Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved

ANNE NITSCHE,[1,2] DOMINIC ROSE,[3,4] MARIO FASOLD,[2,5] KRISTIN REICHE,[6,7] and PETER F. STADLER[1,2,7,8,9,10,11]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, D-04107 Leipzig, Germany
[2]Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany
[3]Bioinformatics Group, Department of Computer Science, University of Freiburg, D-79110 Freiburg, Germany
[4]MML, Munich Leukemia Laboratory GmbH, D-81377 München, Germany
[5]ecSeq Bioinformatics, D-04275 Leipzig, Germany
[6]Young Investigators Group Bioinformatics and Transcriptomics, Department of Proteomics, Helmholtz Centre for Environmental Research–UFZ, D-04318 Leipzig, Germany
[7]Department of Diagnostics, Fraunhofer Institute for Cell Therapy and Immunology–IZI, D-04103 Leipzig, Germany
[8]Max Planck Institute for Mathematics in the Sciences, D-04103 Leipzig, Germany
[9]Department of Theoretical Chemistry, University of Vienna, A-1090 Wien, Austria
[10]Center for non-coding RNA in Technology and Health, University of Copenhagen, DK-1870 Frederiksberg C, Denmark
[11]Santa Fe Institute, Santa Fe, New Mexico 87501, USA

ABSTRACT

Large-scale RNA sequencing has revealed a large number of long mRNA-like transcripts (lncRNAs) that do not code for proteins. The evolutionary history of these lncRNAs has been notoriously hard to study systematically due to their low level of sequence conservation that precludes comprehensive homology-based surveys and makes them nearly impossible to align. An increasing number of special cases, however, has been shown to be at least as old as the vertebrate lineage. Here we use the conservation of splice sites to trace the evolution of lncRNAs. We show that >85% of the human GENCODE lncRNAs were already present at the divergence of placental mammals and many hundreds of these RNAs date back even further. Nevertheless, we observe a fast turnover of intron/exon structures. We conclude that lncRNA genes are evolutionary ancient components of vertebrate genomes that show an unexpected and unprecedented evolutionary plasticity. We offer a public web service (http://splicemap.bioinf.uni-leipzig.de) that allows to retrieve sets of orthologous splice sites and to produce overview maps of evolutionarily conserved splice sites for visualization and further analysis. An electronic supplement containing the ncRNA data sets used in this study is available at http://www.bioinf.uni-leipzig.de/publications/supplements/12-001.

Keywords: long noncoding RNAs; lncRNA; splice sites; multiple sequence alignments; evolution; conservation; evolutionary plasticity

INTRODUCTION

The large genomes of higher eukaryotes are pervasively transcribed, although protein-coding sequence forms only a tiny fraction of the genome (Kapranov et al. 2007). A substantial portion of the transcriptome appears as mRNA-like nonprotein-coding transcripts (Maeda et al. 2006; The ENCODE Project Consortium 2007), although there is ample evidence for the existence of many other classes of transcripts, ranging from small structured ncRNAs (Washietl et al. 2005) to intronic transcripts (Louro et al. 2009), independently transcribed UTRs (Mercer et al. 2011), and giant "macroRNAs" (Kapranov et al. 2010; Hackermüller et al. 2014). Despite their abundance, the evolutionary history of these transcripts is still poorly understood. Apart from a few detailed case studies, global statistical analyses showed that, as a group, the mRNA-like ncRNAs are under stabilizing selection (Ponjavic et al. 2007; Guttman et al. 2009; Marques and Ponting 2009). The level of sequence conservation, however, is very low compared with other functional transcripts (Pang et al. 2006; Marques and Ponting 2009). As a consequence it is hard to identify homologs in genome-wide searches based on sequence similarity. The low levels of sequence conservation provide only very limited contrast between intronic and exonic parts, so that it is difficult at best to infer complete gene structures for orthologs. More importantly, stabilizing selection maintaining the small, well-conserved sequence

Corresponding author: studla@bioinf.uni-leipzig.de
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.046342.114.

elements often located within the gene boundaries of long noncoding RNAs (lncRNAs) cannot be unambiguously attributed to the RNA product. Instead, such "phylogenetic footprints" may just as well be functional at the DNA level, e.g., as enhancer. The *HOX* clusters may serve as a particularly impressive example. On the one hand, many well-characterized mRNA-like lncRNAs (Mainguy et al. 2007; Rinn et al. 2007), including *HOTAIR* (Rinn et al. 2007; Tsai et al. 2010), *HOTTIP* (Wang et al. 2011), and several microRNA precursors (Tanzer et al. 2005), are transcribed from the intergenic regions, on the other hand, the region is packed with conserved functional DNA elements (Lee et al. 2006; Punnamoottil et al. 2010; Natale et al. 2011). The observable conservation of genomic sequence thus does not in itself provide sufficient information to disentangle the evolutionary history of lncRNAs.

Beyond global sequence conservation, however, we can also utilize the conservation of gene structure, in particular the conservation of splice sites, to establish homology. Indeed, novel transcripts can be predicted successfully from multiple genome alignments based exclusively on conserved splice-site patterns (Hiller et al. 2009; Rose et al. 2011). A considerable fraction of the transcripts detected in this manner shows very little sequence conservation and resembles lncRNAs. Probably they would not have been detected based on sequence homology alone.

The rapid development of sequencing technology has made it feasible to obtain high coverage transcriptome data sets for a wide variety of cell and tissue types. In addition to the systematic efforts to exhaustively catalog the human transcriptome in the ENCODE project and large cDNA resource amassed by the FANTOM project (Suzuki and Hayashizaki 2004), rapidly growing resources are also becoming available for a diversity of model organisms. As a consequence, comparative transcriptomics approaches become feasible (see, e.g., Baldo et al. [2011] and Bräutigam et al. [2011] and the review Hashimshony and Yanai [2010]). A recent study demonstrated that 30%–40% of nearly 2000 human lncRNAs show conserved expression in rodents or ungulates (Washietl et al. 2014), based on direct comparison of transcriptome sequencing data for six mammalian species. A similar approach investigating 11 tetrapod species reported 11,000 primate-specific lncRNAs contrasted by 2500 highly conserved ones (Necsulea et al. 2014). These numbers are somewhat lower (19% of lncRNAs are older than primates), presumably because only one nonprimate mammal was included and direct, BLAST-based homology search was used in this study. A maximum likelihood approach to estimate the number of lncRNAs from publicly available data resulted in an estimate of 40,000–50,000 lncRNAs of which ~60%–70% are conserved between man and mouse (Managadze et al. 2013).

In this contribution we make use of genome-wide multiple sequence alignments together with transcriptomics data to construct a comparative map of splice sites. We then use this resource to systematically study the conservation patterns of lncRNAs and their gene structures.

MATERIALS AND METHODS

Genome and transcriptome data

We use two reference alignments: (1) the MULTIZ-based alignment (Blanchette et al. 2004) of 46 vertebrate genomes provided through the UCSC genome browser and (2) the EPO (Paten et al. 2008) multiple alignment of 12 eutherian mammals downloaded from ENSEMBL (Release 63). We reduced the latter alignment to those eight species for which ENSEMBL and UCSC utilize the same genome versions: *Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Equus caballus*, and *Canis familiaris*. In the following we will refer to these two multiple sequence alignments as the UCSC and the ENSEMBL alignment, respectively. (3) In order to investigate the conservation of zebrafish lncRNAs we use the eight-way zebrafish MULTIZ alignment (containing five teleosts, frog, mouse, and human) since the 46-way vertebrate alignment contains only sequences that are alignable to the human genome.

As basis set for transcripts we use a recent RefSeq track (10/2012, 40,373 transcripts) obtained from UCSC as well as the GENCODE v.14 collection of transcripts (Harrow et al. 2006). In addition we extracted all splice sites supported by at least one expressed sequence tag (EST) in the data collection of the UCSC genome browser (downloaded 08/2012). MicroRNA and snoRNA annotations were taken from ENSEMBL.

Comparative map of splice sites

The exon annotations from RefSeq and the EST collection define the coordinates of validated splice sites. Alignment blocks with fewer than 20 nt on intronic side of the splice site are omitted. This excludes too short introns, which are likely artifacts (Hong et al. 2006), and allows us to score splice-site quality.

For each validated splice site, we use a multiple sequence alignment to determine the corresponding (homologous) position in all other genomic sequences. This results in a collection of genomic positions that are known to be a functional splice site in at least one of the aligned species. For each splice-site position, we store for each of the aligned sequences, whether it is a donor or an acceptor, its `MaxEntScan` splice-site score (Yeo and Burge 2004) and information whether the potential splice site has been experimentally validated in a taxon that is included in the sequence alignment. A graphical representation of all splice sites in an interval of the genomic sequence of any of the included taxa can be generated using our web server (http://splicemap.bioinf.uni-leipzig.de); see Figure 1 for an example. We computed and evaluated splice-site maps separately from the UCSC and the ENSEMBL alignments.

Assessment of conservation rate via `MaxEntScan` scoring

The evolution of splice sites cannot be studied meaningfully based only on the annotated splice sites because the transcriptomes of many species are poorly covered in current databases, in particular

validated: RefSeq or >1 EST missing data

<-20 -10 -5 -1 1 2 3 4 5 10 15 >20

FIGURE 1. Splice-site map of the *GAS5* locus. Each line represents a splice site, each column a vertebrate genome arranged in increasing phylogenetic distance from human; MaxEntScan scores for splice-site quality are shown as grayscale; missing data are indicated as gray background.

in their noncoding regions. Therefore we use `MaxEntScan` scoring by Yeo and Burge (2004) to draw conclusions on the conservation rates, in the following way. In brief, `MaxEntScan` models short sequence motifs—here the donor and acceptor motifs of splice junctions—using a probabilistic model based on the "Maximum Entropy Principle." In contrast to position weight matrices or (inhomogeneous) Markov models, this makes it possible to account for both adjacent and non-adjacent dependencies between positions. The resulting gain in accuracy has been shown reliably predict missplicing mutations (Eng et al. 2004).

A splice site is "predicted" to have a functional ortholog, if there is an orthologous site in the relevant genome with a `MaxEntScan` score >3.0. This value is estimated from score distributions in Figure 2. It shows the distribution of donor and acceptor scores of all splice sites in the human lncRNA set as well as the scores of all aligned and all validated orthologous splice-site candidates in the UCSC human–mouse alignment. While the majority of known splice sites has a score >3, we observe a clearly bimodal distribution composed of a large peak conforming to functional splice sites and a second broad distribution of scores ≤3 belonging to positions that most likely have lost their capability of acting as splice donors or splice acceptors. We emphasize that the score cutoff >3 is restrictive and will tend to underestimate the number of conserved splice sites since the `MaxEntScan` scores are gauged so that sites with positive scores are more likely to be functional than not (Yeo and Burge 2004). This is also consistent with the comparison of predicted and validated splice sites of human and mouse coding regions in Table 1 below. A splice site counts as "validated" if it is confirmed by RefSeq or EST annotation. It is considered as "conserved" if it has a predicted and/or validated functional ortholog in the concerned genome. We refer to the supplement for more details on this method.

Conservation rates on the transcript level are derived from its splice sites. A transcript is considered conserved if at least one splice site of the human transcript corresponds to a predicted or validated splice site.

lncRNA transcripts

Since many RefSeq noncoding transcripts are associated with coding loci, we focus our analysis on a restrictively filtered subset of the GENCODE data to ensure conservative estimates of lncRNA conservation. In order to prepare a high-quality set of human lncRNAs we started from the 21,271 well-characterized "Gencode v14 lncRNA" transcripts and applied a series of filtering steps.

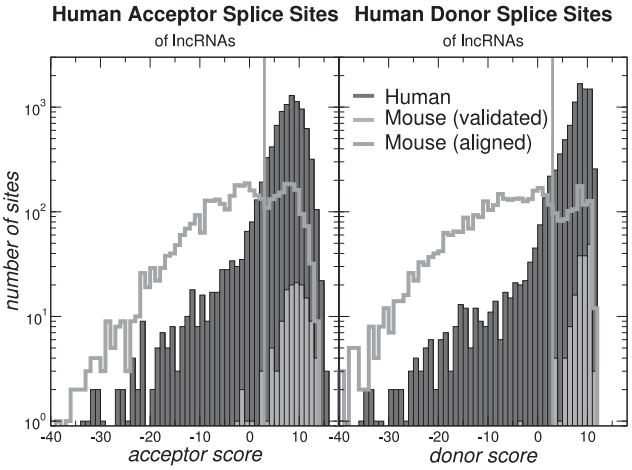

FIGURE 2. Conservation of splice sites of human lncRNAs in the mouse. Filled curves designate the distributions of MaxEntScan scores for human splice sites (dark gray) and orthologous positions that are known to be splice sites in mouse (light gray). The score distribution of all orthologous positions in mouse (gray) is a superposition of conserved functional splice sites and positions that have been destroyed by substitutions. The cut-off value of 3.0 is indicated by a thick light gray line.

TABLE 1. Conservation of splice sites between human and mouse

Data Set	Human *N*	Mouse Aligned	Mouse Predicted	Mouse Validated	Mouse Conserved
RefSeq coding	355,573	340,327	325,323	326,401	333,661
RefSeq 5′ UTR	16,035	11,737	8200	6908	8339
RefSeq 3′ UTR	1124	828	680	607	693
GENCODE lncRNAs	17,163	7339	2179	295	2188
miRNA host	602	282	105	40	108
snoRNA host	335	141	83	46	85

We report the conservation of splice sites for different annotation sets. We give an overview on the total number (*N*) of splice sites present in human, the number of aligned, predicted, validated, and total number of conserved splice sites.

A set of mouse lncRNAs involved in the circuitry controlling pluripotency and differentiation is described in Guttman et al. (2011). It consists of 2076 spliced transcripts with 6975 splice sites, of which 77% are also validated by EST or RefSeq data.

A conservative set of 1133 lncRNAs expressed in zebrafish embryos was recently reported in Pauli et al. (2012). A second, smaller set of 691 zebrafish lincRNAs is expressed in brain development (Ulitsky et al. 2011); only 449 of them are spliced. Since the overlap of the two sets is small, we consider their union consisting of 1508 spliced transcripts with 5415 splice sites.

We discarded transcripts that overlapped within annotated protein-coding sequences or pseudogenes in sense or antisense direction annotated by at least one of GENCODE, ENSEMBL, UCSC, or RefSeq. For GENCODE, we could rely on the annotation with biotype classification for transcripts and genes. In the case of ENSEMBL, RefSeq, and UCSC we used the annotation of coding exons. Since some of the transcripts overlapping in sense-direction might just be noncoding isoforms of protein-coding transcripts, we opted to remove them. We also excluded transcripts located in antisense direction of these coding sequences since conservation of the coding sequence also constrains the sequence of the opposing transcripts, even though they are annotated as noncoding. We used `RNAcode` (Washietl et al. 2011), a tool that efficiently detects conserved open reading frames in multiple sequence alignments, and TBLASTN (Altschul et al. 1990) to remove transcripts with putative coding regions. We only kept those transcripts that did not contain exons overlapping with significant `RNAcode` hits (*P*-value <0.05) or, if an exon could not be scored by `RNAcode` due to low sequence conservation, TBLASTN hits (*E*-value <0.05). We also removed all unspliced entries. At this stage we retained 5703 transcripts. The last filtering step included the application of `PhyloCSF` (Lin et al. 2011). All remaining transcripts with a `PhyloCSF` score >100 and a possible ORF of length ≥30 were sorted out. These cutoffs were chosen accordingly to Cabili et al. (2011). This affected another 290 transcripts. Our final data set thus contains 5413 transcripts with 17,163 splice sites.

We note that this lncRNA data set exhibits substantial overlap with the integrative compilation of 14,274 spliced human noncoding transcripts from different sources covering 24 tissues and cell types by Cabili et al. (2011). Three thousand one hundred forty-five of them are identically (99% reciprocal strand-specific overlap) represented in our set; the agreement increases to 3924 loci when a sequence overlap of at least 70% is required. We will refer to this collection of lncRNAs as the Cabili data set.

An important subclass of spliced lncRNAs with a well-understood function are host genes of microRNAs and snoRNAs. We thus identified lncRNAs that overlapped known microRNAs and snoRNAs as annotated by ENSEMBL. This resulted in 128 transcripts hosting microRNAs (containing 602 unique splice sites) and 73 transcripts hosting snoRNAs (335 unique splice sites). Interestingly, snoRNA host genes and, to a lesser extent also microRNA host genes, on average have more introns than other lncRNAs (3.7 versus 2.9 versus 2.0 introns/transcript in all lncRNAs).

RESULTS

Predicted conservation of protein-coding splice sites shows specificity of the method

Table 1 and Figure 3 summarize the conservation of splice sites between human and mouse, calculated in the described way. Similar data are observed for other mammalian species (see Supplemental Table S4). The overwhelming majority of splice sites in the RefSeq data set delimits coding exons. More than 95% of these are alignable, and nearly 92% have experimentally validated orthologous splice sites in the mouse. A comparison of conserved splice sites that are experimentally confirmed with computationally predicted ones shows that both sets almost perfectly coincide; in fact, the predicted set is even slightly smaller than the validated one, emphasizing that the score cutoff of 3.0 is highly specific.

Only a small fraction of the RefSeq splice sites falls into UTRs, with >14-fold difference between 5′ and 3′ UTRs.

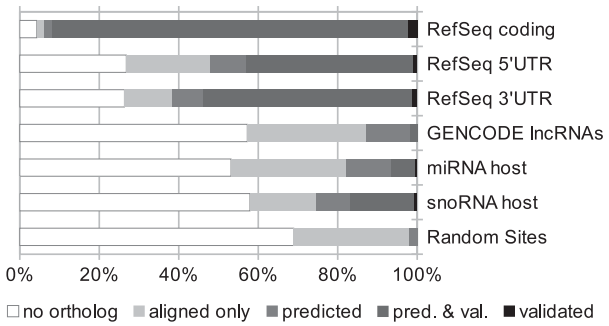

FIGURE 3. Conservation of splice sites between human and mouse in different contexts. In nonwhite the fraction of all alignable splice sites is shown. Dark gray scales display the estimated conservation rate. Consequently, the fraction of alignable but likely nonconserved splice sites is shown in light gray. In protein-coding RNAs 95% of the splice sites are at least alignable to mouse, and of those almost all are conserved, while in lncRNAs the rate of alignable sites drops to ~40%. The fraction of validated splice sites among the predicted ones turned from nearly 98% to only 13%, indicating that there are a lot of unannotated splice sites.

Only about three quarters of these regions are aligned between human and mouse in the UCSC alignments. Still, most of the predicted splice sites are backed up by experimental data. The strong depletion of introns in the 3′ UTRs has been described previously and can be explained as a consequence of nonsense-mediated decay (NMD) or a larger tolerance for intron retention (see, e.g., Scofield et al. [2007]).

Conservation of splice sites provides lower bounds on the number of conserved lncRNAs

Only a moderate fraction of ~3% of the splice sites of human lncRNAs are orthologous to known splice sites of annotated transcripts in other nonprimate Eutheria. This estimate is consistent with the observation that ~12% of the lincRNAs compiled in Cabili et al. (2011) are syntenically paired with a corresponding transcript in another mammalian species as detectable by `TransMap` (Zhu et al. 2007). Furthermore noncoding transcripts are typically expressed at lower levels than their coding counterparts and are often restricted to specific cell-lines or tissues (The ENCODE Project Consortium 2007).

Clearly, the poor sequence conservation of the lncRNAs (Marques and Ponting 2009) limits the number of human splice sites for which sequences from other eutherian families can be aligned. As a consequence, we can only determine a lower bound on the numbers of evolutionarily conserved splice sites in lncRNAs. The estimates therefore are limited by alignment coverage and quality. We refer to the Supplemental Material for a more detailed comparison of UCSC and ENSEMBL alignment.

The small fraction of conserved lncRNAs, however, is mainly the result of the incompleteness of the transcript catalogs in nonhuman species. We therefore use the conservation of splice sites as measured by `MaxEntScan` scores to obtain more accurate estimates. As detailed in the Materials and Methods, a cutoff of 3.0 for the `MaxEntScan` scores is sufficiently specific that we already tend to underestimate the number of conserved splice sites.

lncRNAs with many introns, such as *GAS5* in Figure 6 below, tend to enrich poorly conserved splice sites with only marginal support by low `MaxEntScan` scores. At least some of these are probably mapping artifacts that artefactually reduce the estimates of splice-site conservation. Since we consider an lncRNA as conserved if at least one splice site of the human transcript corresponds to a predicted or experimentally known splice site, the high-scoring splice sites are sufficient to establish the ancient origin of lncRNAs. The biases introduced by spurious and low-scoring splice sites in the GENCODE data thus have little impact on the results at transcript level. Furthermore, we observe no strong dependence of splice-site conservation on the number of exons, although the average splice-site score slightly increases with the number of exons; see Supplemental Table S5. The Cabili data set (Cabili et al. 2011) yields very similar results as the filtered GENCODE data; see Supplemental Figure S6.

The nearly constant conservation rate of ~30% suggests that there is a population of highly conserved splice sites in ancient lncRNAs. On the other hand, it also indicates that sequence conservation in the remaining ~70% of these highly conserved loci is unrelated to splicing and may not be conserved because of a function at the transcript level.

More than half of the GENCODE lncRNAs are conserved across the Eutheria

Table 2 summarizes our data for several mammalian species for which larger sets of transcriptome data are available.

These data indicate that >38% (6541/17,163) of the individual splice sites and 71% (3862/5413) of the transcripts are conserved across the major eutherian families. When we include 15 available nonprimate vertebrate genomes, this number increases further to 4511 transcripts (83%) and 53% of the splice sites. This reveals the massive gap to an estimation of only 3% (506/17,163) conservation of splice sites and 9% (462/5413) of transcripts, where only orthologs in annotated transcripts are considered as conserved.

Most recently, a subset of 1898 GENCODE lncRNAs expressed in a certain collection of human tissues was investigated for conserved expression in five other mammalian species (chimp, rhesus, cow, mouse, and rat) (Washietl et al. 2014). Expression from orthologous loci was observed for 35% (rat) to 80% (chimp) of the human transcripts. In these RNAs, conservation of between 20% and 60% of the observed human splice junctions were directly confirmed as conserved by dedicated transcriptome sequencing data.

TABLE 2. Conservation of GENCODE lncRNAs in the UCSC alignment

Species	Splices sites		Transcripts	
	17,163		5413	
Human	Cons.	Val.	Cons.	Val.
Mouse	2188	295	1910	308
Rat	2005	164	1777	185
Cow	3856	300	2845	268
Dog	4234	146	3053	146
Union 5	6541	515	3862	462
Union 15	9047	506	4511	462

The number of conserved and validated splice sites and transcripts in selected species gives an overview of conservation of human lncRNAs in vertebrates. A validated conserved splice site is defined as known splice site orthologous to the reference, whereas the category conserved includes in addition the predicted functional orthologs. *Union 5* refers to conservation in either mouse, rat, cow, or dog; *Union 15* refers to conservation in at least one of the following species: mouse, guinea pig, rabbit, cow, horse, dog, elephant, armadillo, opossum, chicken, frog, fugu, zebrafish, and lamprey.

This is in good agreement with the estimated conservation of mouse splice sites in Table 1. Our numbers, furthermore, are in agreement with the estimate that 60%–70% of the intergenic lncRNAs are conserved between human and mouse (Managadze et al. 2013). This estimate is based on the comparison of lncRNA expression from syntenically conserved loci, without regard to gene structure. Thus we do expect our estimate to appreciably more conservative.

A surprisingly large number of lncRNAs can be traced even further: 784 transcripts (14.5%) are conserved in at least one of the two marsupials (opossum, wallaby) and 446 can be found in the platypus genome.

Nearly 80% of the human lncRNAs may be older than the primates

At least a crude upper bound on the conservation of lncRNAs can be estimated by discarding all missing data and considering only the conservation of splice sites in those sequences that are present in the multiple sequence alignments. As expected, the estimates for individual species in Supplemental Table S3 are substantially larger than the conservative estimates of Table 2, which interprets all missing data as nonconservation (for GENCODE transcripts conserved in mouse, 50.7% compared with 35.3%). Surprisingly, the discrepancy, however, is rather small for the number of transcripts that are conserved in at least one of the four species: 79.6% versus 71.3% (see Fig. 4).

Most human lncRNAs are either primate-specific or they date back to the origin of the Eutheria

Figure 5 summarizes the gains and losses of human GENCODE lncRNAs. The primate subtree is left unresolved in this analysis because the evolutionary distances within this clade are too small to distinguish splice sites under stabilizing selection from fortuitous conservation due to insufficient divergence time. Only 6.3% (343/5413) of the transcripts are primate specific, while >54% (2905/5413) arose with the *Eutheria* and another 21% (1114/5413) can be traced back to the origins of the *Theria*.

Lineage-specific losses of lncRNAs are common

In contrast to the 71% of transcripts that are conserved between human and at least one of the other four eutherian species (see Table 2), there are few transcripts that are ubiquitously present. Rose et al. (2011) recently introduced a method to detect novel evolutionarily conserved splice sites and provided a collection of predicted splice sites that are well-conserved across the Eutheria. 2061 GENCODE lncRNAs have at least one splice site that is contained in this set of predictions. This fits well with only 814 transcripts that are conserved between human and "all" four eutherian species listed in Table 2. This suggests that lineage-specific losses are frequent.

Indeed, we miss 12.2% (660/5413) of the ancestral lncRNAs in mouse and >19% (1047/5413) in armadillo. These numbers have to be taken with caution, however. Our conservative cutoffs tend to over-emphasize losses and misplace origination events toward the tips of the tree.

Gene structures of conserved lncRNAs evolve rapidly

Conserved lncRNAs exhibit a rapid evolution of their gene structure. To estimate the turnover of individual splice sites in conserved transcripts we consider the 814 transcripts present in human, mouse, rat, cow, and dog. The human transcripts comprise 3080 splice sites. Of these, 87% were ancestrally present. Most of the novel splice sites were gained throughout primate evolution. Complementarily, a comparable number of donors and acceptors have been lost in Glires (Supplemental Fig. S7). In some examples the changes of transcript structure are quite dramatic. In the *ANRIL* isoforms, entire groups of exons are primate specific, while only a few splice sites, mostly located at the 5′ and the 3′ ends are at least as old as the Eutheria (see Fig. 7A below). The visibly higher conservation until marmoset, is consistent with the finding that *ANRIL* is first fully developed in simians, after it went through a two-stage evolution (He et al. 2013). Another famous example is *HOTAIR*, where the 5′-most exons appear to be lacking in the mouse (Schorderet and Duboule 2011).

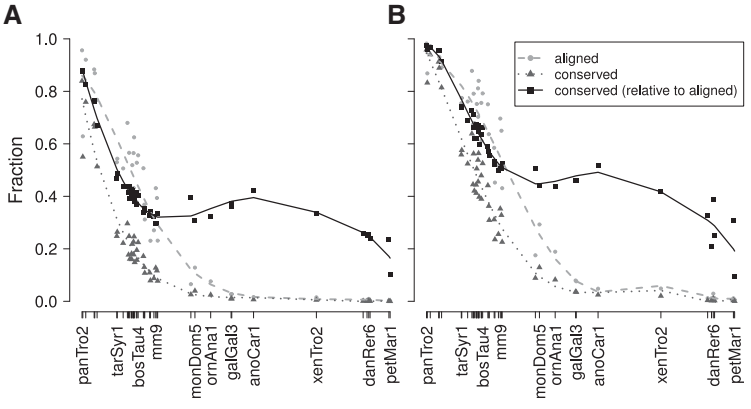

FIGURE 4. Conservation of lncRNAs across 46 vertebrates. Indicated in light gray is the fraction of aligned splice sites, in dark gray the fraction of splice sites that are validated and/or predicted to be a functional splice site in the regarding species. In black the upper bounds on the fraction of conserved splice sites are shown. The numbers are estimated from the fraction of conserved splice sites within aligned sequence blocks only. *A* shows the conservation rate of 17,163 single splice sites, while *B* illustrates the conservation on the level of transcripts for 5413 lncRNAs.

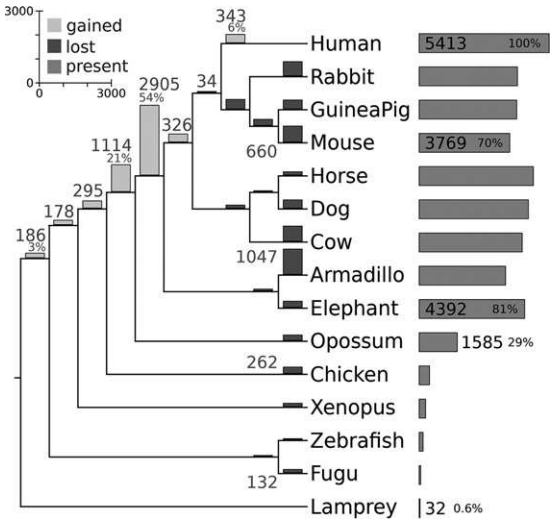

FIGURE 5. Gains and losses of human GENCODE lncRNAs across the vertebrates. Events are inferred by the parsimony criterion: A gene is deemed lost along the edge leading to a maximal subtree for which it is not observed at any leaf; a gain event is placed on the edge leading to the last common ancestor of all observed occurrences. The vertebrate phylogeny is the `phyloFit` tree provided by the UCSC browser. The primate subtree is omitted.

Alternative data sets lead to consistent results

Host genes of microRNAs and snoRNAs form subgroups with well-defined functions. Both groups of small structured RNAs are typically rather well conserved at least across the Eutheria. This is also true for their host genes, Table 3. There is little difference in the conservation of snoRNA and microRNA host genes, even though microRNAs can be processed from both exonic and intronic parts of a primary transcript (Kim 2005), while snoRNAs are obligatorily intronic at least in mammals (Maxwell and Fournier 1995). Interestingly, a much larger fraction of snoRNA host genes has experimentally validated conserved splice sites compared with microRNAs. This is probably due to their different expression patterns: MicroRNAs are often tissue or cell-type specific, while the snoRNAs are required ubiquitously.

The fractions of alignable positions and predicted splice sites among the mouse pluripotency lncRNAs (Guttman et al. 2011) is comparable with the GENCODE data. At the level of transcripts we again find substantial conservation across the Eutheria: More than half of the transcripts are predicted to be conserved in human, and for 40% of these experimental evidence is available.

For the zebrafish lncRNAs, a much lower conservation level of only 34% is observed among the other teleosts. The divergence of zebrafish and the Euteleostei is much older than the divergence of major eutherian groups (150 My versus 95 My from paleontological data (Benton and Donoghue 2007), or 230–333 My (Yamanoue et al. 2006) versus ~100–120 My (Hasegawa et al. 2003) estimated from molecular data). This readily explains the smaller fraction and the lower conservation of alignable splice sites. Interestingly, >11% of transcripts are conserved also in Tetrapoda.

Hundreds of lncRNAs are conserved throughout the vertebrates

Only three GENCODE splice sites in three lncRNAs show conservation to the lamprey, namely *AC011995.1-001*, *RP11-423H2.3-003*, and *RP11-123M21.1-001*. These are neither microRNA nor snoRNA host genes. We find 87 conserved transcripts (including one snoRNA host genes) in at least one of the teleosts. 26% of them are even experimentally validated. Two hundred seventy-one transcripts (including 14 microRNA, 10 snoRNA) are conserved in at least one of the Sauropsida. The deep conservation of host genes does not come as a surprise since in many cases their payload is conserved at least throughout the vertebrates (Hertel et al. 2006; Lestrade and Weber 2006; Sempere et al. 2006; Marz et al. 2011).

GAS5 is probably the best-studied snoRNA host gene, harboring ~10 distinct snoRNAs in its introns (Smith and Steitz 1998). It has recently attracted considerable attention since its in general poorly conserved exonic product acts as a riborepressor that binds to the DNA-binding domain of the glucocorticoidreceptor (Kino et al. 2010; Williams et al. 2011). Its chicken homolog is described in detail in Shao et al. (2009). Large clusters of ESTs are easily identified as *GAS5* homologs in frog (xenTro2 scaffold 1:6,870,168–6,878,818) and zebrafish (*ENSDARG00000092337*). The example of *GAS5* clearly shows the limitations of genome-wide alignments. Although *GAS5* is clearly conserved and functional (at least) across the gnathostomes (Fig. 6) the 46-way MULTIZ alignment does not contain the regions around the splice sites outside the Amniota; even in Sauropsida most parts are missing. Other well-studied examples of deeply conserved snoRNA host genes include *UHG* (*SNHG1*) (Fig. 7) and *U87HG* (Makarova and Kramerov 2009).

Not surprisingly, primary precursors of microRNAs are found among the best conserved lncRNAs. A well-studied case is *Rmst*, which harbors *mir-1251*. The human ortholog was described as differentially expressed in rhabdomyosarcoma subtypes (Chan et al. 2002). The mouse ortholog appeared as *Pax-2* related gene in early hind-brain development (Bouchard et al. 2005). Its evolution was investigated in detail in Chodroff et al. (2010), demonstrating conservation of both the transcript and its expression patterns in opossum and chicken brains. The comparative splice-site map shows that *Rmst* is conserved also in *Xenopus* (Fig. 7). The imprinted *MEG3* lncRNA exhibits a large number of differentially expressed isoforms (Zhang et al. 2010). It is an eutherian innovation apparently associated with the emergence of imprinting at the *Dlk1* locus (Weidman et al. 2006). Indeed, only a single splice site close to the 3′ end of the transcripts is shared with a putative evolutionary precursor in the marsupials (Fig. 7). It hosts the snoRNA *SNORD112* as well as the microRNA *mir-770*.

TABLE 3. Conservation of special subsets

	Aligned	Predicted	Validated
128 human transcripts hosting microRNAs			
Mouse	102	63	19
Dog	118	92	3
Five Eutheria	122	110	26
73 human transcripts hosting snoRNAs			
Mouse	56	49	35
Dog	66	59	20
Five Eutheria	69	63	41
2076 mouse lncRNAs[a]			
Human	1770	1113	446
Dog	1628	944	185
Four Eutheria	1776	1237	472
1508 zebrafish lncRNAs[b]			
Teleostei	953	513	112
Tetrapoda	476	170	56

We tabulate the number of conserved lncRNAs in selected species and in at least one of five Eutheria (human, mouse, rat, cow, dog), four Eutheria (mouse, human, cow, dog), Teleostei (tetraodon, stickleback), or Tetrapoda (human, mouse, frog). We decided to disregard rat for the mouse lncRNA subset calculations, as the two species are too closely related.
[a]Guttman et al. (2011).
[b]Ulitsky et al. (2011); Pauli et al. (2012).

The majority of the lncRNAs implicated in chromatin-based regulation can be traced throughout the Eutheria, although it is very likely that many of them are evolutionarily even older. A good example is *HOTTIP* (Fig. 7; Wang et al. 2011), where we lose the sequence conservation in most parts of the locus outside of the placental mammals. Although there are a few deeply conserved elements these do not include one of the splice-site sequences. Nevertheless, the transcript functions also in chick limb-buds (Wang et al. 2011), suggesting that the gene is considerably older than the Eutheria.

Two zebrafish lncRNAs that are conserved across vertebrates were investigated in detail (Ulitsky et al. 2011). *cyrano* (oip5 antisense transcript) is required for proper embryonic development. Our splice-site map identifies conservation of splice sites across mammals. The sequence is not conserved enough, however, to support an alignment between teleosts and tetrapods. *megamind* (located antisense in an intron of *birc6*) regulates brain morphogenesis and eye development. The last acceptor site is conserved across gnathostomes in the eight-way zebrafish centered alignment.

SpliceMap web service

The splice-site maps based on several multiple sequence alignments are available as a web service. It can be used to produce overview maps such as those in Figures 1 and 7 and to export text files of predicted and validate splice sites. Either a list of splice-site coordinates or a genomic interval can be used as input.

The website and the computation results are served by a set of Python scripts and rendered into static HTML using the Mako template engine. The jobs are scheduled in a queued fashion. Upon completion, the results are available under a personalized link for 2 wk. The service can be accessed at http://splicemap.bioinf.uni-leipzig.de.

DISCUSSION

The majority of the human long noncoding RNAs dates back at least to the radiation of the Eutheria, and thousands of these transcripts arose even earlier. The conservation of parts of their transcript structure constitutes compelling evidence for stabilizing selection, despite the often negligible constraints on the sequence itself. Utilizing the conservation of splice sites rather than measures of sequence similarity, furthermore, disentangles for a given locus the selective pressures on DNA elements from those that refer to the transcript. Our analysis, which suggests that some 70% of human lncRNAs date back to the eutherian ancestor is in agreement with an independent estimate of the conservation of lincRNAs conservation between man and mouse (Managadze et al. 2013) and with a direct comparison of lncRNA expression in six diverse mammals (Washietl et al. 2014).

Despite the conservation at transcript level we observed a surprising amount of turnover at the level of individual splice sites, again in agreement with Washietl et al. (2014). We observe that many of the lncRNA loci exhibit a large number of splicing isoforms. As a consequence of the lack of detailed transcriptomics data for most species, it is

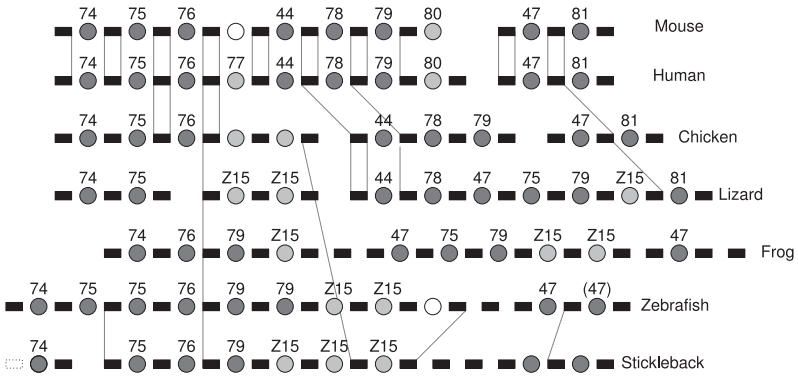

FIGURE 6. Conserved splice sites of the *GAS5* lncRNA. The *GAS5* snoRNA host gene is among the most highly conserved lncRNAs. Its homologs are easily identifiable via the well-conserved snoRNAs (circles) located within its introns. Members of the *SNORD80/Z15* family are shown in light gray. Black boxes indicate the major exons supported by RefSeq and/or EST data. Thin gray lines indicate splice sites that can be traced manually in at least one of the genome-wide alignments available in the UCSC browser. Note that only a subset of these is represented in any individual alignment (cf. Fig. 1). The transcript structure as well as its snoRNA payload has changed also by means of duplications and deletions.

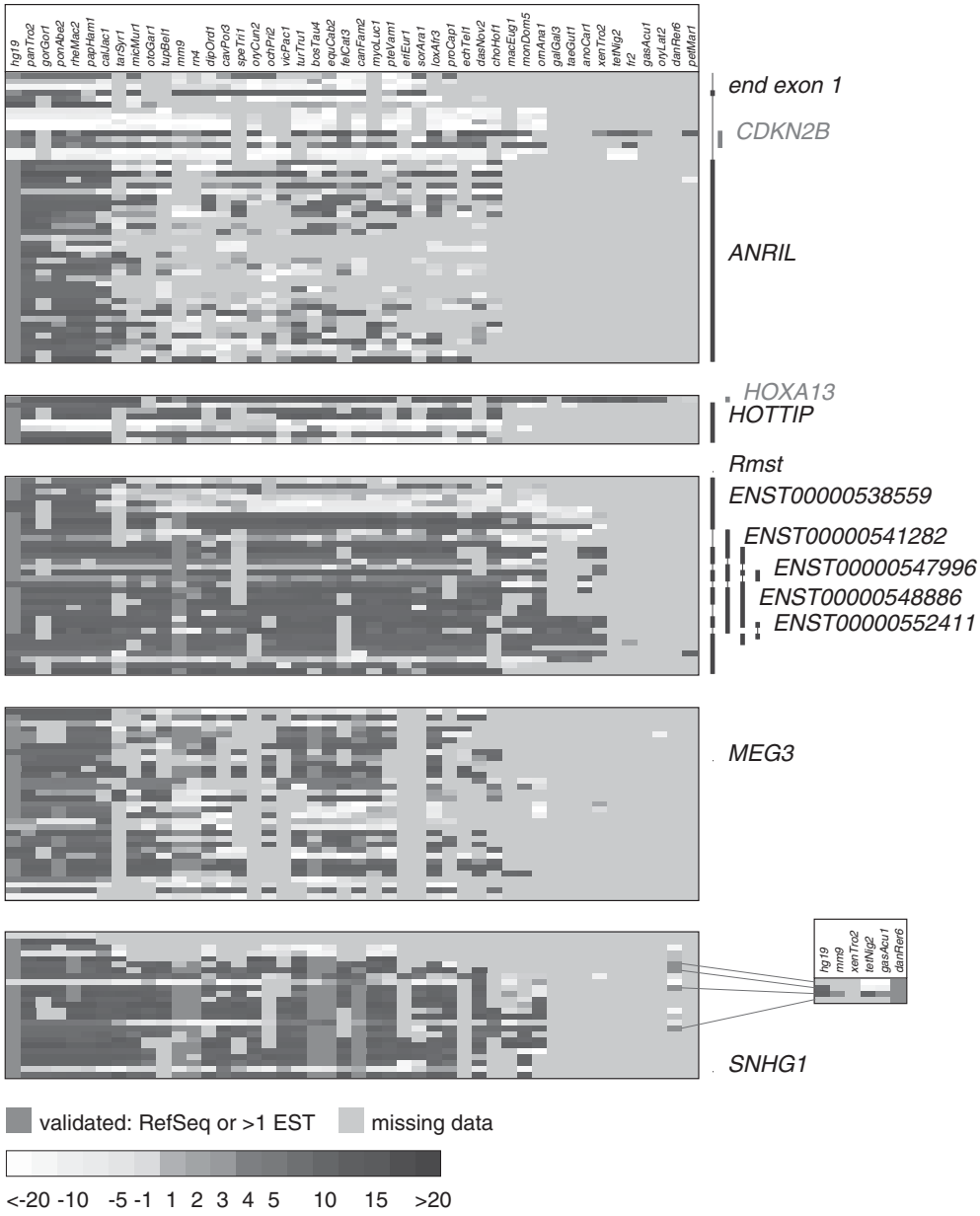

FIGURE 7. Variation of splice-site conservation. The patterns of splice-site conservation vary substantially between different lncRNAs, even when their evolutionary age is comparable. The main panel refers to the UCSC 46-way alignment. In the case of *ANRIL*, only a few splice sites are conserved outside the primates. Although the mouse ortholog shares at least some functions with human *ANRIL* (Pasmant et al. 2010), there are only four shared conserved splice sites. *HOTTIP*, with few exons that are partially conserved, is also a rather typical chromatin-related lincRNA. In contrast, the overwhelming majority of splice sites is conserved in *Rmst*. *MEG3* shows an intermediate pattern, with more lineage-specific losses. The snoRNA host gene *SNHG1* contains several splice sites that are deeply conserved among vertebrates. Some are even found in teleosts. Experimentally known splice sites from zebrafish *SNHG1* were searched also in the six-way zebrafish MULTIZ alignment (*inset*). Additional homologous splice sites in two teleosts demonstrate once more the limitations arising from alignment quality. The grayscales are explained in Figure 1. Thick vertical bars on the *right* mark splice sites that belong to a specific transcript (black: plus strand, gray: minus strand). Thin lines between these bars indicate conserved splice sites, which are not part of the annotated transcripts.

currently impossible to trace the evolution of individual isoforms. The discrepancies among individual splice sites, however, leads us to hypothesize that differential selection of isoforms caused the observed rapid divergence of transcript structures. Together with a prolific innovation of new splice sites this process can quickly obscure the evolutionary relationships. Our analysis may still drastically underestimate the evolutionary age of lncRNAs.

We suspect that, as in the case of *HOTAIR* or *ANRIL*, major changes of transcript structure go hand in hand with

functional changes. This view is supported by major differences between isoforms, e.g., in the association of their expression levels with disease phenotypes (Burd et al. 2010; Holdt et al. 2010, 2013) or the change of function of *HOTAIR* in mouse that correlates with the loss of several exons (Schorderet and Duboule 2011). If our hypothesis is true, lncRNAs are likely to be the root cause for rapid phenotypic evolution, as their often chromatin-associated mode of action is subject to large functional changes by easy-to-achieve changes in gene structure. The selective inclusion or exclusion of protein binding sites would affect the composition of complexes of enhancers and chromatin modifiers (see, e.g., Mercer et al. [2009]), and thus rapidly alter the rules of transcriptional regulation without affecting the proteins machinery. A similar scenario can be drawn for the post-transcriptional regulation of the pool of microRNA composition by sponges such as *HULC* (Wang et al. 2010). We conclude that lncRNAs are an ancient component of vertebrate genomes with an unexpected and unprecedented evolutionary plasticity.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was funded in part by the European Union FP-7 project QUANTOMICS (no. 222664). MMML-seq project of the International Cancer Genome Consortium (ICGC) is funded by German Federal Ministry of Education and Research. LIFE-Leipzig Research Center for Civilization Diseases is funded by the State of Saxony and the European Union.

Received May 12, 2014; accepted December 24, 2014.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Baldo L, Santos ME, Salzburger W. 2011. Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biol Evol* **3:** 443–455.

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol* **24:** 26–53.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit A, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14:** 708–715.

Bouchard M, Grote D, Craven SE, Sun Q, Steinlein P, Busslinger M. 2005. Identification of PAX2-regulated genes by expression profiling of the mid-hindbrain organizer region. *Development* **132:** 2633–2643.

Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Lercher MJ, et al. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiol* **155:** 142–156.

Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, Sharpless NE. 2010. Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet* **6:** e1001233.

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.

Chan AS, Thorner PS, Squire JA, Zielenska M. 2002. Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes. *Oncogene* **21:** 3029–3037.

Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnár Z, Ponting CP. 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* **11:** R72.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, Dörk T, Burge C, Gatti RA. 2004. Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *Hum Mutat* **23:** 67–76.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458:** 223–227.

Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, et al. 2011. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477:** 295–300.

Hackermüller J, Reiche K, Otto C, Hösler N, Blumert C, Brocke-Heidrich K, Böhlig L, Nitsche A, Kasack K, Ahnert P, et al. 2014. Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein-coding RNAs. *Genome Biol* **15:** R48.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen C, Chrast J, Lagarde J, Gilbert J, Storey R, Swarbreck D, et al. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1:** S4.1–S4.9.

Hasegawa M, Thorne JL, Kishino H. 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes Genet Syst* **78:** 267–283.

Hashimshony T, Yanai I. 2010. Revealing developmental networks by comparative transcriptomics. *Transcription* **1:** 154–158.

He S, Gu W, Li Y, Zhu H. 2013. ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC Evol Biol* **13:** 247.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, The Students of Bioinformatics Computer Labs 2004 and 2005. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7:** 25.

Hiller M, Findeiss S, Lein S, Marz M, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, et al. 2009. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res* **19:** 1289–1300.

Holdt LM, Beutner F, Scholz M, Gielen S, Gäbel G, Bergert H, Schuler G, Thiery J, Teupser D. 2010. ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. *Arterioscler Thromb Vasc Biol* **30:** 620–627.

Holdt LM, Hoffmann S, Sass K, Langenberger D, Scholz M, Krohn K, Finstermeier K, Stahringer A, Wilfert W, Beutner F, et al. 2013. Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet* **9:** e1003588.

Hong X, Scofield DG, Lynch M. 2006. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* **23:** 2392–2404.

Kapranov P, Willingham AT, Gingeras TR. 2007. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genetics* **8:** 413–423.

Kapranov P, St Laurent G, Raz T, Ozsolak F, Reynolds P, Sorensen PHB, Reaman G, Milos P, Arceci RJ, Thompson JF, et al. 2010. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol* **8:** 149.

Kim VN. 2005. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* **6:** 376–385.

Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. 2010. Noncoding RNA GAS5 is a growth arrest- and starvation-associated repressor of the glucocorticoidreceptor. *Sci Signal* **3:** ra8.

Lee AP, Koh EG, Tay A, Brenner S, Venkatesh B. 2006. Highly conserved syntenic blocks at the vertebrate HOX loci and conserved regulatory elements within and outside HOX gene clusters. *Proc Natl Acad Sci* **103:** 6994–6999.

Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* **34:** D158–D162.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27:** i275–i282.

Louro R, Smirnova AS, Verjovski-Almeida S. 2009. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* **93:** 291–298.

Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engström PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, et al. 2006. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* **2:** e62.

Mainguy G, Koster J, Woltering J, Jansen H, Durston A. 2007. Extensive polycistronism and antisense transcription in the mammalian *Hox* clusters. *PLoS One* **2:** e356.

Makarova JA, Kramerov DA. 2009. Analysis of C/D box snoRNA genes in vertebrates: the number of copies decreases in placental mammals. *Genomics* **94:** 11–19.

Managadze D, Lobkovsky AE, Wolf YI, Shabalina SA, Rogozin IB, Koonin EV. 2013. The vast, conserved mammalian lincRNome. *PLoS Comput Biol* **9:** e1002917.

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

Marz M, Gruber AR, Höner zu Siederdissen C, Amman F, Badelt S, Bartschat S, Bernhart SH, Beyer W, Kehr S, Lorenz R, et al. 2011. Animal snoRNAs and scaRNAs with exceptional structures. *RNA Biol* **8:** 938–946.

Maxwell ES, Fournier MJ. 1995. The small nucleolar RNAs. *Annu Rev Biochem* **64:** 897–934.

Mercer TR, Dinger ME, Mattick JS. 2009. Long noncoding RNAs: insights into function. *Nat Rev Genet* **10:** 155–159.

Mercer TR, Wilhelm D, Dinger ME, Soldà G, Korbie DJ, Glazov EA, Truong V, Schwenke M, Simons C, Matthaei KI, et al. 2011. Expression of distinct RNAs from 3′ untranslated regions. *Nucleic Acids Res* **39:** 2393–2403.

Natale A, Sims C, Chiusano ML, Amoroso A, D'Aniello E, Fucci L, Krumlauf R, Branno M, Locascio A. 2011. Evolution of anterior HOX regulatory elements among chordates. *BMC Evol Biol* **11:** 330.

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.

Pang KC, Frith MC, Mattick JS. 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22:** 1–5.

Pasmant E, Sabbagh A, Vidaud M, Bièche I. 2010. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB J* **25:** 444–448.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18:** 1814–1828.

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 non-coding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22:** 577–591.

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

Punnamoottil B, Herrmann C, Pascual-Anaya J, D'Aniello S, Garcia-Fernàndez J, Akalin A, Becker TS, Rinkwitz S. 2010. *Cis*-regulatory characterization of sequence conservation surrounding the HOX4 genes. *Dev Biol* **340:** 269–282.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129:** 1311–1323.

Rose D, Hiller M, Schutt K, Hackermüller J, Backofen R, Stadler PF. 2011. Computational discovery of human coding and noncoding transcripts with conserved splice sites. *Bioinformatics* **27:** 1894–1900.

Schorderet P, Duboule D. 2011. Structural and functional differences in the long non-coding RNA *hotair* in mouse and human. *PLoS Genet* **7:** e1002071.

Scofield DG, Hong X, Lynch M. 2007. Position of the final intron in full-length transcripts: determined by NMD? *Mol Biol Evol* **24:** 896–899.

Sempere LF, Cole CN, McPeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol* **306:** 575–588.

Shao P, Yang JH, Zhou H, Guan DG, Qu LH. 2009. Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics* **10:** 86.

Smith CM, Steitz JA. 1998. Classification of GAS5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5′terminaloligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol* **18:** 6897–6909.

Suzuki M, Hayashizaki Y. 2004. Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *Bioessays* **26:** 833–843.

Tanzer A, Amemiya CT, Kim CB, Stadler PF. 2005. Evolution of microRNAs located within *Hox* gene clusters. *J Exp Zool B Mol Dev Evol* **304:** 75–85.

Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329:** 689–693.

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.

Wang J, Liu X, Wu H, Ni P, Gu Z, Qiao Y, Chen N, Sun F, Fan Q. 2010. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* **38:** 5366–5383.

Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120–124.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci* **102:** 2454–2459.

Washietl S, Findeiß S, Müller S, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: robust prediction of protein coding regions in comparative genomics data. *RNA* **17:** 578–594.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24:** 616–628.

Weidman JR, Maloney KA, Jirtle RL. 2006. Comparative phylogenetic analysis reveals multiple non-imprinted isoforms of opossum Dlk1. *Mamm Genome* **17:** 157–167.

Williams GT, Mourtada-Maarabouni M, Farzaneh F. 2011. A critical role for non-coding RNA GAS5 in growth arrest and rapamycin inhibition in human T-lymphocytes. *Biochem Soc Trans* **39:** 482–486.

Yamanoue Y, Miya M, Inoue JG, Matsuura K, Nishida M. 2006. The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. *Genes Genet Syst* **81:** 29–39.

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

Zhang X, Rice K, Wang Y, Chen W, Zhong Y, Nakayama Y, Zhou Y, Klibanski A. 2010. Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: isoform structure, expression, and functions. *Endocrinology* **151:** 939–947.

Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, Haussler D. 2007. Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* **3:** e247.