein-coding exons, we sought evidence for functionality across lncRNA loci from diverse species. We found that such patterns across multiexonic lncRNA loci mirror those of protein-coding genes, although to a lesser degree: Specifically, compared with introns, lncRNA exons are GC rich. Additionally we report evidence for the action of purifying selection to preserve exonic splicing enhancers within human multiexonic lncRNAs and nucleotide composition in fruit fly lncRNAs. Our findings provide evidence for selection for more efficient rates of transcription and splicing within lncRNA loci. Despite only a minor proportion of their RNA bases being constrained, multiexonic intergenic lncRNAs appear to require accurate splicing of their exons to transact their function.

Keywords: intergenic long noncoding RNAs; splicing; ESEs; GC content; selection

INTRODUCTION

Nucleotide composition has long been known to vary greatly among long genomic regions (Eyre-Walker and Hurst 2001; Duret and Galtier 2009). It also varies at shorter scales between coding regions and their flanking sequences (Louie et al. 2003), and shorter still within protein-coding genes between their exons and introns (Louie et al. 2003; Schwartz et al. 2009). The trend of higher GC content in exons over introns is a hallmark of coding sequences (Amit et al. 2012) and has been interpreted as implying more efficient transcription, splicing, or translation (Kudla et al. 2006; Amit et al. 2012). Nucleotide compositional variation across exons has been associated with short motifs proximal to exon–intron boundaries that either enhance or inhibit splicing (Mount et al. 1992; Fairbrother et al. 2002; Wang et al. 2004).

While these features are well known for protein-coding sequences, as are their molecular functions, much remains to be learned for the thousands of intergenic long (≥200 nt) noncoding RNAs (lncRNAs) that have been predicted to be transcribed from animal genomes (Ulitsky and Bartel 2013). Few such loci have been experimentally characterized, but those that have possess roles in dosage compensation in human and fruit fly (Kay et al. 1993; Kelley et al. 1999), splicing regulation, phosphorylation, chromatin remodeling, and pluripotency maintenance (for reviews, see Ulitsky and Bartel 2013; Ponting et al. 2009). Function, for some lncRNA loci, is conveyed by the act of transcription across the locus (transcriptional interference and/or chromatin remodeling) with the resulting transcript being functionally inert (Latos et al. 2012; Yoo et al. 2012), whereas for others it is mediated by the RNA transcript itself (for example: *Xist*, *Paupar*, Brockdorff et al. 1992; Vance et al. 2014). It remains unclear, however, how prevalent are these two classes of lncRNA mechanism. LncRNA loci are found in diverse genomic contexts (enhancer, promoter-associated, intergenic, intronic, antisense; for review, see Qureshi and Mehler 2012) and their transcripts, either polyadenylated or nonpolyadenylated, can be located in diverse cellular compartments (Derrien et al. 2012; van Heesch et al. 2014). They also vary widely in size, ranging from as few as 200 nt to >8 kb for loci such as *Malat1*. In addition lncRNA loci can either be composed of a single exon (for example, *Malat1*, *Paupar*) or of multiple exons (*Hotair* and *Xist*).

An improved understanding of the potential biological functions of intergenic lncRNA loci could derive from interrogating their nucleotide sequence and composition. Previous studies have attempted to identify functional domains

Corresponding author: wilfried.haerty@dpag.ox.ac.uk

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.047324.114. Freely available online through the *RNA* Open Access option.

© 2015 Haerty and Ponting This article, published in *RNA*, is available under a Creative Commons License (Attribution 4.0 International), as described at http://creativecommons.org/licenses/by/4.0/.

within lncRNA by predicting RNA secondary structures or the potential interaction forces between lncRNA loci and protein-coding sequences (Bellucci et al. 2011). RNA secondary structure predictions have tentatively ascribed functional regions to lncRNAs (Smith et al. 2013) but these have tended to suffer from high false-positive rates (Babak et al. 2007). In general, while lncRNA exons are better conserved than their introns, most strikingly so for *Drosophila* sequences, their conservation is exceedingly modest in human, whose contrast to fruit flies is a presumed result of humans' much lower effective population size (Haerty and Ponting 2013).

We sought to further understand the degree by which intergenic lncRNAs contribute to biological function in diverse animal species. We looked beyond the known modest sequence conservation in lncRNA exons by examining their nucleotide composition, and their hallmarks of active transcription, splicing, and evolution, specifically in comparison with protein-coding genes. Unexpectedly, we find that selection acts on exonic splicing enhancers in human and on nucleotide substitutions in lncRNAs in fruit fly.

Our findings of signatures of efficient splicing and selection, similar to those evident within protein-coding genes, indicate that despite their exons' low degrees of sequence conservation, many multiexonic lncRNA loci are likely to possess spliced RNA-dependent functions.

RESULTS

We focused our studies on 66,500 intergenic lncRNAs from fruit fly, zebrafish, coelacanth, mouse, and human, respectively (Table 1). This wide phyletic range allowed us to obtain a broad perspective of lncRNA evolution and function across metazoan evolution. In previous studies on nematode (Nam and Bartel 2012), zebrafish, and mouse (Ulitsky et al. 2011), intergenic lncRNA loci were found to exhibit a GC content that is higher than their flanking intergenic region yet that is significantly lower than that for protein-coding sequences. Exonic and intronic GC contents and splicing motifs were, however, not compared in these previous studies.

Elevated GC content in lncRNA exons over introns for diverse metazoans

Exons of intergenic multiexonic lncRNAs for all five species that we considered contain significantly higher GC content than either their intronic or flanking intergenic sequences; this pattern of GC content variation across the gene model mirrors that seen for their neighboring protein-coding gene models, albeit to a reduced extent (Fig. 1A–E). Single exon intergenic lncRNAs exhibit lower GC content that is only marginally greater than that of their untranscribed flanking sequences, and thus significantly lower than for multiexonic intergenic lncRNAs ($P < 0.001$) (Fig. 1F). The nucleotide composition of exons from multiexonic lncRNA loci is between that for $5'$ or $3'$ UTRs ($P < 0.05$ in both comparisons), while monoexonic loci have the lowest GC content of all exonic categories we considered ($P < 0.05$) (Fig. 1F) with the exception of $3'$-UTR exons ($P > 0.05$). In their elevated GC content, therefore, multiexonic lncRNAs are more similar to protein-coding loci than they are to monoexonic lncRNAs.

The elevated GC content seen for both multiexonic lncRNA and protein-coding gene exons might indicate that predicted lncRNA exons are, instead, protein-coding. Indeed, a number of proposed lncRNA loci, at least in *Drosophila* (Ladoukakis et al. 2011), zebrafish, and human (Bazzini et al. 2014), may encode small open-reading frames (smORFs <100 amino acids). Because of the high GC content of protein-coding exons, we compared the nucleotide content of these smORFs to the remaining intergenic lnRNAs. Noncoding multiexonic lncRNAs exhibit GC levels that are higher than for introns yet that are significantly lower than for these multiexonic smORFs (Mann–Whitney test, $P = 0.032$, Supplemental Fig. 1) arguing that the two locus classes are distinct. In all subsequent analyses all smORF transcripts were excluded from consideration.

Another potential factor influencing the nucleotide composition of lncRNAs could be the embedding of DNA functional elements within them. Indeed, nucleotide composition and GC content more specifically have been directly associated with human transcription factor binding sites (TFBS) (White et al. 2013). We thus considered associations between lncRNA exonic GC content and the presence of TFBS or enhancer elements defined by the FANTOM5 consortium (Andersson et al. 2014) within these exons. Within both monoexonic and multiexonic lncRNAs we observed a significant association between nucleotide composition and the occurrence of TFBS or enhancers ($P < 0.001$). Even within the same locus, exons with a transcription factor binding site have a significantly higher GC content than exons without such sites (Mann–Whitney test $P < 2.2 \times 10^{-16}$, Supplemental Fig. 2). However, multiexonic lncRNAs remain GC rich compared with monoexonic loci even when they

TABLE 1. Numbers of multi- and monoexonic intergenic lncRNAs from five metazoan species

Species	Multiexonic	Monoexonic	References
Drosophila melanogaster	789	330	Young et al. (2012)
Danio rerio	809	4	Ulitsky et al. (2011); Pauli et al. (2012)
Latimeria chalumnae	1046	168	Amemiya et al. (2013)
Mus musculus	2350	15	Belgard et al. (2011); Flicek et al. (2013)
Homo sapiens	14,484	46,517	Cabili et al. (2011); Derrien et al. (2012); Hangauer et al. (2013)

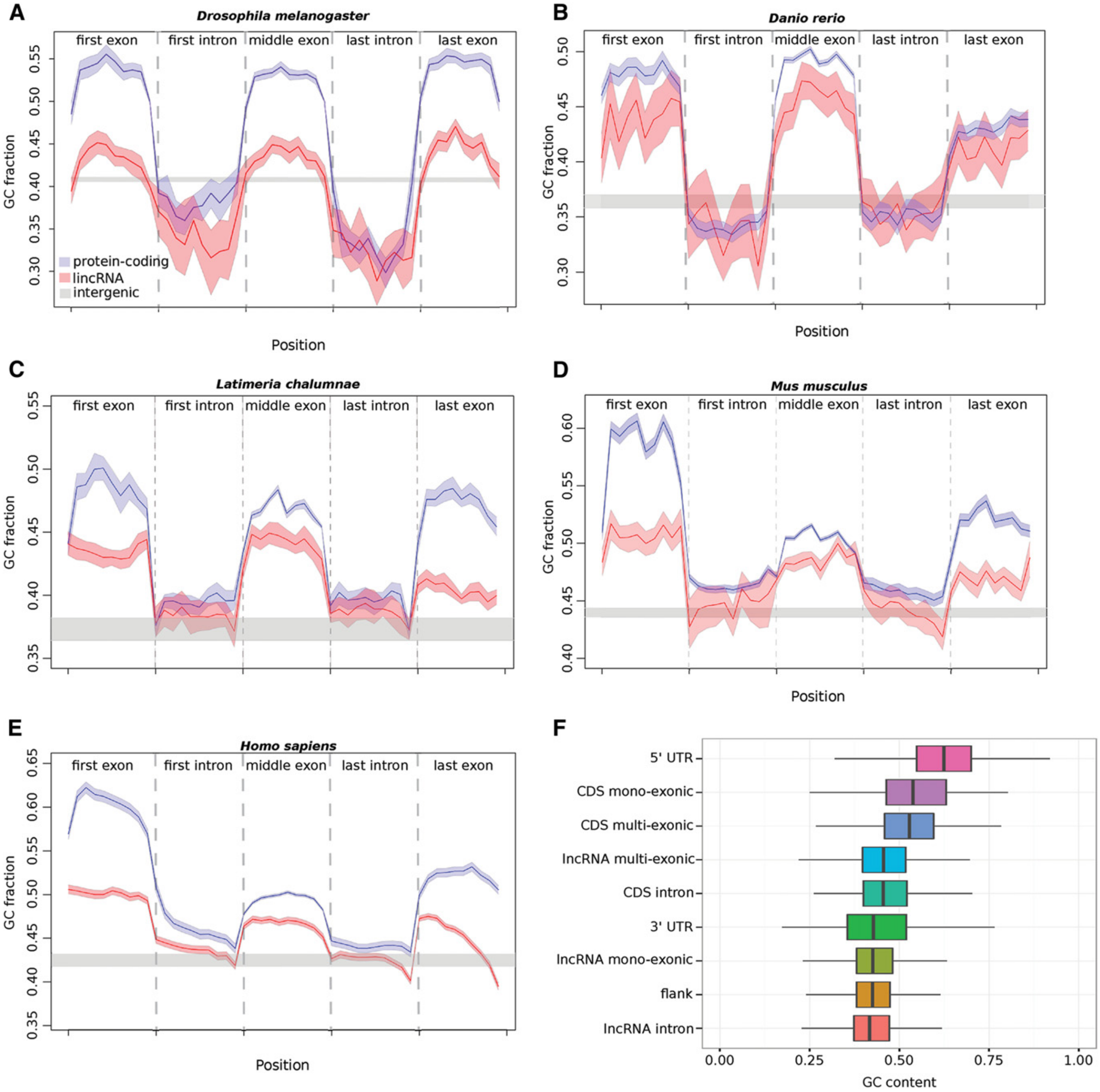

FIGURE 1. GC content variation across multiexonic protein-coding (blue) and intergenic lncRNA loci (red) in (*A*) *Drosophila melanogaster* (Young et al. 2012), (*B*) *Danio rerio* (Ulitsky et al. 2011; Pauli et al. 2012), (*C*) *Latimeria chalumnae* (Amemiya et al. 2013), (*D*) *Mus musculus* (Belgard et al. 2011), and (*E*) *Homo sapiens* (Cabili et al. 2011; Derrien et al. 2012). Nonoverlapping windows each sampling 10% of the sequences were used. The gray band indicates GC content of flanking intergenic sequences. (*F*) Comparison of exonic GC content between multiexonic and monoexonic intergenic lncRNA loci as well as 5′- and 3′-UTR exons from protein-coding genes in human.

do not contain experimental TFBS or enhancers (Mann–Whitney test $P < 2.2 \times 10^{-16}$, Supplemental Fig. 3).

GC content across multiexonic gene models

Protein-coding exons exhibit particularly elevated exonic GC content and strong splice sites when they are flanked by long introns (Amit et al. 2012). We found, similarly, that the elevated GC content of lncRNA exons over their flanking introns is significantly greater for long (third quartile) introns than for shorter (first quartile) introns (Mann–Whitney tests, $P = 1.4 \times 10^{-12}$ and $P = 3.4 \times 10^{-6}$ for 5′ and 3′ introns, respectively, Supplemental Fig. 4). This exon–intron differential GC content is significantly correlated with the length of the flanking introns ($\rho = 0.085$, $P = 9.9 \times 10^{-14}$ and $\rho = 0.057$, $P = 6.6 \times 10^{-7}$ for 5′ and 3′ introns, respectively). If multiexonic

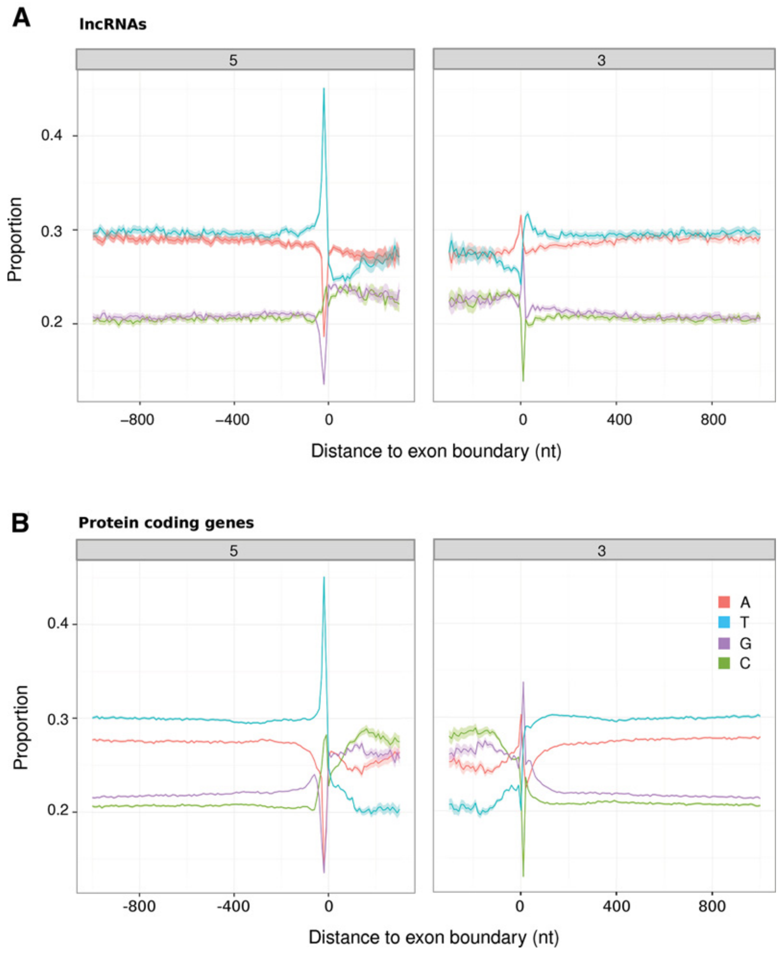

FIGURE 2. Comparison of nucleotide composition at the exon–intron boundaries of internal exons within human intergenic lncRNAs (*A*) and protein-coding genes (*B*). The areas designate the fifth and 95th percent confidence intervals.

lncRNA loci are commonly functionally spliced then it might be expected that splice sites would be significantly stronger (as inferred from their information content) (Yeo and Burge 2004) when flanked by long introns. This, indeed, was found to be the case (Mann–Whitney test, $P = 2.8 \times 10^{-6}$ and $P = 2.1 \times 10^{-5}$ for the 5′ and 3′ splice sites, respectively, comparing the 25% shortest or longest introns). Additionally, we observed a significant positive correlation between intron size and splice site strength in both coding ($\rho = 0.081$, $P < 2.2 \times 10^{-16}$, and $\rho = 0.076$, $P < 2.2 \times 10^{-16}$ for 5′ and 3′ splice sites, respectively) and lncRNA ($\rho = 0.087$, $P = 2.2 \times 10^{-14}$, and $\rho = 0.081$, $P = 1.2 \times 10^{-14}$ for 5′ and 3′ splice sites, respectively) loci.

We next examined sequence adjacent to lncRNA exon–intron boundaries and focused on 6136 human intergenic lncRNAs with at least three exons and constitutively spliced introns. Although not as dramatic as for the protein-coding exons, the compositional bias of their internal exons is due to increases of both G and C nucleotides equally. Less expected was the strong T enrichment within 40 nt upstream of the acceptor splice site signal which indicates the presence of the polypyrimidine tract as well as the depletion in T (but not A) toward both 5′ and 3′ exonic boundaries (Fig. 2). To consider whether this depletion reflected the preferential location of purine-rich exonic splicing enhancers (ESEs) toward exon–intron boundaries we predicted hexamers that are enriched near (<50 nt) to these boundaries. These lncRNA exon hexamers were found to be highly concordant with the set of ESE motifs previously identified by Fairbrother et al. (2002) for protein-coding exons (3.80- and 3.87-fold enrichment, respectively, for the 5′ and 3′ exonic boundaries if $\Delta E \geq 2.5$). The ESE scores (calculated as per Fairbrother et al. 2002) were also highly correlated between protein-coding and lncRNA sequences (Spearman's rank correlation coefficient, $\rho = 0.75$ and $\rho = 0.76$ for the exonic 5′ and 3′ ends, respectively) (Fig. 3A).

Evolutionary conservation of ESESs in multiexonic lncRNAs

Conservation of human lncRNA transcription across mammalian evolution is the exception rather than the rule, which contrasts with the high evolutionary conservation of protein-coding mRNA transcription (Kutter et al. 2012; Necsulea et al. 2014). It was unexpected, therefore, that exonic predicted ESE levels were equivalent to those of highly conserved protein-coding genes for macaque, dog or mouse genomic sequence orthologous to predicted ESEs in human multiexonic lncRNAs (Fig. 4). We conclude, at least for 171 human multiexonic intergenic lncRNAs whose ESEs could be aligned to genomic sequence of all these other mammals, that the density of predicted ESEs was preserved across 100 million years of mammalian evolution, as they are for protein-coding genes.

Using multiple species' genome alignments, we investigated the pattern of nucleotide conservation associated with exon splicing. We found that nucleotide conservation in lncRNA exons across primates is greatest near to the exon boundaries in comparison to sequences of the same length from the middle of the same exons (5′: $P = 6.6 \times 10^{-16}$ and 3′: $P = 7.1 \times 10^{-10}$ after Bonferroni correction; Materials and Methods). Additionally, we observed sequence adjacent to the splice acceptor site at the 5′ ends of lncRNA exons significantly better conserved than sequence at 3′ ends (Kruskal–Wallis test, $P = 1.0 \times 10^{-5}$). This finding, which

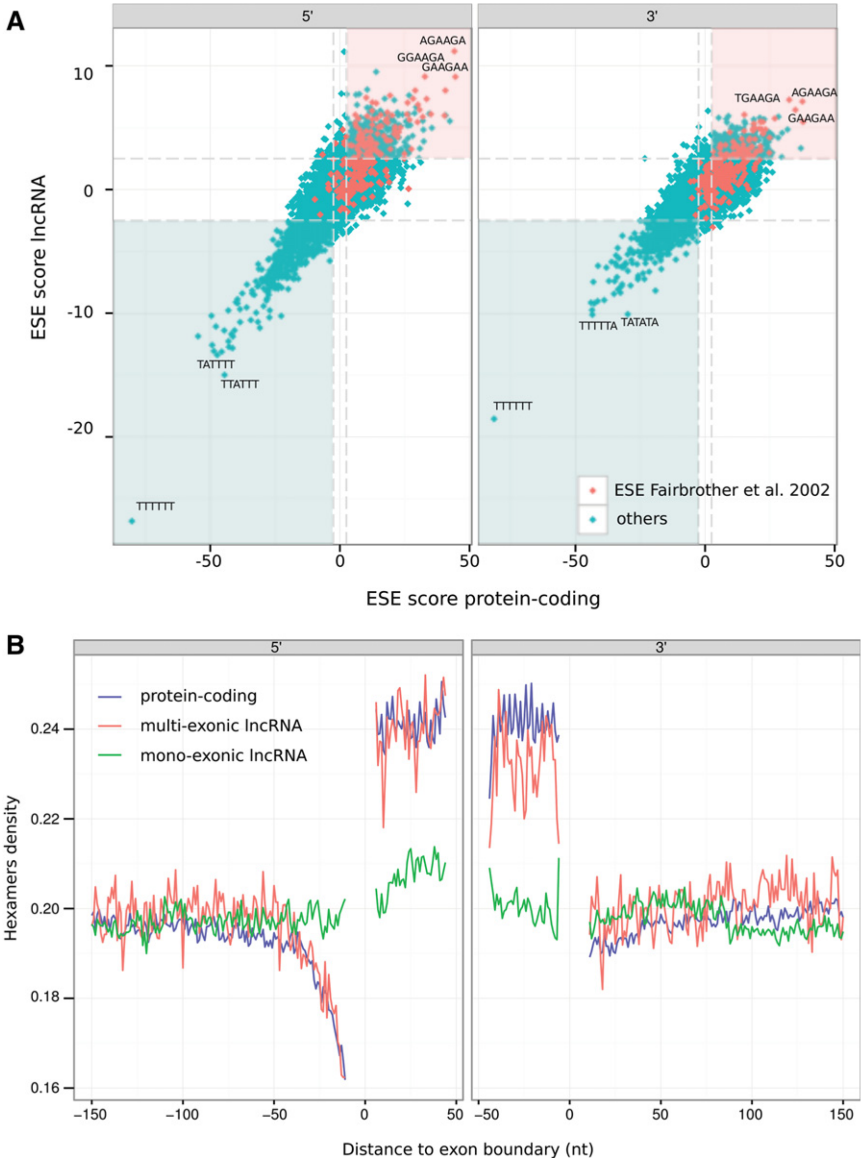

FIGURE 3. (*A*) Comparison of hexamer scores (as per Fairbrother et al. 2002) between human protein-coding and lncRNA exons. Red dots represent the ESEs previously identified by Fairbrother et al. (2002). Red and blue areas contain motifs that are enriched within exons or within introns, respectively, of both protein-coding and lncRNAs (ESEscore > |2.5|, Fairbrother et al. 2002). (*B*) Proportion of hexamers that are ESEs (Fairbrother et al. 2002; Zhang and Chasin 2004; Goren et al. 2006) at the 5′ and 3′ exon–intron boundaries within human protein-coding (blue), multiexonic lncRNAs (red), and monoexonic lncRNA (green) loci.

was previously seen for protein-coding sequences (Chamary and Hurst 2005; Chamary et al. 2006; Parmley et al. 2007; Warnecke and Hurst 2007), reflects the sequence requirements for the spliced RNA-dependent functions of multiexonic lncRNAs.

To investigate whether purifying selection has acted on human multiexonic lncRNA ESE sequences we compared derived allele frequency distributions (Vitti et al. 2013) between single nucleotide variants (SNV) occurring within ESE or non-ESE motifs within 50 nt of exon boundaries. No differences were found, perhaps owing to the method being underpowered to detect selection from such small sequence samples. Nonetheless, the density of human polymorphic sites within these multiexonic lncRNA predicted ESEs was significantly lower (empirical $P < 0.001$) (Fig. 5A; Supplemental Fig. 5) relative to composition matched random samples (see Materials and Methods), and substitutions at these sites, when comparing human sequence with either chimpanzee or macaque, are significantly depleted (empirical $P < 0.001$ and $P < 0.001$, respectively).

Next we extended the analysis to consider whether nucleotide substitutions conserve ESE (or non-ESE) status. We observed significant differences in the nucleotide substitution patterns characterized by a depletion of ESE creating and ESE disrupting mutations within lncRNA exon boundaries relative to the center of exons (χ^2, $P = 0.014$) (Fig. 5B). Importantly, this nucleotide substitution pattern mirrors that previously reported by Fairbrother et al. (2004a) for protein-coding ESEs. A similar analysis contrasting nucleotide substitution patterns between multiexonic and monoexonic lncRNA loci shows mutations preserving the ESEs and fewer ESE creating and ESE disrupting mutations within multiexonic lncRNA (Fig. 5C).

The identification of increased density and nucleotide conservation of ESEs at orthologous loci in mammalian species, as well as mutation and substitution biases affecting discrete sequences, indicate that even though lncRNA loci are rapidly evolving as a whole, constraint on short nucleotide sequences has acted, over relatively recent human evolution, on ensuring efficient splicing of these multiexonic lncRNAs.

Multiexonic lncRNA loci share with protein-coding genes several genomic and transcriptional features

The unexpected splicing and GC content similarities between human multiexonic lncRNAs and protein-coding genes motivated us to consider additional features, specifically those that have been attributed to transcriptional regulation. Protein-coding exons exhibit an unusually high level of binding to nucleosomes, perhaps because high GC-content DNA

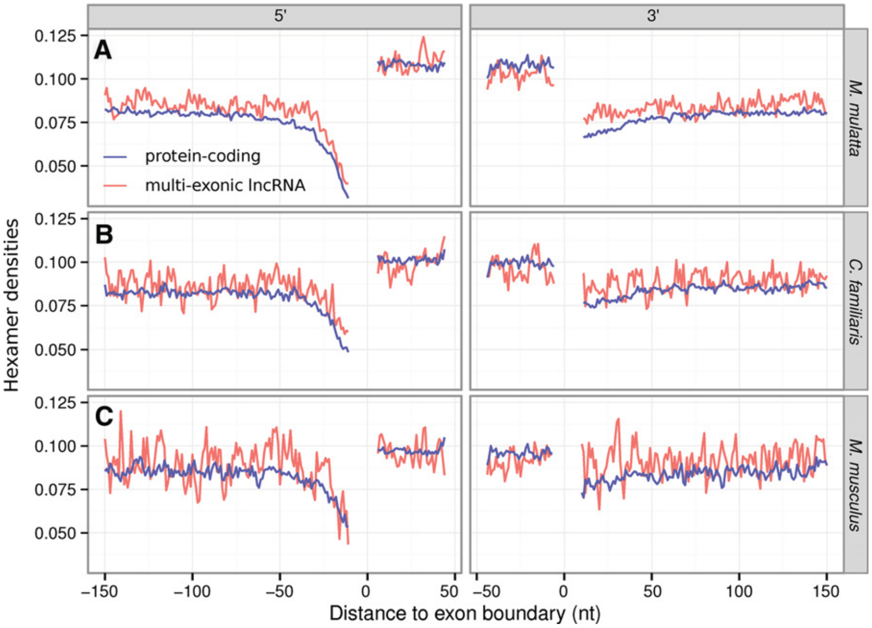

FIGURE 4. Proportion of hexamers that are predicted exonic splicing enhancers (Fairbrother et al. 2002; Zhang and Chasin 2004; Goren et al. 2006) within (*A*) *Macaca mulata*, (*B*) *Canis familiaris*, (*C*) *Mus musculus* genomic sequences that align to predicted exonic splicing enhancers in human intergenic lncRNA exons.

structure is less rigid, and they also appear to slow the procession of RNA polymerase II, perhaps to assist in the proper recognition of regulatory elements by the splicing machinery (Schwartz et al. 2009; Gelfman and Ast 2013). Coding exons tend also to have higher fractions of CpG dinucleotides that are methylated than introns (Gelfman et al. 2013), and higher densities of trimethylated histone 3 lysine 36 (H3K36me3) marks (Schwartz et al. 2009; Luco et al. 2010).

Using data for the human lymphoblastoid cell line Gm12878 (ENCODE Project Consortium et al. 2012), we found that lncRNA exons exhibit elevations in nucleosome binding, and RNA polymerase II read densities relative to their introns (Fig. 6A,B). In contrast, the density of H3K36me3 marks, is only marginally higher for multiexonic lncRNA exons than introns (Fig. 6C). In comparison to a set of protein-coding genes selected to match the expression levels of lncRNAs, we observed comparable nucleosome binding at the lncRNA exon–intron boundary, and lower density of H3K36me3 marks. Surprisingly, we found the density of Pol II reads to be higher for lncRNA loci than expression level matched protein-coding sequences. This increased density could reflect greater polymerase pausing at the lncRNA exons relative to protein-coding exons or highlight the greater instability of lncRNA transcripts (Clark et al. 2012). In contrast to protein-coding exons, we did not observe an increased proportion of methylated CpG within lncRNA exon boundaries. This is likely the consequence of the low number of lncRNA loci identified in this data set (604) leading a reduced power to detect such events, because a significant enrichment is observed when using the full lncRNA annotation regardless of the expression calls in H1 cells (Supplemental Fig. 6; Lister et al. 2009).

In summary, multiexonic lncRNA loci are associated with signatures of enhanced splicing efficiency, some of which are similar to those previously observed in protein-coding genes, which indicates their transcripts' functionality.

Evolution of intergenic lncRNA exon nucleotide composition

We then sought to analyze the evolutionary forces shaping the nucleotide composition of intergenic multiexonic lncRNAs. To do so we analyzed both the mutation and substitution patterns (A,T → G,C and G,C → A,T) at different genomic features. If base composition is at equilibrium, we expect sequences to have similar ratios of polymorphism to

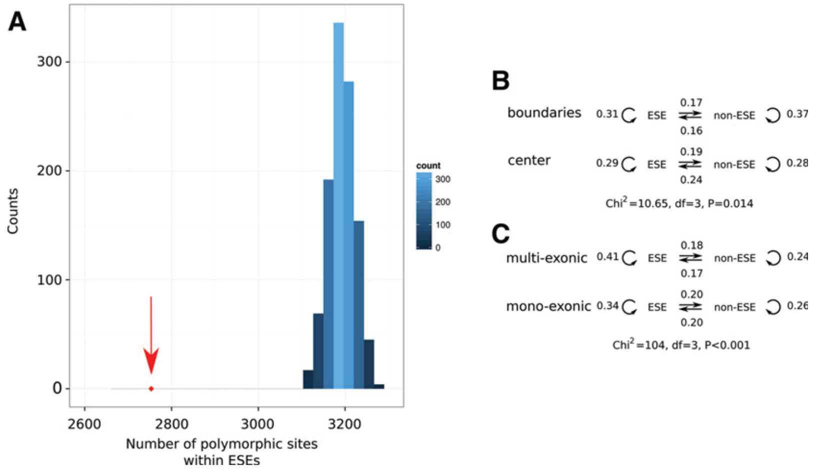

FIGURE 5. (*A*) Comparison of the observed (2743, red arrow) and the number of polymorphic sites within ESE sequences (predicted as per Fairbrother et al. 2002; Zhang and Chasin 2004; Goren et al. 2006) expected by simulation within human intergenic multiexonic lncRNAs. The distribution of expected number of polymorphic sites within ESEs is based on 1000 randomizations of SNPs within multiexonic lncRNA exons accounting for base composition. (*B*) Comparison of the nucleotide substitution patterns with respect to ESE motifs within 20 nt of exon boundaries and 20 nt located at the center of lncRNA exons. (*C*) Comparison of the nucleotide substitution patterns with respect to ESE motifs identified within multiexonic and monoexonic lncRNAs.

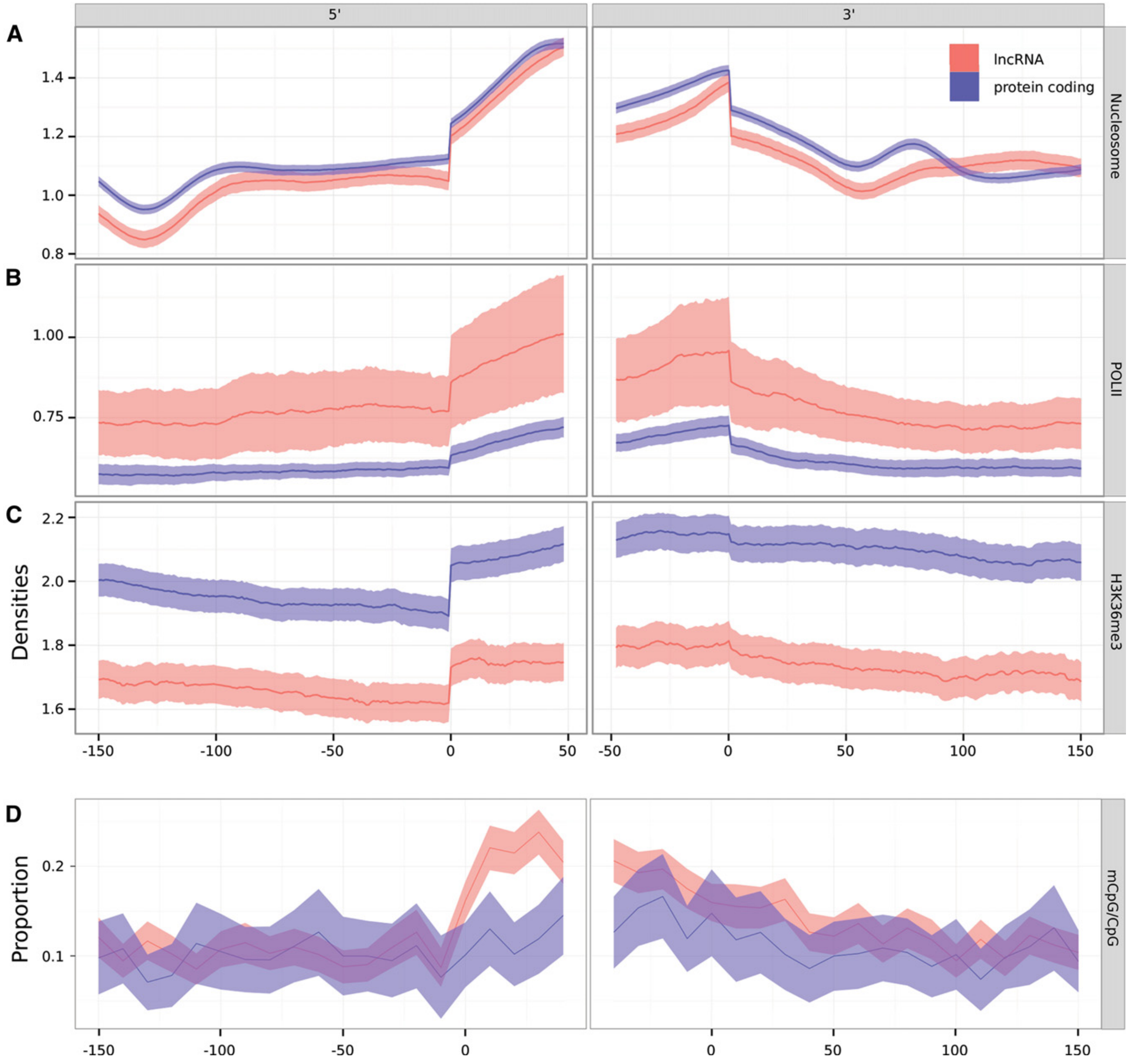

FIGURE 6. Comparison of (*A*) nucleosome, (*B*) RNA polymerase II (POLII), and (*C*) H3K36me3 reads densities, and (*D*) the proportion of methylated CpG (mCpG) over the number of CpG at the 5′ and 3′ boundaries of intermediate protein-coding (blue) and intergenic lncRNA (red) exons.

divergence for the different classes. Any departure would be indicative of action of nonselective (biased gene conversion) or selective forces acting on the nucleotide composition (Haddrill and Charlesworth 2008). Our analysis focused on *Drosophila* (Mackay et al. 2012), a species chosen owing to its larger effective population size and therefore greater efficiency of selection than human (Eyre-Walker et al. 2002; Li and Durbin 2011).

As expected from previous studies in *D. melanogaster* we observed a general excess of GC → AT relative to AT → GC that is indicative of the AT biased mutation rate in this species (DuMont et al. 2009; Singh et al. 2009). Most importantly we observed strong differences between nondegenerate and lncRNA exonic sites relative to intronic sites (Fig. 7). LncRNA exons and nondegenerate sites are characterized by a significant excess of polymorphism over divergence (χ^2 test, $P < 0.001$) (Fig. 7) which indicates the action of purifying selection (Haerty and Ponting 2013). Furthermore for these sites we found the ratio polymorphism$_{GC \rightarrow AT}$/divergenc$_{GC \rightarrow AT}$ to be significantly greater than polymorphism$_{AT \rightarrow GC}$/divergenc$_{AT \rightarrow GC}$. These results could be interpreted as the consequence of purifying selection acting on nucleotide composition that disfavors the fixation of G,C to A,T mutations within multiexonic lncRNA exons. High GC content thus tends to confer functional benefit on spliced lncRNA exons in fruit flies.

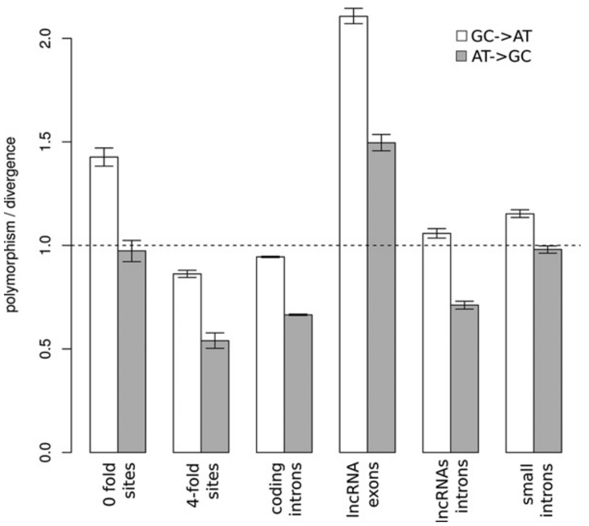

FIGURE 7. Comparison of polymorphism to divergence ratio depending upon mutation bias (AT → GC, GC → AT) between different genomic categories. Small introns include introns from protein-coding genes that are <86 nt.

Finally, we sought to provide evidence for the reasons underlying the elevated GC content of human multiexonic lncRNA exons. After taking into account, using partial correlation analysis, the potential association among multiple factors (recombination rate, intron size, number of exons, expression level) (Supplemental Fig. 7), GC content remained weakly but significantly positively correlated with multiexonic lncRNA expression levels, specifically their maximum expression values across 16 tissues ($P < 0.001$) or their tissue expression breadth ($P < 0.001$). High GC sequence in protein-coding exons has been associated with greater transcriptional activity (Kudla et al. 2006). Our results thus imply that human multiexonic lncRNAs exhibit protein-coding gene-like signatures of selection for enhanced transcriptional and splicing efficiencies.

DISCUSSION

Human multiexonic lncRNA loci are very modestly constrained in their exons, relative to their introns, which we have interpreted as implying either that their functions contribute little to organismal fitness or that their functionality is conveyed by only a small minority of their sequences (Ponjavic and Ponting 2007; Haerty and Ponting 2013). The first of these two possibilities is further suggested by the frequent lack of conserved transcription of orthologous lncRNAs across diverse mammals (Kutter et al. 2012; Necsulea et al. 2014). However, the latter possibility is supported by our observations that, akin to protein-coding genes, (i) evolutionary constraint is more concentrated near to human spliced lncRNA intron–exon boundaries, (ii) such regions contain an unusually high density of ESEs, and (iii) these ESEs are unexpectedly preserved in orthologous sequence in sequence-divergent mammals. Consequently, in contrast to a previous study which indicated that transcription of predominantly monoexonic lncRNAs is not conserved in a single tissue, adult liver, across eutherian mammals (Kutter et al. 2012), our findings indicate that transcription and efficient splicing of multiexonic lncRNAs will often be conserved across mammals, perhaps in spatiotemporally distinct tissues.

Intergenic lncRNA exons are GC rich compared with introns or monoexonic intergenic lncRNAs

Elevated GC content within exonic sequences has long been a hallmark of protein-coding sequences. Unexpectedly, we showed that this signature is also shared by intergenic multiexonic lncRNAs across five eukaryotic species although to a lower extent than the variation observed between protein-coding exons and introns.

However, multiple neutral processes could also explain the observed elevated GC content within lncRNA exons and in many cases their effects have been wrongly ascribed to the action of selection. The most important mechanism involves GC-biased gene conversion, a process by which, during recombination, the mismatch repair mechanism favors the propagation of G,C over A,T alleles. GC-biased gene conversion within lncRNA loci would be consistent with the positive correlation of genomic GC content and recombination rate (Duret and Galtier 2009) as we observed for lncRNA loci. Nevertheless, because GC-biased gene conversion does not differentiate between exons and introns, it alone cannot explain their different GC contents. For the same reason, transcription-coupled repair (Polak et al. 2010), whose effects on nucleotide content extend across complete transcribed loci, also cannot explain the variation in GC content between exons and introns.

Higher GC content within protein-coding exons has been proposed to be a consequence of selection on the efficiency or robustness of translation (Akashi 1995; Drummond and Wilke 2009; Tuller et al. 2010). The vast majority of the lncRNA loci that we analyzed are unlikely to be translated (Guttman et al. 2013; Wilhelm et al. 2014), hence their high GC exonic content cannot be explained in this manner. Increased lncRNA exon GC content might also be ascribed to selection for the formation and the maintenance of GC-rich RNA secondary structures. Previous analyses reported enrichment for conserved secondary structures within lncRNA (Smith et al. 2013) and a correlation between folding propensity and lncRNA expression (Managadze et al. 2011). The major difficulty still resides in the high false discovery rate associated with de novo predictions. However, the experimental identification of regions involved in secondary structures either using Parallel Analysis of RNA Structures (PARS), Parallel Analysis of RNA structures with Temperature Elevation (PARTE), or ds/ssRNA-seq (for review, see Mortimer et al. 2014) should overcome this issue in the future.

Finally, the elevated GC content within exonic sequences could also reflect that efficient transcription of these loci is under selection across multiple species. The positive correlation we found between nucleotide composition and expression levels of human intergenic lncRNAs is consistent with this hypothesis. Previously Kudla et al. (2006) showed that manipulation of the GC3 content of a sequence dramatically affects its transcription and translation rates in human cells.

In comparison to multiexonic lncRNAs, we failed to observe elevated GC content in large numbers of proposed monoexonic lncRNAs, for example the 45,905 (89.2%) that have single exons in the set of Hangauer et al. (2013). These models also show significantly weaker biases in nucleosome binding, polymerase occupancy or H3K36me3 occupancy relative to multiexonic loci. Consequently, we conclude either that the GC content profile of multiexonic lncRNAs can be explained solely by their splicing requirements or that sets of monoexonic lncRNAs contain false-positive predictions that greatly outnumber well-described monoexonic lncRNA such as *Malat1*, *Neat1*, and *Paupar* that have demonstrated cellular functions (Clemson et al. 2009; Bernard et al. 2010; Tripathi et al. 2010; Vance et al. 2014). Monoexonic loci can be wrongly predicted because of genomic DNA contamination in RNA sequencing libraries, or serendipitous transcription. It is also possible that such predictions represent RNAs with functions that are very different from multiexonic lncRNAs, such as those derived from transcription across active genomic regions such as enhancers (Marques et al. 2013).

Splicing-associated purifying selection in multiexonic lncRNAs

The mechanism of exon splicing within protein-coding transcripts has long been associated with specific chromatin marks (Schwartz et al. 2009; Luco et al. 2011) but most importantly with purifying selection on discrete exonic and intronic sequences within protein-coding genes in eukaryotes (Chamary et al. 2006; Warnecke and Hurst 2007). It is therefore not fully unexpected that we found enrichments at the lncRNA exon–intron boundaries of factors or features that previously have been associated with "optimal" splice site choice (Pol II, or H3K36me3 densities, Schwartz et al. 2009; Luco et al. 2010; Gelfman and Ast 2013).

We previously reported no significant evidence for purifying selection acting on human intergenic lncRNAs when studying full loci (Haerty and Ponting 2013). However, an absence of evidence for selection at the full gene model level does not preclude short functional regions selectively constrained. Indeed, lncRNA exonic sequences predicted to encode splicing regulatory elements show increased conservation and significant substitution biases against ESE disruption and ESE creation. Remarkably, the same patterns of elevated nucleotide conservation and substitution biases are well known for protein-coding sequences. For instance, within human populations, Fairbrother et al. (2004a) identified purifying selection acting on ESEs. The authors reported lower ESE disruption and ESE creation than expected, highlighting the effect of natural selection on these regulatory motifs. Additionally, similar to our conclusions of a greater conservation of ESEs close to lncRNA exon boundaries, the authors also showed that the selective pressure on ESEs was also stronger close to splice sites. Several subsequent analyses of ESEs within protein-coding sequences reached similar conclusions either using multispecies comparisons or larger SNV data sets in human (Parmley and Hurst 2007; Ke et al. 2008; Cáceres and Hurst 2013).

Overall, and following on previous analyses of splicing in protein-coding genes (Fairbrother et al. 2004b; Schwartz et al. 2009; Amit et al. 2012; Gelfman and Ast 2013) and despite low selective constraints detected at the full locus level (Ponjavic et al. 2007; Marques and Ponting 2009; Ulitsky and Bartel 2013), we have identified strong indications that splicing of multiexonic lncRNA loci is likely to often be required for their function. These include (1) a significantly increased level of nucleotide conservation across primates for human lncRNA exonic sequence near to their intron boundaries relative to their internal regions (see Haerty and Ponting 2014) and (2) a significant depletion of polymorphic sites within predicted ESE motifs. These observations are surprising because intergenic lncRNA sequence tend to be very poorly conserved, and evidence of selection at the whole locus level is at the best weak (Marques and Ponting 2009; Haerty and Ponting 2013).

Our results indicate that these lncRNA loci predominantly possess spliced RNA sequence-dependent functions that are conveyed by only minor proportions (<5%) of their exonic sequence (Ponjavic et al. 2007); moreover, as for protein-coding loci, their regulatory sequence and composition appear to have been under selection for efficient transcription and splicing. Together, our observations imply that multiexonic lncRNA loci often convey spliced RNA-dependent functions that are widely conserved among mammals. Our findings cannot, however, distinguish between RNA exon sequence-dependent functions that act locally, near to their sites of synthesis, and those that act more distally, such as chromatin guides and scaffolds (Kung et al. 2013) or competitive endogenous lncRNAs (Marques et al. 2011).

Our observations thus better discriminate functional lncRNAs and identify functional elements, namely splice sites, which can now be targeted for disruption, for example, using CRISPR/Cas9 technology in high-throughput phenotypic assays.

MATERIALS AND METHODS

To allow comparisons with protein-coding sequences, whose compositional features are well-established, we focused our analyses on intergenic lncRNAs. Antisense, overlapping, and intronic lncRNAs were all discarded from published lncRNA data sets prior to analyses.

Long intergenic noncoding RNA data sets

Intergenic lncRNAs were acquired from published sets (Table 1) for the fruit fly (*Drosophila melanogaster*), zebrafish (*Danio rerio*), coelacanth (*Latimeria chalumnae*), mouse (*Mus musculus*) and human (*Homo sapiens*). We considered a nonredundant, nonoverlapping set of human intergenic lncRNAs identified by Cabili et al. (2011), the ENCODE consortium (Derrien et al. 2012) or Hangauer et al. (2013) resulting in a total of 61,001 intergenic lncRNAs. We note that the Hangauer et al. data set is comprised of 45,905 (89.2%) single exon lncRNA models.

In order to avoid contamination in lncRNA sets derived from protein-coding sequences either through gene duplication or pseudogenization that might inflate their computed nucleotide content, we removed, prior to analyses, any locus that shared weak or strong sequence similarity with annotated protein sequences (as detected using BLASTN ($E < 0.1$) or whose genomic loci overlapped (in multigenome alignments, UCSC genome database http://genome.ucsc.edu/) with annotated protein-coding loci in other vertebrates species.

Nucleotide sequence composition

All transcripts in each lncRNA locus were collapsed in order to derive their maximal extended exons. Exons (or introns) of lncRNA loci or their closest genomically neighboring protein-coding genes were partitioned into their first, middle, or last or their sole exons (single exon gene models) and their G + C proportions computed within windows that each represents a 10% portion (decile) of exonic sequence. This procedure was repeated for introns of multiexonic lncRNAs. It was also repeated for protein-coding genes, but only for those genes lying in genomic sequence adjacent to intergenic lncRNA loci, in order to match for nucleotide composition which is known to vary on the 100 kb–1 Mb scale. We also ran the same analysis on the flanking intergenic sequences after masking 500 nt adjacent to annotations. Our analyses made use of the hg19, mm10, danRer7, latCha1, and dm3 versions of genome assemblies, all acquired from the UCSC genome database.

Exonic splicing regulatory elements

We implemented the RESCUE-ESE algorithm developed by Fairbrother et al. (2002, 2004b) to identify hexamers that are significantly enriched or depleted within exonic sequences relative to their flanking intronic sequences. We focused our analysis on all internal ("middle") lncRNA exons that are longer than 100 nt and flanked on both sides by introns including at least 400 nt. Following Fairbrother et al. (2002) and Yeo et al. (2004), prior to analysis we masked nucleotides flanking the splice sites (5 nt for the 5′ exonic, intronic [donor], and 3′ exonic sequences and 20 nt for the 3′ intron [acceptor site]).

We took advantage of sets of 607 exonic splicing enhancer hexamers (ESEs) previously predicted by Fairbrother et al. (2002), Zhang and Chasin (2004) or Goren et al. (2006) to identify the density of predicted splicing regulatory elements in exonic sequence lying adjacent to exon–intron boundaries for internal exons within human lncRNA loci or protein-coding genes. Fifty nucleotides of exonic sequence and 150 nt of intronic sequence were considered: The first five exonic nucleotide and 10 intronic nucleotides were discarded from our analyses because of the string composition bias associated with these splice sites. In order to assess the rate of cross-species conservation of these regulatory sequences we used UCSC liftOver files (obtained from http://genome.ucsc.edu/) to project the human lncRNA sequence onto orthologous macaque (rheMac3 assembly), dog (canFam3), and mouse (mm10) genomes and calculated the density of predicted ESE motifs identified within these orthologous sequences.

Epigenetic marks associated with splicing

Instead of considering predictions of nucleosome occupancy (Kaplan et al. 2009), which depend on nucleotide composition, we took advantage of experimental evidence for nucleosome binding (ENCODE Project Consortium et al. 2012).

Multiple factors have been proposed to regulate the splicing of exons in mammals (Brown et al. 2012). We used the signal tracks generated by the ENCODE consortium (ENCODE Project Consortium et al. 2012) to quantify the enrichment or depletion of nucleosome location, RNA polymerase II reads and H3K36me3 marks at multiexonic lncRNA exon–intron boundaries. These data were acquired in the Gm12878 cell line and only a subset was considered, namely for those loci that had uniquely mapped reads from RNA-seq for this cell line. For comparison, we performed the same analyses for all protein-coding genes expressed in either of these two cell lines. The difference in occupancy or enrichment between exonic and intronic sequences was tested using a Mann–Whitney test, focusing on the 50 exonic and 50 intronic nucleotides flanking each splice site. Because of the large difference in expression levels between protein-coding and lncRNA genes, we selected a subset of 2642 protein-coding sequences whose gene expression distribution in Gm12878 matches the expression distribution of lncRNA loci.

The proportion of methylated CpG dinucleotides near to lncRNA exon–intron boundaries was computed for each exonic position using data generated by Lister et al. (2009) (http://neomorph.salk.edu/human_methylome/data.html). As before, the protein-coding genes were sampled to match the expression of the lncRNAs in the H1 cell line (Lister et al. 2009).

lncRNA composition and constraint

Factors such as recombination rate, gene structure, or gene expression level have previously been shown to correlate with protein-coding nucleotide composition (Chamary et al. 2006). Consequently, we used the corpcor partial correlations package from R (http://strimmerlab.org/software/corpcor/) to test, using a partial correlations approach, for associations between these genomic factors and GC content within lncRNA loci in fruit fly and/or human genomes. Significance of these partial correlations was assessed through randomization of values while keeping one parameter constant. For computation of partial correlations, we collected for sequence intervals their (i) recombination rate values (Fiston-Lavier et al. 2010; Kong et al. 2010), (ii) GC content for exons and introns, (iii) numbers of introns, (iv) expression level from 16 organs as part of the Human Body Map, and (v) the expression breadth across tissues (τ) (Larracuente et al. 2008).

To assess the distribution of polymorphic sites within lncRNA exons with respect to splicing regulatory elements we used polymorphism data from the 1000 Genomes Consortium (1000 Genomes Project Consortium et al. 2012). The SNV density within exonic spliced enhancers (ESEs, 607 motifs, Fairbrother et al. 2002; Zhang and Chasin 2004; Goren et al. 2006) was then compared with the expected SNV density based on 1000 random samples from the same sequences taking into account nucleotide composition. In order to also account for mutation biases associated with dinucleotide composition, the resampling analysis was performed again conserving the dinucleotide composition of the sequences.

Additionally, using the pairwise alignments between the human genome and the chimpanzee or the macaque genomes (UCSC genome database http://genome.ucsc.edu/), we identified the substitutions occurring within the human lineage using maximum parsimony and assessed the number of substitutions within the human lineage that conserve ESE status (ESEs → ESEs) versus those that either create or erase an ESE. We compared the substitution patterns within the 20 nt flanking the 5′ and 3′ exon boundaries to the middle regions (20 nt) of lncRNA exons longer than 100 nt. The differences between internal regions and exon boundaries were tested using a χ^2 test with 3 degrees of freedom.

We assessed potential selection on nucleotide composition within intergenic lncRNAs by implementing a test developed by Haddrill and Charlesworth (2008) derived from the McDonald and Kreitman test (McDonald and Kreitman 1991). At equilibrium, the following equality is expected:

$$\frac{\text{Polymorphism}_{\text{GC}\rightarrow\text{AT}}}{\text{Divergence}_{\text{GC}\rightarrow\text{AT}}} = \frac{\text{polymorphism}_{\text{AT}\rightarrow\text{GC}}}{\text{divergence}_{\text{AT}\rightarrow\text{GC}}}.$$

Any significant deviation could be attributed to the action of selection or other nonselective processes such as GC gene biased conversion (Haddrill and Charlesworth 2008).

To infer, using parsimony, the nucleotide composition of ancestral sequences, we used the genomic alignments of *D. melanogaster* with either *D. simulans* or *D. yakuba*. Sites whose ancestral state could not be inferred were discarded. We used these ancestral sequences to infer AT → GC and GC → AT directional substitution rates within lncRNA exons and introns (DuMont et al. 2009). For comparison we also computed rates for small introns (<86 nt) as a proxy for neutrally evolving sequences (Clemente and Vogl 2012).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

W.H. and C.P.P. are supported by the Medical Research Council, and this work was also supported by a European Research Council Advanced Grant to C.P.P.

Received July 16, 2014; accepted November 25, 2014.

REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139:** 1067–1076.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496:** 311–316.

Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, Lev-Maor G, Burstein D, Schwartz S, Postolsky B, et al. 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1:** 543–556.

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

Babak T, Blencowe BJ, Hughes TR. 2007. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* **8:** 33.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33:** 981–993.

Belgard TG, Marques AC, Oliver PL, Abaan HO, Sirey TM, Hoerder-Suabedissen A, García-Moreno F, Molnár Z, Margulies EH, Ponting CP. 2011. A transcriptomic atlas of mouse neocortical layers. *Neuron* **71:** 605–616.

Bellucci M, Agostini F, Masin M, Tartaglia GG. 2011. Predicting protein associations with long noncoding RNAs. *Nat Methods* **8:** 444–445.

Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, Zhang MQ, Sedel F, Jourdren L, Coulpier F, et al. 2010. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J* **29:** 3082–3093.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S. 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved orf and located in the nucleus. *Cell* **71:** 515–526.

Brown SJ, Stoilov P, Xing Y. 2012. Chromatin and epigenetic regulation of pre-mRNA processing. *Hum Mol Genet* **21:** R90–R96.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915–1927.

Cáceres EF, Hurst LD. 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol* **14:** R143.

Chamary JV, Hurst LD. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet* **21:** 256–259.

Chamary JV, Parmley JL, Hurst LD. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7:** 98–108.

Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, Moscato P, Dinger ME, Mattick JS. 2012. Genome-wide analysis of long noncoding RNA stability. *Genome Res* **22:** 885–898.

Clemente F, Vogl C. 2012. Unconstrained evolution in short introns?—an analysis of genome-wide polymorphism and divergence data from *Drosophila*. *J Evol Biol* **25:** 1975–1990.

Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, Lawrence JB. 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol Cell* **33:** 717–726.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775–1789.

Drummond DA, Wilke CO. 2009. The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* **10:** 715–724.

DuMont VLB, Singh ND, Wright MH, Aquadro CF. 2009. Locus-specific decoupling of base composition evolution at synonymous sites and introns along the *Drosophila melanogaster* and *Drosophila sechellia* lineages. *Genome Biol Evol* **1:** 67–74.

Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10:** 285–311.

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet* **2:** 549–555.

Eyre-Walker A, Keightley PD, Smith NG, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol Biol Evol* **19:** 2142–2149.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297:** 1007–1013.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004a. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol* **2:** E268.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004b. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* **32:** W187–W190.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. *Gene* **463:** 18–20.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Gelfman S, Ast G. 2013. When epigenetics meets alternative splicing: the roles of DNA methylation and GC architecture. *Epigenomics* **5:** 351–353.

Gelfman S, Cohen N, Yearim A, Ast G. 2013. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon–intron structure. *Genome Res* **23:** 789–799.

Goren A, Ram O, Amit M, Keren H, Lev-Maor G, Vig I, Pupko T, Ast G. 2006. Comparative analysis identifies exonic splicing regulatory sequences–the complex definition of enhancers and silencers. *Mol Cell* **22:** 769–781.

Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. 2013. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154:** 240–251.

Haddrill PR, Charlesworth B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in *Drosophila*. *Biol Lett* **4:** 438–441.

Haerty W, Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biol* **14:** R49.

Haerty W, Ponting CP. 2014. No gene in the genome makes sense except in the light of evolution. *Annu Rev Genomics Hum Genet* **15:** 71–92.

Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9:** e1003569.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366.

Kay GF, Penny GD, Patel D, Ashworth A, Brockdorff N, Rastan S. 1993. Expression of Xist during mouse development suggests a role in the initiation of X chromosome inactivation. *Cell* **72:** 171–182.

Ke S, Zhang XH, Chasin LA. 2008. Positive selection acting on splicing motifs reflects compensatory evolution. *Genome Res* **18:** 533–543.

Kelley RL, Meller VH, Gordadze PR, Roman G, Davis RL, Kuroda MI. 1999. Epigenetic spreading of the *Drosophila* dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98:** 513–522.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467:** 1099–1103.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* **4:** e180.

Kung JT, Colognori D, Lee JT. 2013. Long noncoding RNAs: past, present, and future. *Genetics* **193:** 651–669.

Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet* **8:** e1002841.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in *Drosophila*. *Genome Biol* **12:** R118.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends Genet* **24:** 114–123.

Latos PA, Pauler FM, Koerner MV, Şenergin HB, Hudson QJ, Stocsits RR, Allhoff W, Stricker SH, Klement RM, Warczok KE, et al. 2012. Airn transcriptional overlap, but not its lncRNA products, induces imprinted igf2r silencing. *Science* **338:** 1469–1472.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475:** 493–496.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462:** 315–322.

Louie E, Ott J, Majewski J. 2003. Nucleotide frequency variation across human genes. *Genome Res* **13:** 2594–2601.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. *Science* **327:** 996–1000.

Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T. 2011. Epigenetics in alternative pre-mRNA splicing. *Cell* **144:** 16–26.

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* **482:** 173–178.

Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. 2011. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol* **3:** 1390–1404.

Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10:** R124.

Marques AC, Tan J, Ponting CP. 2011. Wrangling for microRNAs provokes much crosstalk. *Genome Biol* **12:** 132.

Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR, Ponting CP. 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol* **14:** R131.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.

Mortimer SA, Kidwell MA, Doudna JA. 2014. Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15:** 469–479.

Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C. 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res* **20:** 4255–4262.

Nam JW, Bartel DP. 2012. Long noncoding RNAs in *C. elegans*. *Genome Res* **22:** 2529–2540.

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.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. *Mol Biol Evol* **24:** 1600–1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. *PLoS Biol* **5:** e14.

Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, et al. 2012. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22:** 577–591.

Polak P, Querfurth R, Arndt PF. 2010. The evolution of transcription-associated biases of mutations across vertebrates. *BMC Evol Biol* **10:** 187.

Ponjavic J, Ponting CP. 2007. The long and the short of RNA maps. *Bioessays* **29:** 1077–1080.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long noncoding RNAs. *Cell* **136:** 629–641.

Qureshi IA, Mehler MF. 2012. Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat Rev Neurosci* **13:** 528–541.

Schwartz S, Meshorer E, Ast G. 2009. Chromatin organization marks exon–intron structure. *Nat Struct Mol Biol* **16:** 990–995.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila. Mol Biol Evol* **26:** 1591–1605.

Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* **41:** 8220–8236.

Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, Freier SM, Bennett CF, Sharma A, Bubulya PA, et al. 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39:** 925–938.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141:** 344–354.

Ulitsky I, Bartel DP. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154:** 26–46.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147:** 1537–1550.

van Heesch S, van Iterson M, Jacobi J, Boymans S, Essers PB, de Bruijn E, Hao W, Macinnes AW, Cuppen E, Simonis M. 2014. Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol* **15:** R6.

Vance KW, Sansom SN, Lee S, Chalei V, Kong L, Cooper SE, Oliver PL, Ponting CP. 2014. The long non-coding RNA Paupar regulates the expression of both local and distal genes. *EMBO J* **33:** 296–311.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet* **47:** 97–120.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119:** 831–845.

Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol* **24:** 2755–2762.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the *cis*-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci* **110:** 11952–11957.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509:** 582–587.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377–394.

Yeo G, Hoon S, Venkatesh B, Burge CB. 2004. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc Natl Acad Sci* **101:** 15700–15705.

Yoo EJ, Cooke NE, Liebhaber SA. 2012. An RNA-independent linkage of noncoding transcription to long-range enhancer function. *Mol Cell Biol* **32:** 2020–2029.

Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, Ponting CP. 2012. Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol Evol* **4:** 427–442.

Zhang XH, Chasin LA. 2004. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev* **18:** 1241–1250.