Cell Reports

Supplemental Information

Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species

Hadas Hezroni, David Koppstein, Matthew G. Schwartz, Alexandra Avrutin, David P. Bartel, and Igor Ulitsky

Supplemental Information for "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species" by Hezroni et al.

Supplemental Experimental Procedures

RNA-Seq and 3P-seq data sources

RNA-seq reads were obtained from the NCBI short read archive (SRA), with the following accessions, unless indicated otherwise (see Table S1 for read statistics): Human (ERP003613); Rhesus (SRP009818, SRP016501 and SRP017517); Marmoset (obtained from NHPRTR (Pipes et al., 2013)); Mouse (Mouse ENCODE project CSHL Long RNA-Seq data); Dog (SRP009687); Ferret (SRP009667, obtained from Ensembl); Opossum (SRP023152); Chicken (ERP003988 and SRP016501); Lizard (SRP009831); Stickleback (SRP012923); Nile tilapia (SRP009911); Zebrafish (SRP009426, SRP024369, ERP000447, and ERP000016); Spotted Gar (SRP042013, obtained from Ensembl); Elephant Shark (SRP013772); and Sea urchin (SRP014690). Zebrafish 3P-seq data were previously described (Ulitsky et al., 2012) and are available at NCBI GEO accession GSE37453.

Annotation data

Genomic sequences with and without repeat masking (see Table S1 for assembly versions) and annotations of genomic sequence gaps were obtained the UCSC genome browser. Protein-coding and small RNA gene annotations were obtained from Ensembl and RefSeq databases and supplemented with RefSeq transcripts and human protein sequences from other species mapped to the respective genome in the UCSC genome browser ("Other RefSeq" and "Human proteins" tracks from the UCSC genome browser, if available). For small RNA annotations, we only considered the "Other RefSeq" entries for which the length of the original transcripts ("xenoRefSeqAli.qSize") was <200 nt. In order to address inconsistencies between Ensembl annotations and our criteria for protein-coding potential, we excluded Ensembl protein-coding genes that did not belong any known Ensembl family and had a predicted coding sequence length of <150 codons. For sea urchin, Ensembl proteins shorter than 80 aa were not considered. Transposable element annotations were taken from the "RepeatMasker" track in the UCSC genome browser, and only non-simple repeats were considered.

Comparisons of lncRNAs and protein-coding genes

One of the challenges in comparing transcriptomes across vertebrate species was the inconsistency in genome assembly and varying completeness of existing protein-coding gene annotations in different species. To address differences in annotation quality, the baseline collection of protein-coding genes that we used was restricted to those proteins annotated in Ensembl that had confident support. To accommodate differences in sequence-assembly quality and RNA-seq depth, when comparing lincRNAs to protein-coding genes, we used the reconstructed full-length mRNA models produced by our pipeline, thereby controlling for the quality of transcript models when comparing lincRNA and protein-coding genes.

PLAR – Pipeline for lncRNA annotation from RNA-seq data

1. Transcriptome assembly

The first step of PLAR uses TopHat2 (Trapnell et al., 2009) to map the RNA-seq reads to the genome in two iterations. In the first, we provided TopHat2 with spliced junction from Ensembl and RefSeq gene annotations (where available) and spliced ESTs and mRNAs obtained from the UCSC genome browser. In the second iteration, we used these junctions together with the splice junctions discovered in the first iteration and combined across all samples. Alignments were used as input for Cufflinks (Trapnell et al., 2010) with default parameters for assembly of transcript models in each sample. The Cufflinks transcript models from all the samples in each species were then merged using CuffMerge with the Ensembl annotations as a reference. Expression levels in each sample in Fragments Per Kilobase per Million reads (FPKM) units where quantified using CuffDiff (Trapnell et al., 2012). All the programs were used with default parameters.

2. Initial filtering

Multi-exon transcript models that had exonic sequences of less than 200 bases, or that were expressed at FPKM<0.1 in all samples were removed. Following analysis of 3P-seq data (**Figure S1A**) we implemented more stringent criteria for single-exon transcripts, and only those single-exon Cufflinks models with exonic length >2,000 nt and an FPKM>5 in at least one sample were retained. Transcripts that 50% of their exonic sequences annotated a single repeat were also removed. In chicken and zebrafish, where 3P-seq data were available, transcripts that did not overlap or end within 200 nt of a 3P-seq-defined polyadenylation site with at least three 3P-seq reads were removed.

3. Identifying transcripts overlapping with annotated genes and small RNA primary transcripts and hosts

A transcript that overlapped the coding sequence of a protein-coding gene by at least one base and overlapped any of its exonic sequence by at least 100 nt was designated as protein-coding. A single-exon transcript contained within an intron of a protein-coding gene on the same strand was annotated as “intron contained”. Transcripts overlapping on the other strand at least one base of the coding sequence of a protein-coding gene were considered as “antisense” transcripts and those overlapping on the same strand a small RNA gene where considered “small RNA primary transcripts”.

4. Filtering protein-coding potential

In order to identify potential protein-coding transcripts among lncRNA candidates, we used three methods for discovering protein-coding potential:

CPC (Kong et al., 2007) was applied to repeat-masked transcript sequences, using the RefSeq database of protein sequences (only “NM_” entries) as database of protein-coding genes.

HMMER (Eddy, 1998) was applied to repeat-masked transcripts translated in all three possible frames using the Hidden Markov Models (HMMs) of protein domains from the Pfam-A and Pfam-B databases (Finn et al., 2014). Any transcript with a Pfam domain prediction with E<0.001 was considered coding.

RNAcode (Washietl et al., 2011) was applied in species where whole genome alignments were available (**Table S2**). Transcripts were designated as coding if their exons overlapped a predicted protein-coding element with $P<10^{-4}$ by at least 10 bases. This filter was not used for antisense lncRNAs.

Any gene that contained an isoform reported as coding by one of those programs was designated as "predicted coding". As expected, the number of such protein-coding genes not yet annotated in Ensembl was small in extensively annotated genomes such as mouse or zebrafish (614 and 945 predicted coding genes, respectively), and larger in poorly annotated ones (3,402 in stickleback, which has 19,318 annotated protein-coding genes compared to 22,507 in zebrafish).

5. Additional filters

Transcripts proximal (within 500 nt for multi-exon or 2 Kb for single-exon) to an annotated or reconstructed protein-coding gene on the same strand were excluded (based on the observations in **Figure S1A**). In species were pseudogene annotations were available in the http://pseudogene.org resource (human, mouse, chicken, dog and zebrafish), we removed transcripts that overlapped any annotated pseudogenes. We removed transcripts that started or ended within 500 nt of a protein-coding gene (2,000 nt for single-exon transcripts). We also removed single-exon transcripts that overlapped a multi-exon transcript by at least 50% of their exonic sequence.

6. Filtering using *de novo* transcript assembly by Trinity

In species with relatively poor genome assemblies (**Table S2**), Trinity (Grabherr et al., 2011) was used with default parameters to reconstruct transcripts *de novo*, without a reference genome. Only Trinity transcripts that matched at least 200 nt of genomic sequence, and overlapped the Cufflinks transcripts model by ≥70 nt and ≥50% of the length of the Cufflinks model exons were considered. A Trinity transcript was considered to be "anchored" in a Cufflinks transcript if one of its termini appeared within 20 nt of an exon of that transcript. We then excluded lncRNA candidates if:

1. A Trinity transcript overlapped at least 50% of the exonic bases of the lncRNA candidate and had both ends anchored in it, but the Trinity transcript contained a prefix or a suffix with at least 100 nt that was not aligned to the reference genome sequence in the same locus.

2. The Trinity transcript had one end anchored in a lncRNA and another anchored in an annotated or reconstructed protein-coding gene.

Only transcripts that passed both filters were annotated as lncRNAs (see **Table S2** for numbers).

Identifying clusters of orthologous lincRNAs

1. Identifying sequence-similar lncRNAs

Two methods were used for comparison of lincRNA sequences. Repeat-masked exonic sequences of lncRNAs were compared using BLASTN from the BLAST+ suite version 2.2.28+, with parameters "-task blastn -word_size 8 -strand plus"). Transcripts with alignments with E-value $<10^{-5}$ were considered as sequence-similar. In addition, whole genome alignments from human, marmoset, mouse, dog, opossum, chicken (galGal3 assembly), zebrafish, stickleback where obtained from the UCSC genome browser and used to project sequences from the respective genomes to other

genomes from our species collection that were part of the specific genome alignment. Transcripts were considered sequence-similar if the projection of any exon of the first transcript overlapped the exon of the other.

2. Putative clusters of orthologous transcripts

In order to identify clusters of orthologous lincRNAs we first constructed a graph where every lincRNA transcript was a node and edges connected transcripts from the same species that had overlapping exons, and sequence-similar transcripts from different species. We then used breadth-first search (Cormen et al., 2009) to identify connected components in this graph and each connected component was considered as a putative cluster of orthologous lincRNAs.

3. Filtering putative clusters using synteny

In order to avoid spurious sequence similarities, we focused only on putative clusters where at least one pair of transcripts was supported by synteny. Annotations of protein-coding gene orthologs were obtained from Ensembl Compara (Vilella et al., 2009). When comparing two genomes A and B, and when considering orthologous protein-coding genes G_1 and G_2 we first identified lincRNAs within $5\times10^5\times\sqrt{(\text{GenomeLength}(A)/10^9)}$ nt of G_1 in A and within $5\times10^5\times\sqrt{(\text{GenomeLength}(B)/10^9)}$ nt of G_2 in B. An lncRNA was considered to be found "upstream" of the protein-coding gene when it overlapped it or ended 5' to its 5' end, and "downstream" when it overlapped it or started 3' to the 3' end of the protein-coding gene. Two lincRNAs L_1 and L_2 from A and B were considered syntenic, if they were both upstream or both downstream of G_1 and G_2, with the same relative orientations. A putative cluster was carried forward only if it contained at least two syntenic lncRNAs from different species.

Identifying stringently syntenic lincRNAs between human and other species

Stringently syntenic lincRNAs (**Figure 6**) between human and other species (all species considered except coelacanth, tilapia, elephant shark, and sea urchin where pairwise alignments with the human genome were not available) were syntenic lincRNAs identified as described above that also passed an additional filtering step. The basic principle of this filtering was that if L_1 and L_2 were syntenic based on their adjacency to the homologous G_1 and G_2 from their respective genomes then it is unlikely that there is a region R_1 in species A that maps to a region R_2 in species B such that R_1 is found in the proximity of G_1 and upstream of L_1 and R_2 is found in the proximity of G_2 but downstream of L_2, or vice versa – R_1 is downstream of L_1 whereas R_2 is upstream of L_2 (**Figure 6A**). We used pairwise alignments of the human genome with other genomes to systematically look for such pairs (R_1,R_2), and if such a pair was found we excluded the pair (L_1, L_2) it from consideration as syntenic transcripts.

Identifying significantly enriched *k*-mers

In each species and for each lncRNA gene, we merged overlapping lincRNA exons, retained only those that were shorted than 1,000 nt, and concatenated those into one sequence per gene. Ten random sequences of the same length and with the same dinucleotide distribution were generated using random shuffling. The number of occurrences of each 6mer was tallied in each of the sequences and in the controls. The significance of the number of appearances of each 6mer was evaluated by using a

Wilcoxon rank sum test comparing the vector of numbers of appearances of the 6mer in the lncRNA sequences with the average counts in the control sequences. P-values were Bonferroni corrected for multiple hypotheses testing.

In order to extract nonredundant *k*-mers, after processing the sequences of all species, for each 6mer, the average rank in a list of 6mers sorted by their Wilcoxon P-values was computed. Nonredundant 6mers were extracted by traversing the sorted list from the 6mer with the highest average rank to the lowest and retaining only 6mers that differed at least two mismatches from any 6mer that is a circular permutation of a higher-ranking 6mer. Specifically, consider a list of 6mers sorted by their average rank: $S_1,\ldots,S_n$. A sequence S_i was considered a nonredundant 6mer if at least two mismatches existed between S_i and any substring of S_j+S_j for any $j<i$.

Supplemental Figures

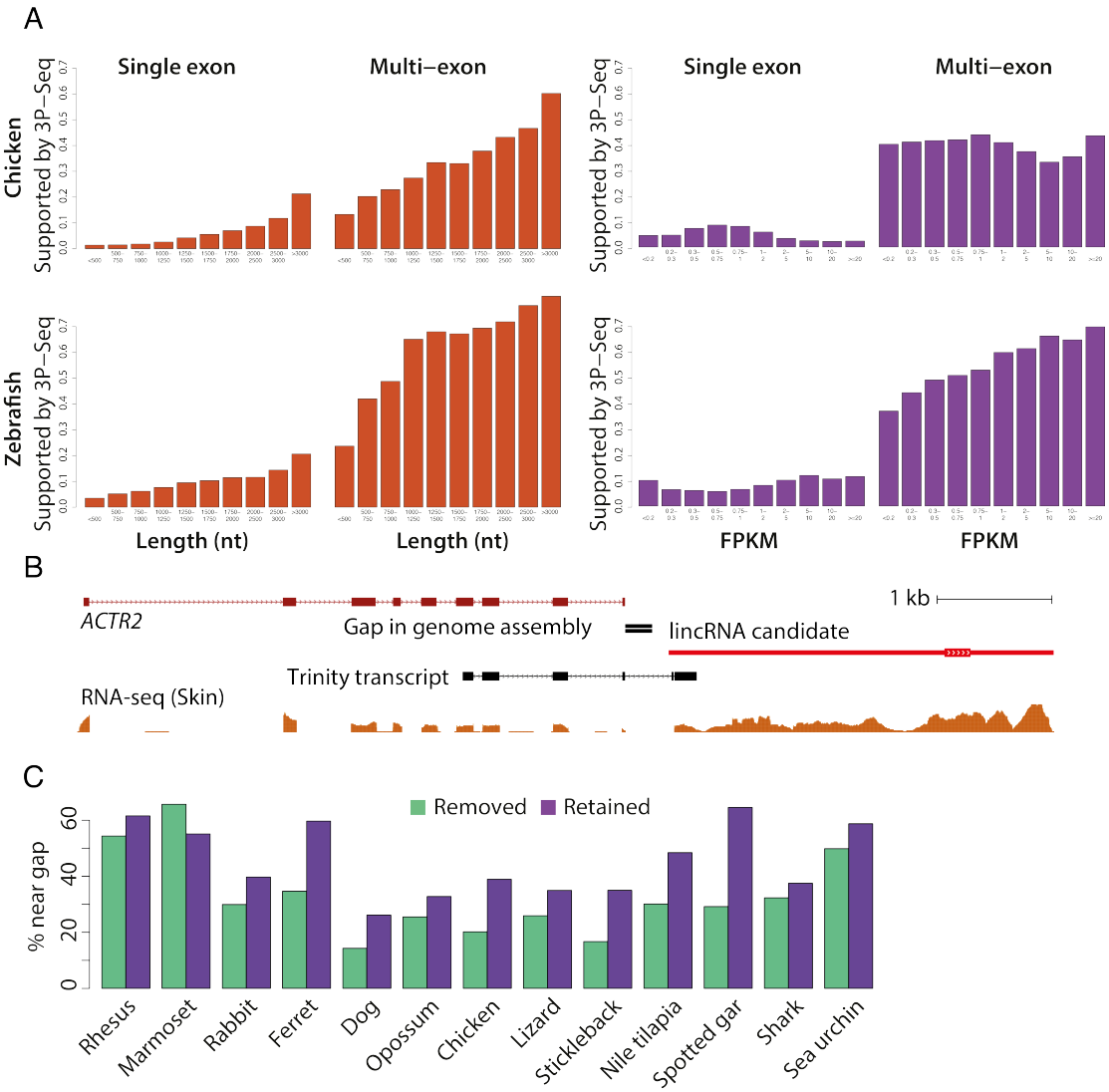

Figure S1. PLAR methodology, related to Figure 1. (A) Fraction of transcripts identified using Cufflinks/Cuffmerge pipeline that are supported by a cluster of 3P-seq reads, as a function of the exonic length (left) or maximal expression level across all samples (right). **(B)** Example of a lincRNA candidate in stickleback removed using the Trinity filter. *ACTR2* annotation taken from Ensembl, Trinity reconstructed transcript was mapped to the genome using BLAT. **(C)** Fraction of transcripts appearing within 200 nt of a gap in the genome assembly, shown separately for those transcripts that were removed by the Trinity filter or retained after it was applied.

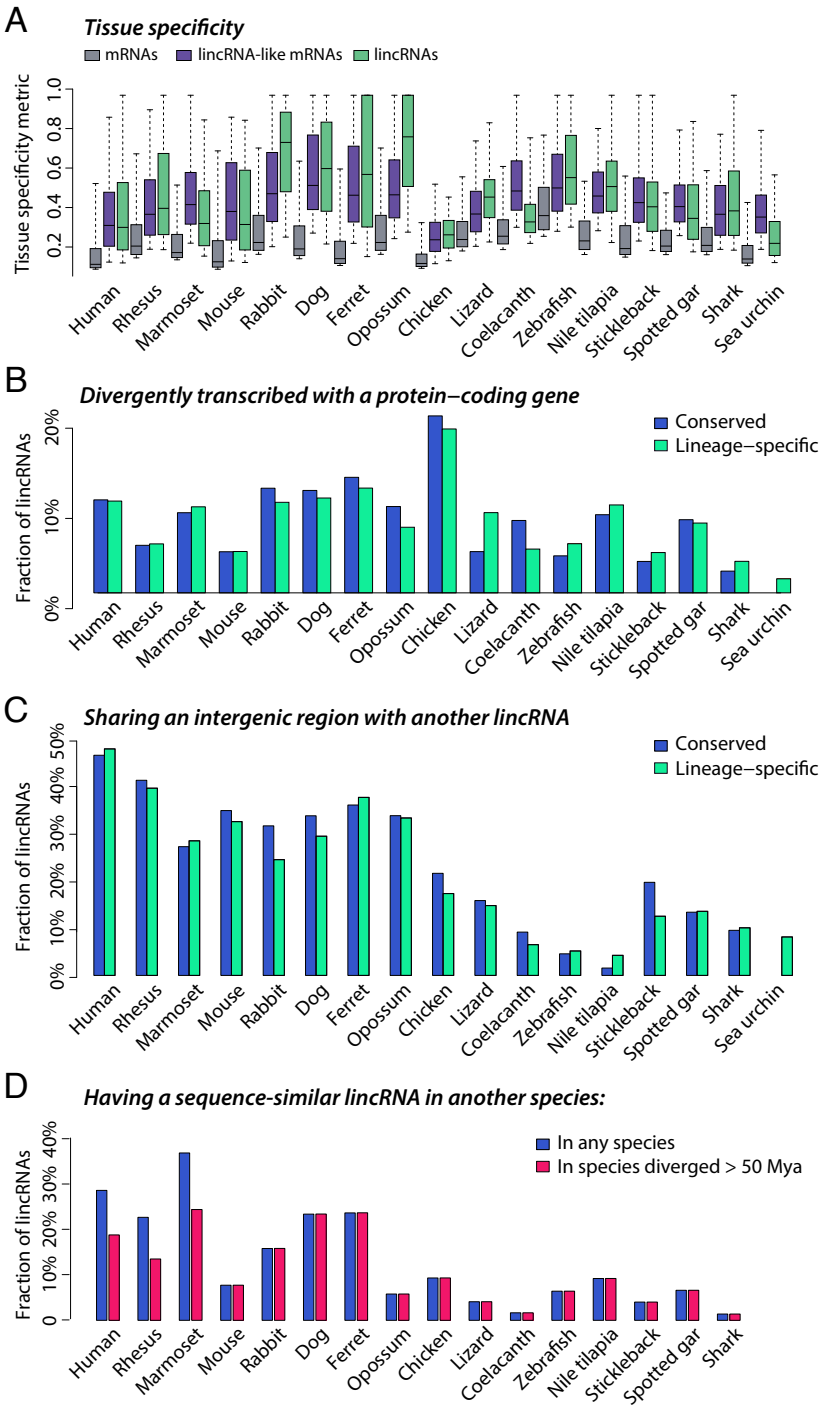

Figure S2. Features of lincRNAs in different species, related to Figure 1. (A) Tissue specificity of lincRNAs compared to protein-coding genes. Gene expression levels were estimated using CuffDiff (Trapnell et al., 2010). Tissue specificity computed as in Cabili et al. (2011). lincRNA-like mRNAs are a subset of mRNAs with average expression levels similar to that of lincRNAs. Specifically, the average expression levels of the mRNAs and lincRNAs were compiled and split into 10 equal-size bins. Then, mRNAs were sampled such that the fraction of mRNAs coming from each bin was the same as the fraction of lincRNAs. Plots indicate the median, quartiles, and 5th and 95th percentiles. **(B)** Fraction of lincRNAs in each species that are divergent with a protein-coding gene, defined as those lincRNAs that have their 5' end within 1 Kb of the 5' end of a protein-coding gene (as reconstructed by PLAR) on

the other strand. **(C)** Fraction of lincRNAs in each species that share the intergenic region in which they are located with another lincRNA (intergenic regions were defined by pairs of adjacent protein-coding genes). **(D)** Fraction of lincRNAs in each species that are conserved in another species from the 16 species we studied, or in one of the species diverged >50 million years ago (Mya).

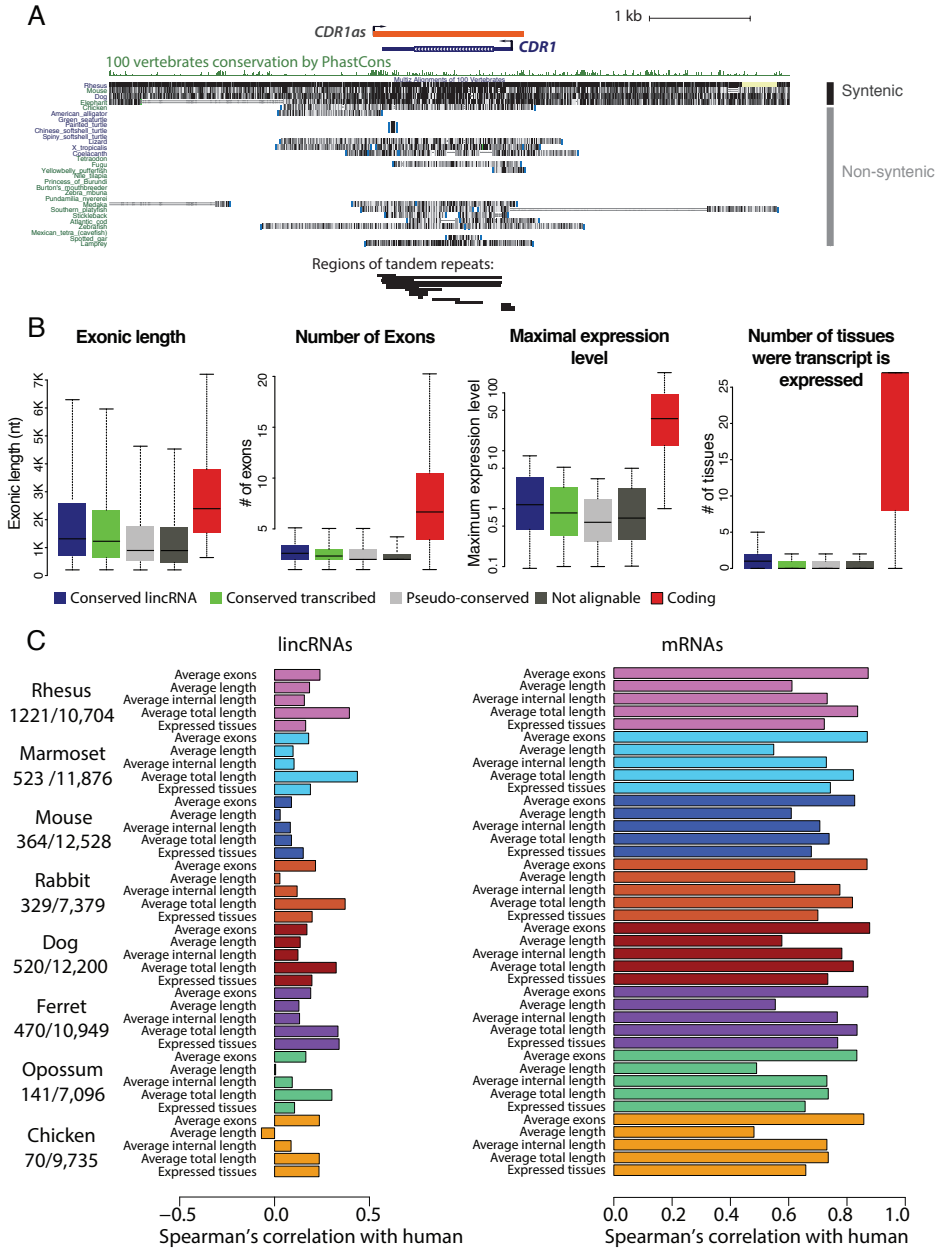

Figure S3. Features of conserved lincRNAs, related to Figure 2. (A) The genomic locus of the CDR1as transcript. Whole genome alignments were taken from the UCSC genome browser 100-way vertebrate genome alignment and the genomes in which the aligning regions are syntenic to the human CDR1as are indicated. Regions of tandem repeats taken from the "Simple repeats" track in the UCSC browser. **(B)** Comparison of features of human lincRNAs with varying levels of conservation. Conserved lincRNAs are those that have a sequence-similar homolog in another species. "Conserved transcribed" are lincRNAs alignable to a transcribed region that is not annotated as a lincRNA in the other species. Pseudo-conserved lincRNAs are

alignable to the genomes of other species, but the corresponding regions have no evidence of transcription. “Coding” refers to protein-coding genes reconstructed by our approach. Plots indicate the median, quartiles, and 5th and 95th percentiles. **(C)** Spearman’s correlations between the indicated genomic features of human lincRNAs and protein-coding genes and those of the indicated species. “Internal length” is the total length of mature transcript without the first and last exon. Numbers of the left indicate the number of lincRNA and protein-coding genes, respectively, used in the comparison.

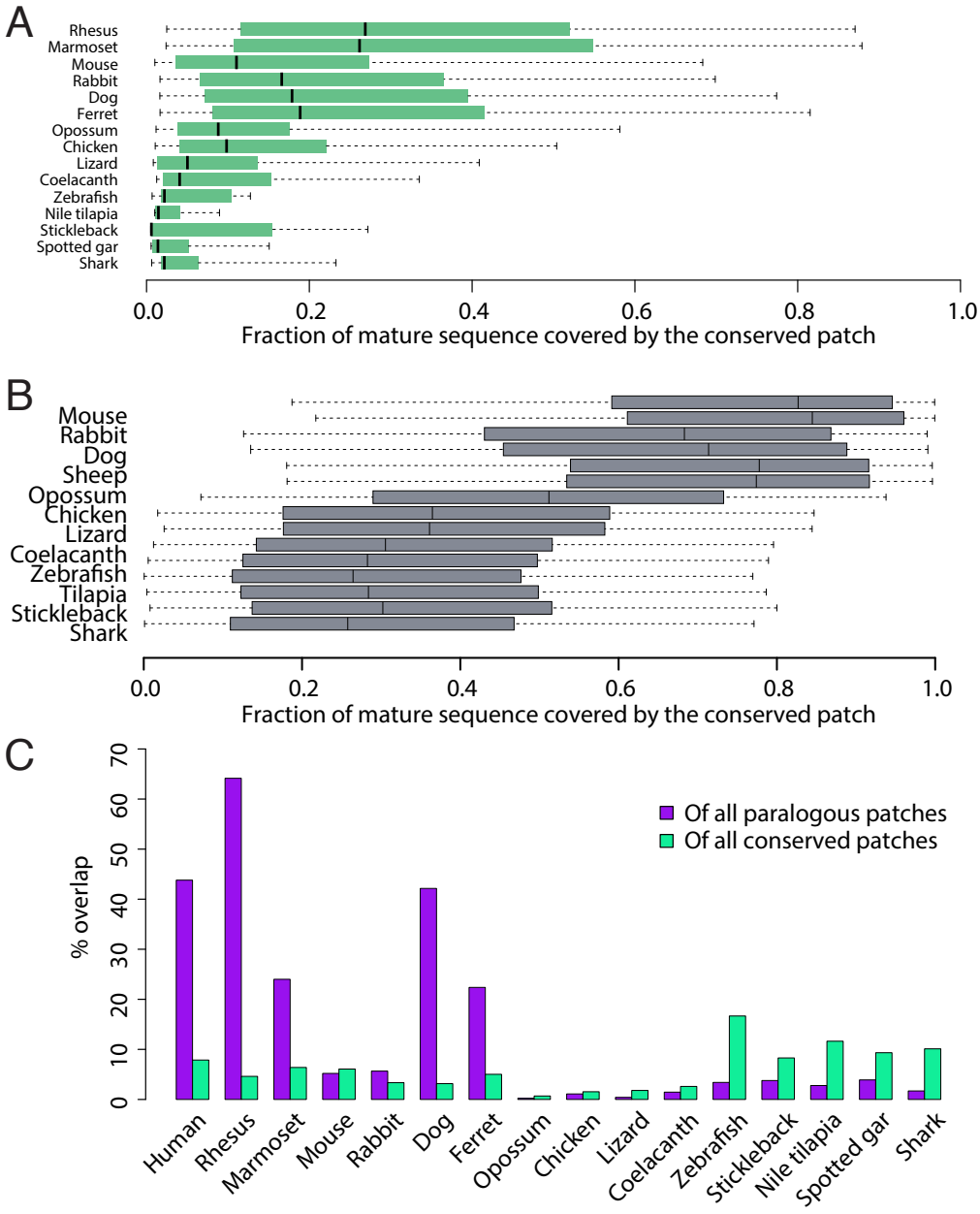

Figure S4. Features of conserved and paralogous patches, related to Figure 3. (A) Fraction of the human mature transcript sequence covered by sequence patches conserved in the indicated genome. Plots indicate the median, quartiles, and 5th and

95th percentiles. **(B)** Same as A, but for reconstructed protein-coding sequences. **(C)** Fractions of paralogous sequence similarity patches that overlapped a conserved sequence patch and vice versa.

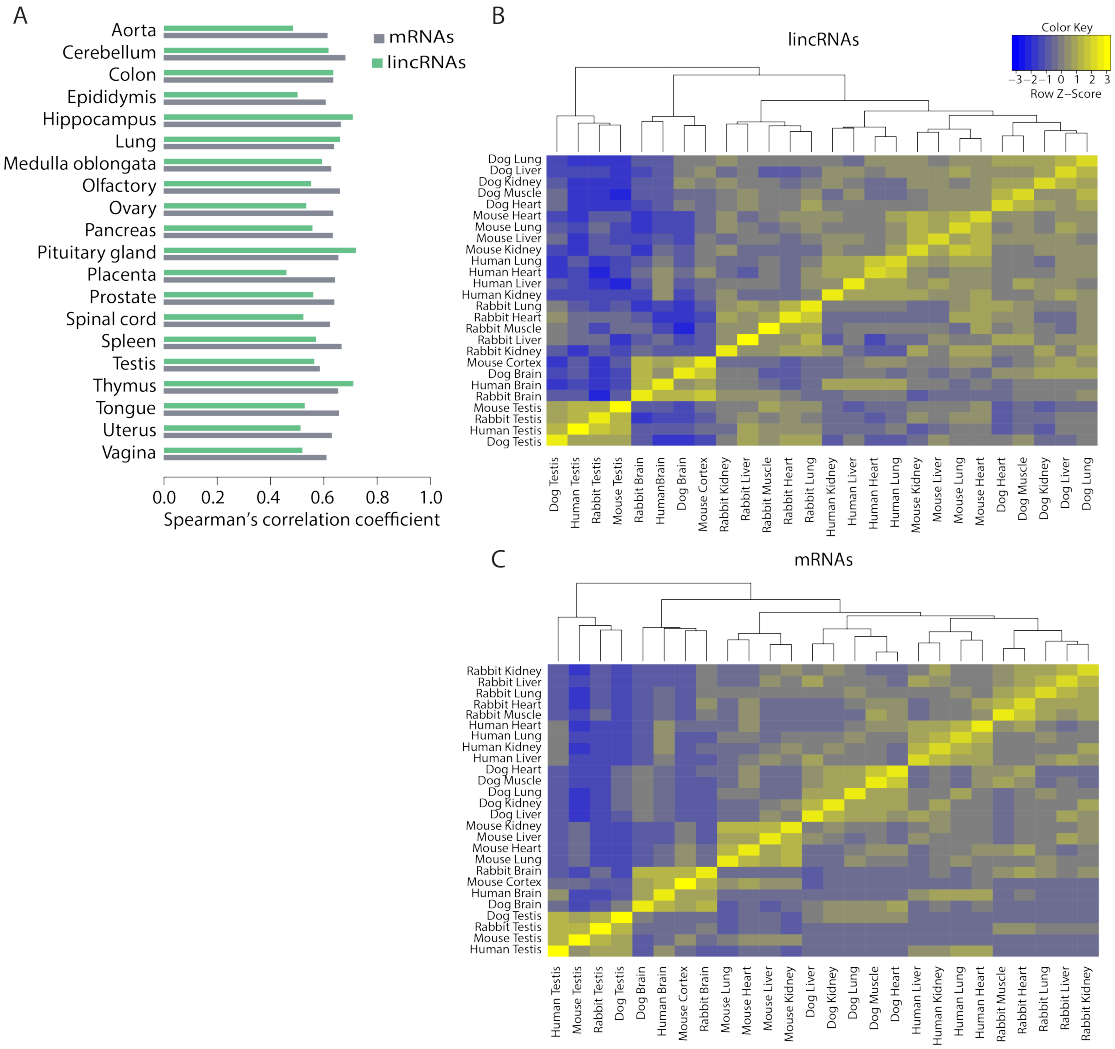

Figure S5. Conservation of lincRNA gene expression, related to Figure 4. (A) Correlation of absolute expression levels between human lincRNAs and mRNAs and their conserved homologs in indicated other species, based on gene expression estimates derived from CAGE data. The expression level of each gene is the sum of the transcripts per million (TPM) metrics taken from FANTOM5 data (Consortium et al., 2014). **(B-C)** Hierarchical clustering of RNA-seq–derived gene expression patterns using lincRNA (B) and mRNA (C) expression.

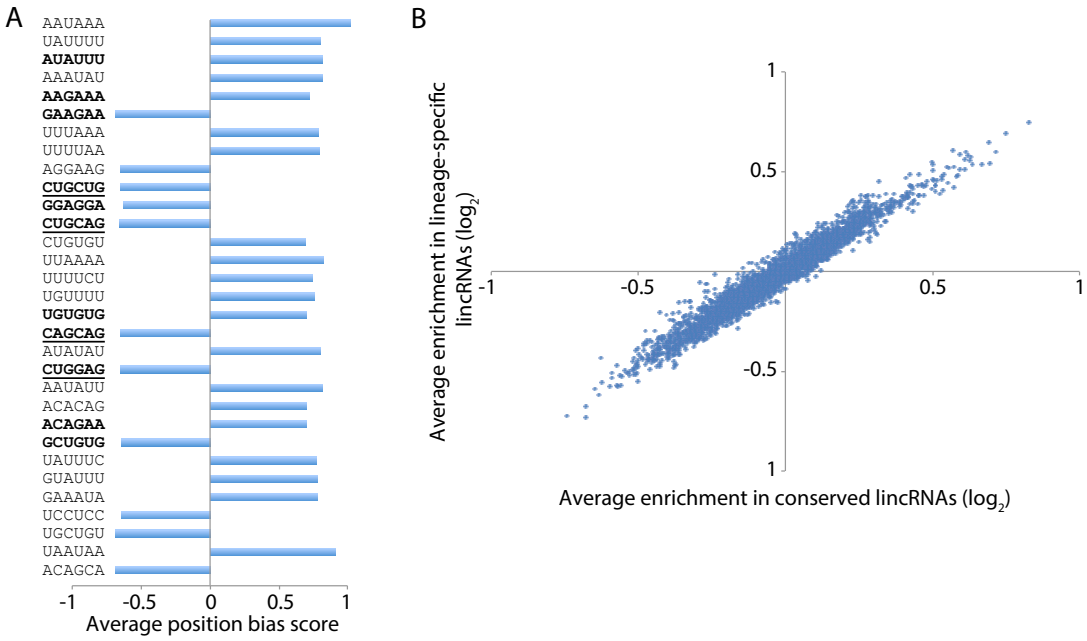

Figure S6. Features of motifs enriched in lincRNA sequences, related to Figure 5. **(A)** Non-redundant motifs significantly enriched in lincRNAs in at least 12 (70%) of the species. Motifs associated with exonic splicing enhancers (Goren et al., 2006) are in bold and CUG/CAG repeats are in bold and underlined. The position bias *p* is the relative position of the motif within the lincRNA sequence (0 for a motif always found in the very beginning of the sequence and 1 for one found always at the very end). The average position bias score is –log(average(p)) for p>0.5 and log(average(p)) for p<0.5. (B) Correlation between the enrichment of 6mers in the sequences of conserved lincRNA and lineage-specific lincRNAs.

Supplementary Tables

Table S1, Related to Figure 1

Details of the genome assemblies and the raw data (RNA-seq and 3P-seq) used in this study

Table S2, Related to Figure 1

Statistics of protein-coding and lncRNA gene numbers in each species, including numbers of genes and isoforms filtered by different steps of PLAR.

Table S3, Related to Figure 1

Genomic positions of the transcripts annotated in 14 vertebrates.

Table S4, Related to Figure 2.

Clusters of potentially orthologous lincRNAs from different species.

Table S5, Related to Figure 5.

Motifs enriched and depleted in lincRNA genes from different species.

Table S6, Related to Figure 6.

Human lincRNAs with sequence conservation outside of amniotes and syntenic lincRNAs in sea urchin.

References

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development *25*, 1915-1927.

Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Lassmann, T., Itoh, M.*, et al.* (2014). A promoter-level mammalian expression atlas. Nature *507*, 462-470.

Goren, A., Ram, O., Amit, M., Keren, H., Lev-Maor, G., Vig, I., Pupko, T., and Ast, G. (2006). Comparative analysis identifies exonic splicing regulatory sequences--The complex definition of enhancers and silencers. Molecular cell *22*, 769-781.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.