Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals

Stefan Washietl,[1] Manolis Kellis,[1,2,5,6] and Manuel Garber[2,3,4,5,6]

[1]*Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02140, USA;* [2]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA;* [3]*Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA;* [4]*Department of Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA*

Long intergenic noncoding RNAs (lincRNAs) play diverse regulatory roles in human development and disease, but little is known about their evolutionary history and constraint. Here, we characterize human lincRNA expression patterns in nine tissues across six mammalian species and multiple individuals. Of the 1898 human lincRNAs expressed in these tissues, we find orthologous transcripts for 80% in chimpanzee, 63% in rhesus, 39% in cow, 38% in mouse, and 35% in rat. Mammalian-expressed lincRNAs show remarkably strong conservation of tissue specificity, suggesting that it is selectively maintained. In contrast, abundant splice-site turnover suggests that exact splice sites are not critical. Relative to evolutionarily young lincRNAs, mammalian-expressed lincRNAs show higher primary sequence conservation in their promoters and exons, increased proximity to protein-coding genes enriched for tissue-specific functions, fewer repeat elements, and more frequent single-exon transcripts. Remarkably, we find that ~20% of human lincRNAs are not expressed beyond chimpanzee and are undetectable even in rhesus. These hominid-specific lincRNAs are more tissue specific, enriched for testis, and faster evolving within the human lineage.

[Supplemental material is available for this article.]

LincRNAs are transcribed by polymerase II and show similar epigenomic, transcriptional, and splicing properties as protein-coding genes, but they do not lead to protein products and act primarily at the RNA level (The FANTOM Consortium et al. 2005; The ENCODE Project Consortium 2007; Amaral et al. 2008; Chodroff et al. 2010; Guttman and Rinn 2012). They play diverse biological roles, including X inactivation (Penny et al. 1996), epigenetic silencing by recruiting chromatin modifying complexes (Rinn et al. 2007; Tsai et al. 2010), retina development (Young et al. 2005), and transcriptional coactivation (Feng et al. 2006). Recent reports have resulted in comprehensive maps of lincRNAs in vertebrates, including human tissues (Cabili et al. 2011), mouse primary cells (Guttman et al. 2010), and zebrafish development (Ulitsky et al. 2011; Pauli et al. 2012). As a class, lincRNAs are highly tissue specific and increasingly recognized as an intrinsic part of the cellular network, where they may serve as modular scaffolds to mediate specific complex protein-RNA-DNA interactions (Tsai et al. 2010; Guttman et al. 2011; Guttman and Rinn 2012).

Across species, lincRNAs have markedly different sequence conservation patterns than protein-coding genes. Although they show clear signs of exonic sequence constraint as a set (Guttman et al. 2009, 2010; Marques and Ponting 2009), they only show small patches of conserved bases surrounded by large seemingly unconstrained sequence (Guttman et al. 2009, 2010). A handful of lincRNAs show sequence conservation across vertebrates (Feng et al. 2006; Chodroff et al. 2010; Ulitsky et al. 2011), but they seem to be the exception rather than the rule (Derrien et al. 2012; Kutter et al. 2012). Previous studies of lincRNA functional conservation included liver lincRNAs between rodents (Kutter et al. 2012) and brain lincRNAs between mouse, chicken, and opossum (Chodroff et al. 2010). However, these studies did not include human lincRNAs for which a comprehensive characterization is still lacking.

Here, we focus on conservation of lincRNA expression levels and characterize their splicing patterns and tissue specificity across nine tissues in six mammals to directly evaluate whether lincRNA activity is evolutionarily constrained, despite their weak primary sequence conservation. We show that a significant subset of human lincRNAs has conserved expression across mammals, with at least 35% showing detectable orthologous transcription across boreoeutheria. These also show conserved tissue-specific gene expression patterns, suggesting the strong tissue specificity of lincRNAs is not fortuitous, but instead selectively maintained across evolutionary time. In contrast, splicing patterns of lincRNAs are highly diverged, suggesting their precise splicing patterns are not essential to their function. Compared to protein-coding genes, we observe extensive gain and loss of lincRNAs across the mammalian lineage with approximately a quarter of lincRNAs becoming expressed after the last common ancestor of human, chimpanzee, and rhesus. However, in spite of the high interspecies turnover, lincRNAs show intra-species expression conservation levels similar to coding genes. For ~20% of lincRNAs, we do not find orthologous expression beyond chimpanzee, even in the closely related rhesus. A detailed comparison of these lincRNAs with conserved expression with those having hominid-specific expression shows several significant differences, including higher tissue specificity, increased repeat content, and accelerated primary sequence evolution across species and even within the human lineage. Our

[5]These authors contributed equally to this work.
[6]Corresponding authors
E-mail manoli@mit.edu
E-mail manuel.garber@umassmed.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.165035.113.

© 2014 Washietl et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

24:616–628 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

study provides the first systematic analysis of human lincRNA evolution and provides an important evolutionary layer to the current annotation of human lincRNAs, which constitutes a rich resource for further experimental and computational studies.

Results

A reference set of lincRNAs in human

The GENCODE catalog is currently the most comprehensive set of manually annotated coding and noncoding gene annotations in human (Derrien et al. 2012). We based our analysis on version 12 of GENCODE, which includes 30,645 noncoding transcripts grouped in 11,790 loci. This set includes transcript types that overlap protein-coding genes such as intronic noncoding RNAs or noncoding isoforms of mRNAs. To best understand the evolutionary properties of noncoding transcripts, we focused on long intergenic noncoding RNAs (lincRNAs) for which we strictly filtered GENCODE noncoding annotations that overlap annotated protein-coding genes in GENCODE, as well as in Ensembl (Flicek et al. 2013) and RefSeq (Pruitt et al. 2012) annotation sets (Fig. 1A).

To further exclude any potential protein-coding transcripts, we removed transcripts with clear evolutionarily conserved coding regions based on RNAcode (Washietl et al. 2011; Methods). At the RNAcode cutoff of $P = 0.01$, we found sensitivity and specificity to be 96% and 96% (Fig. 1B). Although lincRNAs as a class essentially have coding potential indistinguishable from random regions, we found a small number of 397 loci (252 expected false positives) that show signs of protein-coding potential, some of which are compelling novel protein-coding genes candidates (Supplemental Fig. 1; Supplemental Table 1). Importantly, the transcripts with positive RNAcode scores showed clear homology with known protein domains (Finn et al. 2013; Methods). The remaining set of transcripts did not exhibit significant homology with protein compared to random genomic sequence (Methods).

Finally, we only included lincRNAs that were significantly expressed in the human RNA-seq data set. As lincRNAs are known to be highly tissue specific (Cabili et al. 2011; Derrien et al. 2012), we expect only a subset of GENCODE transcripts to be expressed in the tissues we surveyed. We found 1898 loci (37% of 5206 GENCODE intergenic noncoding RNAs) significantly expressed in the tissues surveyed here (significance level 0.05 compared to random regions, see below) (Fig. 1C), which we use for our subsequent analyses. This filter is necessary to select lincRNAs with robustly detectable expression; and indeed, the resulting lincRNA catalog shows significantly higher expression than expected by chance (Methods). As a set however, lincRNAs have significantly lower expression than mRNAs (Fig. 1C), consistent with previous studies (Cabili et al. 2011; Derrien et al. 2012).

The final set consists of 1898 lincRNA loci, including 1375 intergenic loci (GENCODE biotype class "lincRNA," 72%), 434 antisense loci (23%), and 89 unclassified loci (5%). Because of our filters, the antisense transcripts considered here are transcribed from the opposite strand to neighboring protein-coding genes but do not overlap them.

Detection of orthologous lincRNA loci

For each human lincRNA, we identified the best orthologous genomic region in each mammal, using genome-wide pairwise alignments from the UCSC Genome Browser (Karolchik et al. 2014; Methods). These alignments are based on a chaining approach of short conserved segments into long homologous regions (Kent et al. 2003), which is ideal for mapping orthologous transcripts. This approach uses the larger syntenic context to increase sensitivity for the initial alignment step and removes repeats present in ancestral species prior to the alignment to avoid paralogous mapping.

We found aligned sequences for almost all lincRNAs in the primate species, with 98% of lincRNAs in chimpanzee and 93% in

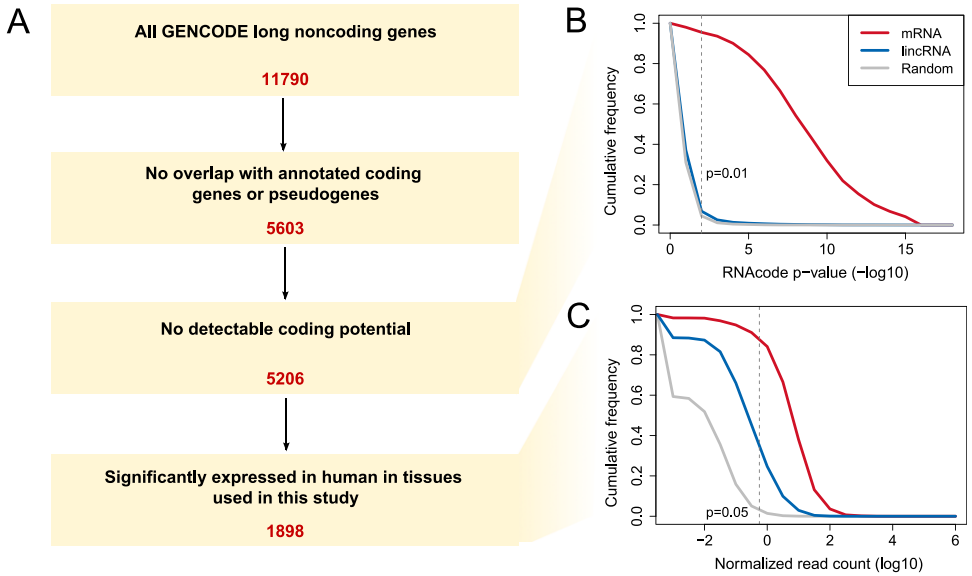


Figure 1. Definition of the lincRNA set. (*A*) Filtering steps of all GENCODE noncoding transcripts to the final set of lincRNAs used for further analysis in this study. (*B*) Cumulative distribution of RNAcode (Washietl et al. 2011) *P*-values measuring the coding potential of transcripts. The *P*-value cutoff of 0.01 is indicated, and for comparison the distributions for coding transcripts and randomized transcripts are also shown. (*C*) Distribution of normalized expression levels in human. The maximum FPKM (fragments per million reads per kb of transcript) over all tissues is shown. The cutoff was chosen empirically using randomized transcripts (Methods) as the background distribution and requiring a significance level of 0.05. If read counts were zero, we set the count to 10^{-3}, explaining the discontinuous shape of the curves.

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

rhesus showing >30% of exonic bases aligned (Table 1; Supplemental Fig. 2). The fraction of loci that can be mapped to the more distantly related mammals rapidly decays, with 73% of lincRNAs in cow, 58% in mouse, and 54% in rat, showing >30% exonic alignment. This fraction is well below that of mRNAs but clearly above random regions (Table 1; Supplemental Fig. 3).

LincRNA expression across mammals

To detect the expression of homologous lincRNAs in other species, we designed a comparative study of multiple tissues and multiple individuals. We used high-coverage RNA-seq data from nine different tissues (colon, spleen, lung, testes, brain, kidney, liver, heart, and skeletal muscle) in four species (rhesus, mouse, rat, and cow) (Supplemental Table 2; Methods). This data set published by Merkin et al. (2012) was previously analyzed for protein-coding genes, and we describe here their initial analysis to study lincRNAs. We complemented this data set with lower-coverage RNA-seq in six tissues in human and chimpanzee (Brawand et al. 2011).

To assess the conservation of human transcription in the other species, we calculated the read counts over orthologous exonic positions. To ensure highest sensitivity, we combined all tissues from all individuals in this analysis. As a control, we also calculated the read counts for mRNAs and for random genomic regions (Methods).

For mRNAs, expression is nearly constant across all species, regardless of their evolutionary distance (Fig. 2A). For lincRNAs, however, expression conservation declines faster than sequence conservation, suggesting a high turnover of lincRNAs compared to mRNAs (Fig. 2A). Interestingly, this trend already starts to show within the primate clade.

We first confirmed that this difference is not due to the lower expression level of lincRNAs reducing our ability to detect their transcripts in other species. For a subset of mRNAs expressed at the same levels as lincRNAs (Fig. 2A, dotted line), expression levels remained essentially unchanged throughout all species.

We continue to observe the same trends when restricting the analysis to lincRNAs that can be reliably (uniquely and reciprocally; Methods) mapped between human and the other species (Supplemental Fig. 4), indicating that lack of orthologous expression is not due to poor mappability, and that the differences we see are indeed due to evolutionary turnover.

Third, we found that despite their low interspecies expression conservation, lincRNAs show remarkably reproducible expression across individuals, similar to that of mRNA genes, showing that their expression is not stochastic; and that the observed interspecies divergence is not due to technical artifacts limiting our ability to measure their expression levels accurately (Supplemental Figs. 5, 6A).

Table 1. Fraction of human loci mapped to other species

	mRNA	lncRNA	Random
Chimp	0.99	0.98	0.96
Rhesus	0.98	0.93	0.86
Cow	0.97	0.73	0.51
Mouse	0.96	0.58	0.36
Rat	0.93	0.54	0.32

A locus is included here if >30% of the exonic bases could be mapped from human to the other species.

Evidence of extensive gain and loss

In addition to these global expression distributions, we sought to identify individual human lincRNAs with conserved expression in orthologous regions in the other five species using an empirical expression level cutoff (Supplemental Fig. 6B; Methods). Of the 1898 human lincRNAs significantly expressed in the tissues surveyed in this study, and consistent with a previous report that focused on rodent liver lincRNAs (Kutter et al. 2012), we found evidence for orthologous transcription for 1523 lincRNAs (80%) in chimpanzee, 1196 (63%) in rhesus, 734 (38%) in cow, 715 (38%) in mouse, and 660 (35%) in rat (Fig. 2B). This shows that rapid turnover of large noncoding transcripts has occurred throughout the phylogeny. We observe a higher turnover than previously reported using sequence mapping only (Derrien et al. 2012). Indeed a surprising large portion of transcripts with a clear ortholog fails to have detectable expression. For example, >93% of lincRNAs are alignable to rhesus but only 63% show significant orthologous expression.

We used a parsimony model to determine gain and loss events for each branch given the species phylogeny (Fig. 2C, top; Methods). The model suggests that 55% of human lincRNAs date back to the last common ancestor of the boreoeutherian mammals studied here; 76% date back to the last common ancestor of human, chimpanzee, and rhesus; and 92% to the last common ancestor of human and chimpanzee. In the rodent branch, 44% of human lincRNAs can be found in the last common ancestor of mouse and rat. As two interesting classes, we point out on one hand evolutionarily young lincRNAs (e.g., Fig. 2D) that are consistently expressed in human and chimpanzee with conserved splice sites, but show turnover in rhesus and are undetectable in more distant mammals, and ancestral lincRNAs (e.g., Fig. 2E) that are consistently expressed in all the tested mammalian species.

Our parsimony approach shows that 62% of lincRNAs can be explained by a single gain event and no loss, 26% require at least one loss event (of which a quarter are lost in the rodent lineage), and 12% require two independent loss events. These results suggest substantial turnover of lincRNAs, but they have to be interpreted in light of inherent limitations to detect all transcripts accurately (low expression levels, errors in read mapping, and genome assembly errors).

Conservation of lincRNA tissue specificity

One of the most striking characteristics of lincRNAs is their extremely tissue-specific expression (Cabili et al. 2011), which may be key to their function (Guttman et al. 2011), but it is unclear whether this tissue specificity is fortuitous or selectively maintained. We and others had previously reported that the level of primary sequence conservation for lincRNA promoters is nonrandom (Ponjavic et al. 2007) and similar to that of protein-coding gene promoters (Guttman et al. 2009), suggesting similar levels of regulatory constraint. However, it is unclear whether this increased constraint would be sufficient to maintain expression levels, or whether new and distinct expression patterns would evolve across different species.

To address this question, we studied the tissue specificity across the nine tissues for the 323 lincRNA loci that are significantly expressed (Methods) in all high-coverage tissue libraries of rhesus, mouse, rat, and cow. We calculated a tissue specificity score (Cabili et al. 2011) for each lincRNA in each species, measuring how strongly the expression is dominated by a single tissue. We observed remarkably similar levels of tissue specificity for orthologous

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

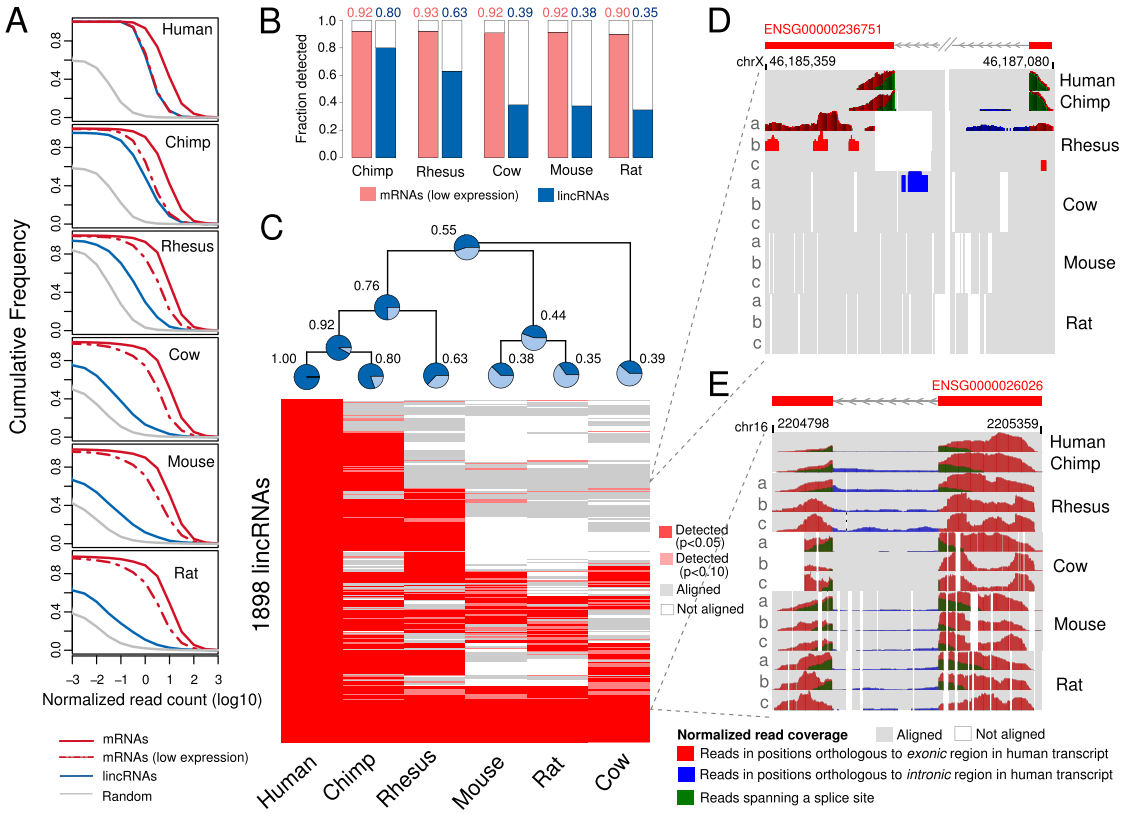


Figure 2. Conservation of lincRNA expression across placental mammals. (*A*) Cumulative distributions of normalized read counts (number of reads per million reads in the library per kb of the transcript portion that could be aligned to the other species). The maximum of this normalized count of all tissues is considered for the distribution shown. We use a floor of 10^{-3} whenever no reads were found in any tissue or the transcript could not be aligned. (*B*) Fraction of human lincRNAs that were detected in other species. A lincRNA is counted as detected if it either was expressed with an empirical *P*-value of $P < 0.1$ compared to random regions or if it is supported by conserved splice sites (Methods). In comparison, the detection rate for mRNAs with similar expression levels as the lincRNAs are shown (to be conservative in this comparison, we only used the expression *P*-value cutoff because mRNAs have more and better conserved splice sites). (*C*) Conservation patterns of individual lincRNAs. The fraction at the tips of the phylogenetic tree corresponds to the fraction of detected lincRNAs in *B*. The fractions for the inner nodes are estimated using a parsimony approach (Methods). *D* and *E* show the actual read patterns observed in the different species for two lincRNA examples. Read counts were normalized between 0 and 1 for each line; only positions with absolute read coverage greater than five are shown. For rhesus, cow, mouse, and rat, all three replicates are shown (indicated by a, b, c). Example *D* shows a lincRNA well-supported in human and chimpanzee but absent in all replicates in the more distantly related mammals. Example *E* shows a transcript conserved in all species also supported by all replicates.

lincRNAs between species (Fig. 3A, right). Ubiquitously expressed lincRNAs in human were ubiquitous across all species (e.g., TUG1) (Fig. 3D) and tissue-specific lincRNAs in human were tissue specific in all species (e.g., Fig. 3E).

Moreover, lincRNAs were consistently expressed in the same tissues across species (Fig. 3A). The correlation coefficients of normalized expression counts across tissues are similar for both lincRNAs and mRNAs (Fig. 3C; Supplemental Fig. 7). The tissue-specific nature of lincRNA expression patterns extended to all nine tissues studied, although the largest clusters of tissue-specific lincRNAs were found in testis and brain, where human lincRNAs are known to be highly expressed (Cabili et al. 2011). Both tissues showed remarkable conservation of tissue specificity across species, suggesting that these are not subject to promiscuous expression, but instead highly regulated expression patterns that are selectively maintained.

An unbiased clustering of lincRNA expression patterns across all tissues and all species resulted in a perfect separation of all nine tissues (Fig. 3B). Consistent groups of colon, spleen, lung, testes, brain, kidney, liver, heart, and skeletal muscle were found, regardless of the species in which they were profiled. These results further indicate that similar to protein-coding genes (Barbosa-Morais et al. 2012; Merkin et al. 2012), the expression profiles of lincRNAs are conserved across species and strongly defined by tissue identity and only to a lesser extent by species identity.

Thus, despite having lower sequence conservation than mRNAs, lincRNAs show similar levels of regulatory conservation as protein-coding genes. These findings are consistent with an earlier study that showed conserved tissue-specific intergenic transcription between human and chimpanzee in brain, heart, and testis (Khaitovich et al. 2006). Conservation of tissue specificity of lincRNAs might be an indirect effect through coregulation with mRNAs. We found, however, that lincRNAs that are expressed in sense and antisense orientation relative to the closest protein gene do not show significantly different conservation of tissue specificity (Supplemental Fig. 7B). Interestingly, lincRNAs close (<10 kb) to protein-coding genes show consistent lower conservation of tissue specificity than lincRNAs distant (>10 kb) to protein-coding genes (Supplemental Fig. 7B). These results suggest that conservation of tissue specificity is not just a by-product of protein-coding gene regulation but rather an inherent property of lincRNAs.

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

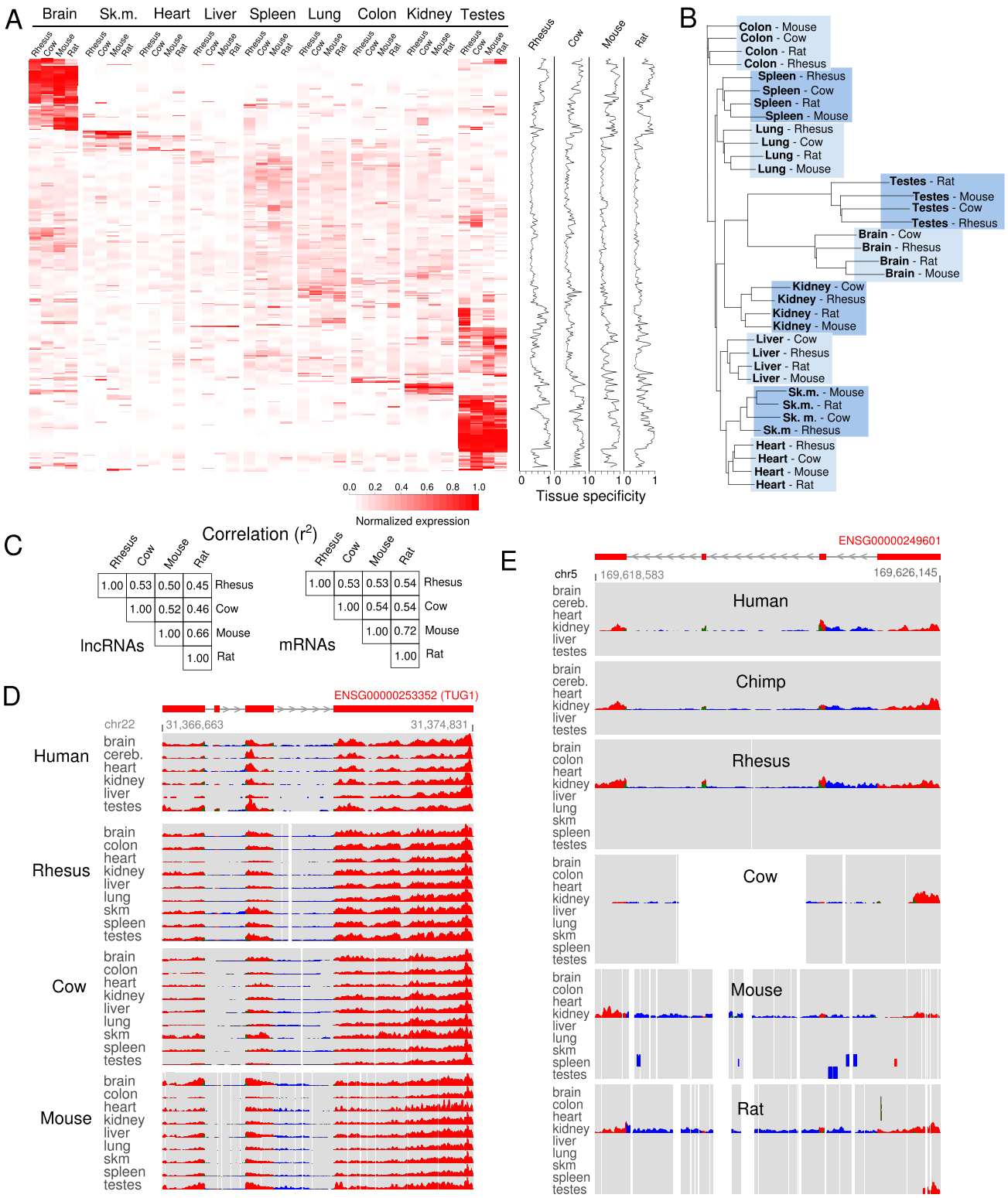


Figure 3. Tissue specificity of lincRNAs across species. (*A*) Heatmap of normalized expression values (see Methods) for all tissues and species. Data is only shown for lincRNAs that have significant ($P < 0.1$; Methods) expression in rhesus, cow, mouse, and rat. On the *right* of the heatmap, a normalized tissue specificity score is shown for all species (Methods). (*B*) Neighbor-joining tree generated from the similarity matrix of expression values across all lincRNAs in all tissues and species. (*C*) Correlation of expression between species across all tissues for lincRNAs and mRNAs. *D* and *E* show examples of a lincRNA ubiquitously expressed in all tissues and a lincRNA highly restricted to kidney, respectively. The same conventions as in Figure 2 are used.

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

Evolution of splicing patterns

Having established that tissue-specific expression patterns are strongly conserved for the set of lincRNAs with clear orthologs, we next investigated the degree of conservation of their gene structure. Previous studies reported primary sequence conservation between human and mouse at splice-site motifs (Ponjavic et al. 2007). Consistent with these findings, we observed that the fraction of splice sites that can be aligned is relatively high in all species: In rhesus, 90% of lincRNA splice sites are conserved at the sequence level (compared to 94% for coding and 91% for UTR splice sites); and in rat, 62% of lincRNAs splice sites are conserved (compared to 89% for coding and 71% for UTRs) (Table 2). It is unclear, however, whether this primary sequence conservation would also result in conservation of splicing events, given the diversity of signals involved in splicing (Wang and Burge 2008).

We first quantified the level to which exons are maintained between species. We assembled transcripts from the high coverage RNA-seq data sets in rhesus, cow, mouse, and rat using Cufflinks (Trapnell et al. 2010) and compared the predicted exons to the human GENCODE reference transcripts (see Methods). We found that 73% of exons in reconstructed transcripts show conserved expression in rhesus, and ~40% show conserved expression in the other species, compared to 83%–89% for coding exons (Supplemental Fig. 8).

We next compared the exon boundaries of orthologous exon pairs. We found that lincRNA exon boundaries show larger and more frequent changes across mammals than for protein-coding genes (Fig. 4A). For example, lincRNAs show 2.3 times fewer orthologous exon boundaries within 25 nt of the reference exon in mouse compared to coding exons. Thus, even for exons with conserved expression, lincRNAs show less constraint on maintaining an exact position of splicing events.

We next compared exonic and intronic read counts surrounding the splice sites of lincRNA exons, coding exons, and untranslated region (UTR) exons of mRNAs (Methods). We found a clear conservation signature for coding exons, consisting of a sharp boundary between high exonic and low intronic read counts (Fig. 4B). In contrast, lincRNA splicing shows a much weaker signature than coding genes, and remarkably, even weaker than in UTRs (Fig. 4B). This difference is clearly visible in the normalized read count around all splice sites (Methods), showing high conservation for coding exons, a gradual decline for UTRs, and an even faster decline for lincRNAs with increasing evolutionary distance (Fig. 4C).

We next sought to identify individual exons with conserved expression, using split reads that span exon junctions (Methods). In human, this approach recovered 89% of annotated human coding splicing events, 72% of lincRNA splicing events, and 71% of UTR splicing events (Table 1), providing a benchmark for our detection rate due to coverage and mappability of split reads. Applying this signature to the other species and restricting our analysis to lincRNA splicing events recovered in human, split reads confirm only 64% of aligned junctions in chimpanzee, although 96% are aligned at the sequence level. In rhesus, 90% of junctions are aligned at the sequence level, but only 56% of the corresponding splicing events are supported by split reads. Outside primates, 62%–72% of lincRNA splice sites can be aligned, but only 21%–29% of the corresponding splicing events are supported by split reads. By comparison, 87%–90% of protein-coding exon splicing events and 50%–55% of UTR splicing events are detected as conserved using the same method (Table 2). This suggests that disruptions of lincRNA structure may result in little functional consequence, and perhaps certain regions of lincRNAs are not necessary for their function, providing a potential explanation for their overall low primary sequence conservation.

We find that lincRNA exon junctions with conserved splicing events between human and mouse also show significantly higher primary sequence conservation than junctions with diverged splicing events ($P < 10^{-21}$, Mann-Whitney) (Fig. 4B, bottom right). This suggests that primary sequence changes are accompanying splicing event changes, and conserved splicing events may be actively maintained by selective constraint at the primary sequence level. These splice junctions may span functionally critical elements of lincRNAs or may be important for splicing-associated regulatory events.

We further asked if splice-site turnover is equally distributed across transcripts or if there are subpopulations of transcripts with particularly high and low splice-site turnover. We found a dramatic range of conservation between different lincRNAs, from lincRNAs with highly constrained splice sites across all species, to lincRNAs with complete splice-site turnover even in closely related primate species (Fig. 4D).

Differences between lincRNAs with conserved expression and lineage-specific expression

We next asked if lincRNAs with lineage-specific expression (e.g., Fig. 2D) and lincRNAs with conserved expression throughout the mammalian lineage (e.g., Fig. 2E) show different characteristics. We defined 376 "hominid-expressed" lincRNAs, for which evidence of transcription could not be found beyond human and chimpanzee, and 549 "mammalian-expressed" lincRNAs, for which transcription was consistently detected in all the primates (human, chimpanzee, rhesus) and in one or more additional mammals (mouse, rat, or cow).

First, we ensured that the set of hominid-expressed lincRNAs are not due to spurious transcripts that were incorrectly annotated in GENCODE or false positives in our expression analysis. We found that the hominid-expressed lincRNAs show comparable

Table 2. Splice-site conservation (see text for details)

	Aligned			Confirmed			Confirmed (of human confirmed sites)		
Species	Coding	UTR	lincRNA	Coding	UTR	lincRNA	Coding	UTR	lincRNA
Human	1.00	1.00	1.00	0.89	0.71	0.72	1.00	1.00	1.00
Chimp	0.97	0.96	0.96	0.82	0.58	0.51	0.90	0.76	0.64
Rhesus	0.94	0.91	0.90	0.82	0.56	0.46	0.89	0.70	0.56
Cow	0.94	0.80	0.72	0.82	0.42	0.23	0.90	0.55	0.29
Mouse	0.92	0.73	0.64	0.82	0.39	0.18	0.90	0.51	0.21
Rat	0.89	0.71	0.62	0.79	0.38	0.18	0.87	0.50	0.22

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

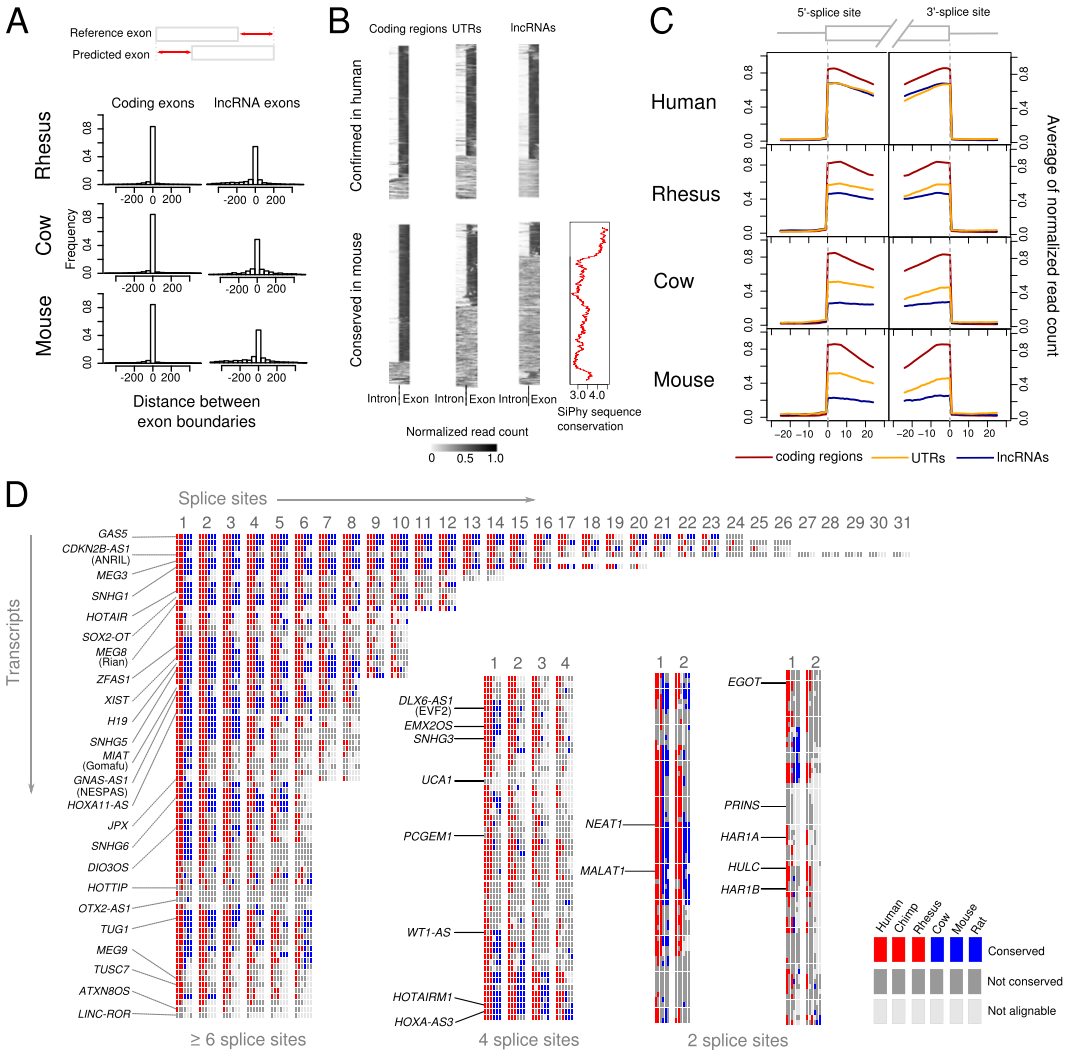


Figure 4. Conservation of splicing patterns across species. (*A*) Conservation of exon boundaries. The distributions show the difference of exon boundaries of reference exons from the human GENCODE annotation and predicted exons in the other species. (*B*) Normalized read density in a window of 50 nucleotides around splice sites in human and mouse. Both 5′- and 3′-splice sites are shown. Only splice sites for which at least half of the positions could be aligned in mouse were considered. The graph at the *bottom right* shows the SiPhy conservation scores for splice sites in mouse. The mean score averaged over all aligned positions in the 50-nt window and a running average over 100 splice sites is shown. (*C*) Averaged normalized read count in a 50-nt window around 3′- and 5′-splice sites in human, rhesus, cow, and mouse. Again, only splice sites with more than half the positions in the window aligned were considered. Also, only "split reads" that map to two regions across an exon/intron boundary were counted. (*D*) Splice-site conservation patterns of individual transcripts. Each line represents a transcript. Each group of boxes represents a splice site (both 3′- and 5′-sites are shown separately, i.e., two splice sites means a transcript has two exons and one intron). Each box within a group indicates the conservation status in the different species. All multiexon lincRNAs are shown for which we could detect significant expression ($P < 0.1$; Methods) in human, chimpanzee, rhesus, cow, mouse, and rat. All known lincRNAs from lincRNAdb are included and highlighted with their name. If a locus had multiple isoforms, the isoform with the most confirmed human splice sites is shown, which is not necessarily the most abundant transcript.

levels of expression to the mammalian-expressed lincRNAs (Fig. 5A). Moreover, we tested what fraction of GENCODE-annotated splice sites in hominid-expressed and mammalian-expressed lincRNAs are independently supported by the RNA-seq data in our study. The fraction of hominid-expressed lincRNAs with supported splice sites is even slightly higher than for the mammalian-expressed lincRNAs (88% and 83%, respectively). Thus, although they do not show conserved expression beyond chimpanzee, hominid-expressed lincRNAs appear to be bona fide transcripts whose annotation is not of lower quality. Mammalian-expressed and hominid-expressed lincRNAs showed little difference in their length, number of isoforms, or relative orientation to the closest protein-coding gene (Supplemental Fig. 9), but several other properties set them apart.

We compared the level of primary sequence constraint across mammals (Lindblad-Toh et al. 2011) as measured by the SiPhy algorithm (Garber et al. 2009) for mammalian-expressed versus hominid-expressed lincRNAs. Mammalian-expressed lincRNAs showed greater constraint than hominid-expressed lincRNAs ($P < 6 \times 10^{-18}$, Mann-Whitney, two-tailed) (Fig. 5C) for their primary sequence both across the transcript and at the predicted transcription start sites (TSS; $P < 4 \times 10^{-17}$, Mann-Whitney, two-tailed), suggesting they are more likely to have conserved functions and conserved regulation. We also evaluated the sequence

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

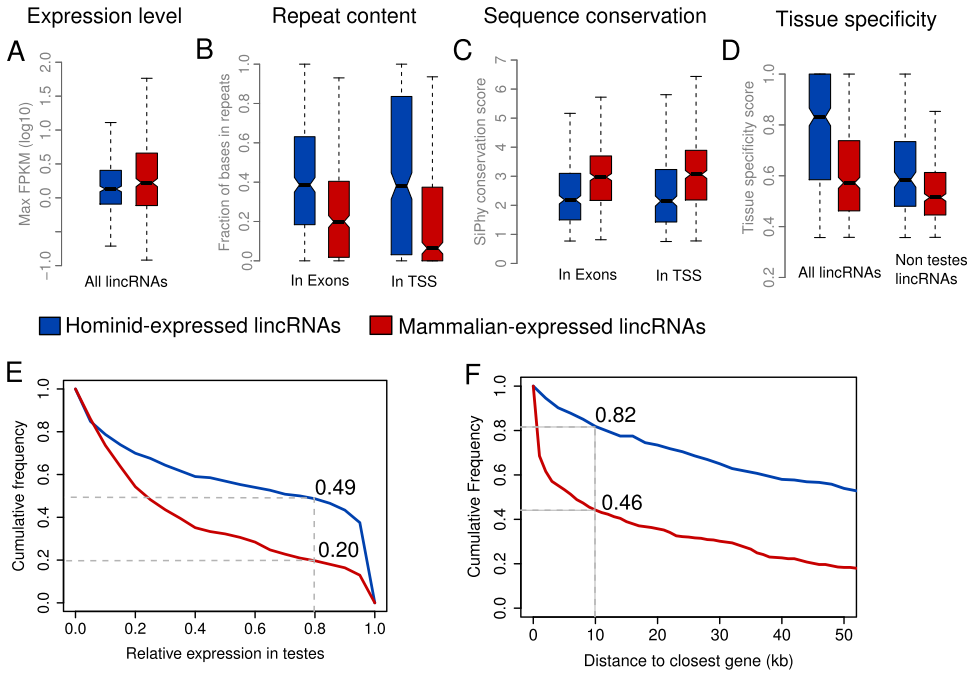


Figure 5. Differences between hominid-specific lincRNAs and lincRNAs conserved across mammals. Distributions are shown as box plots indicating the first quartile, median, and third quartile. Whiskers represent the range of the data without outliers. (*A*) Normalized expression level in human. The highest expression in all tissues is shown. (*B*) Repeat content. The fraction of repeat-masked bases in the exons (union over all isoforms) of a lincRNA locus and in the putative transcription start site (window 350 upstream and 150 around the annotated transcript start) is shown. (*C*) Sequence conservation as measured by SiPhy for exons and putative transcription start site (Methods). (*D*) Tissue specificity score (Methods). (*Left*) All lincRNAs of both sets are considered. (*Right*) lincRNAs that have a relative expression level higher than 0.8 in testis were removed. (*E*) Distribution of relative expression in testis (Methods). (*F*) Cumulative distribution of the distances of human lincRNA loci to the closest annotated (Ensembl version 64) protein-coding gene.

conservation of lincRNAs using alignments made specifically with human, chimpanzee, gorilla, orangutan, and macaque (see Methods). Even with this reduced power, mammalian-expressed lincRNAs are significantly more constrained than randomly sampled genomic sequence ($P < 3 \times 10^{-16}$, Mann-Whitney, two-tailed). In contrast, hominid-expressed lincRNAs are not significantly more conserved at the sequence level than randomly sampled genomic sequence ($P > 0.01$, Mann-Whitney, two-tailed).

We also compared the lincRNA level of sequence constraint within the human lineage using a derived allele frequency (DAF) metric, a commonly used test for measuring lineage-specific selection (Sabeti et al. 2006; Voight et al. 2006). We had previously found that lincRNAs as a group showed lower DAF than control regions, suggesting they are preferentially constrained in human (Ward and Kellis 2012), although their sequence conservation across the mammalian lineage is much weaker (Guttman et al. 2009; Marques and Ponting 2009; Chodroff et al. 2010; Ward and Kellis 2012). With the ability to distinguish mammalian-expressed lincRNAs and hominid-expressed lincRNAs, we asked if they showed differences in their DAF distribution. We calculated DAF using the expanded number of human genomes available from Phase 1 of the 1000 Genomes Project Consortium (The 1000 Genomes Project Consortium 2012) and using improved methods that correct for varying coverage associated with varying GC content (Green and Ewing 2013; Ward and Kellis 2013).

We found that mammalian-expressed lincRNAs show lower DAF than our reference neutral controls (regions not covered by ENCODE annotations), consistent with purifying selection in the human lineage. In contrast, hominid-expressed lincRNAs showed higher DAF than neutral controls, suggesting they may be under positive selection at the sequence level (Supplemental Table 3). We also measured the rate of divergence of hominid-expressed lincRNAs in primate alignments using both the SiPhy omega rate and the LOD score, measuring the significance of that rate. Using both measures, hominid-expressed lincRNAs showed an excess of rapid divergence relative to mammalian-expressed lincRNAs ($P < 2 \times 10^{-6}$, Mann-Whitney, two-tailed). These results are consistent with either positive selection or lower constraint for hominid-expressed lincRNAs relative to mammalian-expressed lincRNAs.

Interestingly, in spite of their similar overall expression levels, mammalian-expressed and hominid-expressed lincRNAs clearly show different repeat content (Fig. 5B). Exons of mammalian-expressed lincRNAs show lower repeat content (25%) than hominid-expressed lincRNAs (42%; $P < 10^{-18}$, Mann-Whitney, two-tailed), and their putative TSS have even fewer repeats than their exons ($P < 10^{-9}$, Mann-Whitney, two-tailed). In contrast, hominid-expressed lincRNAs, show no difference in repeat content between their putative TSS and exonic regions ($P = 0.87$). The reduced repeat content might indicate selection in the mammalian-expressed lincRNAs against disruption by repeat insertions that may disrupt *cis*-regulatory promoter sequence or RNA structure.

Furthermore, we compared tissue specificity to see if one of the two classes is restricted to specific tissues and thus potentially has more specialized functions. We found that hominid-specific lincRNAs are more tissue specific than conserved lincRNAs ($P < 10^{-30}$, Mann-Whitney, two-tailed) (Fig. 5D). They are 2.5-fold enriched for testis-specific transcripts, with 49% showing greater than 0.8 relative expression in testis (see Methods), compared to

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

20% for conserved lincRNAs (Fig. 5E). Even after excluding all testis-specific lincRNAs, hominid-specific lincRNAs are still more tissue specific than mammalian-conserved lincRNAs ($P < 10^{-7}$, Mann-Whitney, two-tailed) (Fig. 5D), an effect present in all tissues similarly.

It was previously reported that protein-coding genes that are neighbors of lincRNAs are enriched in specific functional classes (Guttman et al. 2009). We found that conserved lincRNAs are closer to protein-coding genes than hominid-specific lincRNAs ($P < 6 \times 10^{-43}$, Mann-Whitney, two-tailed) (Fig. 5F), with ~50% within 10 kb of the closest protein-coding gene compared to 20% for hominid-specific lincRNAs. To ensure that proximity to well-conserved protein-coding genes is not a confounding factor, we repeated the previous analyses separately considering genes within and outside 10 kb of the closest protein-coding genes. In both cases, we obtained qualitatively very similar results for the comparisons of sequence constraint, repeat content, and tissue specificity (not shown).

Similarly to neighboring coding gene pairs, lincRNA-coding gene neighbors are frequently coregulated and are enriched in cell type-specific functional categories (Guttman et al. 2009; Cabili et al. 2011). We studied the gene ontology enrichments of neighboring coding genes of conserved and hominid-specific lincRNAs. We found a dramatic difference, with protein-coding genes neighboring conserved lincRNAs enriched in tissue-specific cellular functions. For example, coding genes next to conserved lincRNAs expressed in brain are significantly enriched in brain function or in brain expressed genes. In contrast, we find no significant enrichment for coding genes neighboring hominid-specific lincRNAs (Supplemental Material).

Although the majority of lincRNAs are multiexonic, conserved lincRNAs are 2.5 times more frequently single-exon lincRNAs compared to the hominid-specific set (18% versus 8%, $P < 4 \times 10^{-6}$, Fisher's exact test). Conserved lincRNAs also have a 3.4-fold higher fraction annotated as "known" by GENCODE, which means they have been annotated also by the RefSeq (Pruitt et al. 2012) and HUGO Gene Nomenclature Committee projects (Seal et al. 2011) (7% versus 2%, $P < 5 \times 10^{-4}$, Fisher's exact test). The increased enrichment of conserved lincRNAs in curated annotations may be partly due to an ascertainment bias, as conserved functions are more likely to be curated, but may also suggest that conserved lincRNAs are more likely to be functional than non-conserved lincRNAs.

Discussion

Although it is increasingly recognized that lincRNAs are key components of gene regulation and a diversity of mechanisms of action have been proposed (Rinn and Chang 2012), the selective pressures acting on human lincRNAs are still uncharacterized. Studies of lincRNA conservation have been plagued by distinct lincRNA properties that distinguish them from protein-coding genes. First, although the primary sequence of protein-coding genes is constrained by its amino acid translation, leading to very high and specific sequence conservation, the primary sequence of lincRNAs is significantly less constrained, making orthology search a significant challenge. Second, the expression levels of lincRNAs are significantly lower than those of protein-coding genes, making it difficult to distinguish evolutionary divergence from lack of detection. Last, lincRNAs are highly tissue specific, making it difficult to detect orthologous expression unless matching tissues are available.

In our study, we address these shortcomings by exploiting the extensive conservation of mammalian synteny to detect lincRNAs in orthologous loci by exploiting deeply sequenced RNA-seq libraries only recently made possible, and by surveying multiple tissues in each species. Moreover, access to multiple individuals per species makes it possible to distinguish true evolutionary divergence between species from stochastic or spurious transcription because we find high reproducibility of lincRNA transcription between individuals of the same species.

Our phylogenetic analysis suggests that 55% of lincRNAs date prior to the last common ancestor of the placental mammals tested, an estimate significantly higher than previous estimates of 12%–15% based on public EST data (Cabili et al. 2011). However, we find that the rate of lincRNA turnover is much higher than for mRNAs and also surprisingly high between closely related species, with only 63% of human lincRNAs showing conserved expression in the closely related rhesus. The accelerated evolution of lincRNAs may be due to lower purifying constraint, or positive selection associated with environmental adaptations, as lincRNAs could contribute to regulatory plasticity given the highly conserved functions of protein-coding genes. Consistent with the second possibility, hominid-specific lincRNAs show significantly higher derived allele frequencies within the human population than neutrally evolving regions, suggesting that they have been subject to recent positive selection since divergence from chimpanzee.

We also find striking conservation properties of lincRNAs that give new clues into their function. LincRNAs are known to be highly tissue specific, but our results indicate that their tissue-specific expression is not stochastic or fortuitous; it appears to be tightly regulated and selectively maintained across evolutionary time, as conserved lincRNAs show promoter conservation levels similar to mRNAs and are expressed in the same tissues across distantly related species. In contrast to their conserved tissue-specific expression, however, gene structure is poorly conserved; even for lincRNAs with conserved expression, we find very high levels of splice-site turnover, substantially higher than for protein-coding exons and even UTRs. Not even a quarter of splice sites are supported by spliced reads in the more distantly related mammals compared to almost 90% for protein-coding exons, suggesting that transcript structure and exact splicing patterns are not critical for lincRNA function and that purifying selection is not acting on the linear RNA polymer but more likely on only portions of the molecule or on its folding structure.

We find clear differences between hominid-expressed lincRNAs and mammalian-expressed lincRNAs, suggesting potentially distinct roles. lincRNAs with conserved expression show higher levels of sequence constraint, implying that they contain functional sequence elements beyond simply their property of transcription. Conserved lincRNAs are also situated closer to protein-coding genes and more frequently enriched in genes that are expressed in the same tissue or with function associated with the tissue where the lincRNA is expressed. This is potentially due to regulatory relationships established early in mammalian evolution. Evolutionarily younger lincRNAs are less conserved across both mammals and primates; and within humans are more tissue specific and particularly enriched for testis expression. Testis specificity of lincRNAs was observed previously in various species and suggests roles in sexual selection or testis-specific processes such as piRNA production.

Repetitive sequences are more common in the evolutionarily young lincRNAs. Although there may be selection against disruption by repeat insertions in conserved lincRNAs, hominid-specific

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

lincRNAs may result from exaptation of repetitive sequence or just from stochastic acquisition of a cell type-specific *cis*-regulatory sequence that drives expression. One possible interpretation is that new repetitive elements may replace existing lincRNAs, or make them redundant, by binding similar protein complexes or DNA locations, thus decreasing selective pressures and resulting in the observed high turnover. An alternative model is that younger lincRNAs are less likely to be functional, and their expression is a consequence of fortuitous binding tissue-specific transcription factors. Our catalog of hominid-specific and mammalian-conserved lincRNAs provides an important resource that can guide directed experimental studies to resolve these possibilities.

The very high tissue specificity and the rapid turnover of lincRNA transcripts are both in stark contrast to protein-coding genes that are often widely expressed and nearly always very deeply conserved. This raises a compelling hypothesis of a functional and evolutionary interplay between protein-coding genes and lincRNAs. Although the functions of protein-coding genes are very rigid and slow evolving, lincRNAs could modulate the activity, DNA targets, or interaction partners of protein-coding genes in a tissue-specific way, enabling them to rapidly adapt to new functions, conferred by rapidly evolving lincRNA partners (Guttman and Rinn 2012).

The question of what fraction of lincRNAs has functional roles is still under debate. Our data revealed conserved transcription over evolutionary time scales for a substantial fraction of lincRNAs and thus points to their functional importance. Unfortunately however, we still know very little about these genes and their specific mechanisms of action. As opposed to coding genes for which tests for adaptive evolution are well established (Yang and Bielawski 2000), we do not yet have established statistical methods for evaluating lincRNA adaptive selection. As the field advances and the exact structures and mechanism of function are established, we may be able to dissect the specific aspects of lncRNA function that are under accelerated evolution, purifying constraint, or neutrally evolving, and reconcile their high tissue specificity with their apparently rapid evolutionary turnover.

Methods

Sequence data

All genomic sequences were downloaded from the UCSC Genome Browser (Karolchik et al. 2014). We used the following assemblies: hg19 (human), panTro3 (chimpanzee), rheMac2 (rhesus), bosTau6 (cow), mm9 (mouse), rn4 (rat).

Filtering and selection of a human reference lincRNA set

Starting with all noncoding transcripts in GENCODE 12, we applied several filtering steps. We excluded all lincRNAs that had any overlap with annotated protein-coding genes from GENCODE, Ensembl (version 64), or RefSeq.

In addition, we removed all transcripts that were annotated as a pseudogene of any type (processed, unprocessed, transcribed, etc.) or annotated by Ensembl as "ambiguous_orf," "IG_V_gene," "retained_intron," "retrotransposed," "TEC," or "TR_V_gene." From the resulting set, we only kept GENCODE loci of type "lincRNA," "antisense," "non_coding," and "processed_transcript." It is important to note that because of our filters, the transcripts of type "antisense" are transcribed from the opposite strand to neighboring protein-coding genes but do not overlap them.

We kept all 43 GENCODE loci that were listed in lncRNAdb (Amaral et al. 2011) and added six lincRNAs from RefSeq that were listed in lncRNAdb but not in GENCODE (DISC2, NR_002227; LUST, NR_045388; NRON, NR_045006; SAF, NR_028371; Tsix, NR_003255; ncR-uPAR, NR_028375).

Control sets

As positive controls, we used mRNAs from GENCODE version 12. We randomly selected 6412 loci (roughly one-third of all loci that were annotated as "protein_coding" with status "KNOWN"). In addition, we created a randomized set of transcripts. First we created a list of intergenic regions that do not overlap Ensembl, RefSeq, or GENCODE transcripts. We randomly placed each lincRNA in our set into a random intergenic region at a random position. We repeated this process seven times. We found that these random regions still contained regions that overlap known transcripts in human or other species. We therefore added an additional filtering step and excluded all regions that overlap with the following annotation tracks from the UCSC Genome Browser: "human mRNAs," "transmapped mRNAs," and "xeno-mRNAs." This process finally resulted in a set of 6186 random loci.

Coding potential

We used RNAcode (Washietl et al. 2011) to evaluate the coding potential of GENCODE lincRNAs. RNAcode uses a comparative approach to detect evolutionary signatures of protein-coding regions in multiple sequence alignments. The main signatures are synonymous mutations in the DNA sequence that do not change the amino acid sequence, conservative mutations that change amino acids to biochemically similar amino acids, and conservation of the reading frame. We used alignments of 29 mammalian species (Lindblad-Toh et al. 2011) that were generated by LASTZ (Harris 2007). We extracted all alignment regions corresponding to exonic regions in the lincRNAs (we considered all exons of all isoforms). For efficiency reasons, we divided blocks longer than 400 columns in nonoverlapping blocks of around 200 columns, following protocols in Washietl et al. (2011). Those blocks were directly scored with RNAcode using the parameters "–best-only -p 1.0." That command reports all possible reading frames and their associated *P*-values. If two reading frames overlap, it reports only the higher scoring reading frame. As overall score for a locus, we report the *P*-value of the best scoring reading frame of all blocks of a locus.

Comparative approaches have reduced power when regions are poorly conserved. To further filter transcripts that may have coding potential, we searched for significant homology with known protein domains using PfamScan (Released October 15, 2013) with default parameters against the Pfam database version 27 (Finn et al. 2013). To control for random homology with coding domains, we used size-matched randomly selected nonexonic sequences. Excluding domains that were more or equally frequent in the random set than in our lincRNA sets, only two of the hominid-specific lincRNAs showed homology with a protein domain (one to a Zinc finger) that were at a level similar to that of protein-coding genes. These putative lincRNAs may be pseudogenes, recent duplications, or have random similarity to a coding domain, which would be expected to occur in a set of random sequences of similar size. We therefore did not exclude these two transcripts from our analyses.

Mapping of genomic regions between species

To map genomic regions between species, we used pairwise alignments produced by the UCSC comparative genomics pipe-

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

line. In essence, it produces pairwise alignments between species using LASTZ (Harris 2007). In a process called "chaining" (Kent et al. 2003), alignment blocks from LASTZ are combined to longer consecutive aligned regions that allow for gaps in both species simultaneously. In a step called "netting," the best scoring chains are selected, and regions not covered by the highest scoring chain are filled by lower scoring chains in a hierarchical manner. We downloaded the final chain files that have undergone the netting step between human and all other species (hg19To*.over.chain) from UCSC. The chain file format lists all aligned blocks between two species. To map a genomic position from human to another species, we scanned the chain file and considered all aligned blocks overlapping the human region. If a region was covered by more than one chain, we chose the chain that had the highest coverage, i.e., the most bases aligned. To quantify the ambiguity caused by multiple chains that map to two or more different places in the other genome, we calculated the fraction of coverage of the longest chain of the total coverage by all chains. This fraction is 1 if there is only one chain and, for example, around 0.5 if a locus has two chains with similar coverage. We also tested if the mapping is reciprocal. To this end, we also downloaded the chain files with the nonhuman species as reference (*toHg19.over.chain). Using the same procedure as described before, we mapped the putative orthologous region back to human and tested if the mapped locus is identical to the original locus. This additional quality control of our mapping procedure showed that most of the mappings were unambiguous (i.e., a locus does not map to multiple nonsyntenic regions in the other species) and reciprocal (i.e., mapping back using the same procedure recovers the original locus) (Supplemental Fig. 2).

Expression data, read mapping, and transcript reconstruction

Summary statistics for all RNA-seq data used in this study is shown in Supplemental Table 2. The high coverage data was first described in (Merkin et al. 2012) and directly obtained from the authors. Data (Fastq files) from Brawand et al. (2011) were downloaded from GEO and aligned using TopHat, version 1.3.2 (Trapnell et al. 2009). First, reads were aligned to the genome using default parameters. Second, we used the EST library available for the genome (downloaded from the UCSC Genome Browser) (Karolchik et al. 2014), together with the junction file obtained in the first stage to realign the reads.

For exon predictions shown in Figure 4C and Supplemental Figure 8, we used Cufflinks (Trapnell et al. 2010) with default parameters. Transcript reconstructions were done for each tissue using the combined reads from all individuals.

Expression *P*-values, detection cutoffs, and parsimony analysis

To define cutoffs for the expression level of a putative ortholog, we calculated an empirical *P*-value based on the read count distribution of random genomic regions. The initial set of human lincRNAs was selected to have $P < 0.05$. Requiring the same significance level in other species would be too conservative because we could only detect RNAs that have the same or higher expression levels. We would miss orthologous lincRNAs with slightly lower expression due to natural variability in expression levels. We estimate from the variation of expression levels between individuals of the same species that we would misannotate as nonexpressed about 9%–15% of lincRNAs (Supplemental Fig. 6). Because it is reasonable to assume that expression level variation and associated loss in sensitivity is even higher for interspecies comparisons, we therefore set a less conservative cutoff of 0.1 at which we can reliably recover >95% between individuals of the same species. These *P*-values are used to define comparable and consistent cutoffs throughout the paper and are not corrected for multiple testing.

For the analysis shown in Figure 2, we also considered a lincRNA ortholog to be detected in a species if at least one splice site of the human transcript can be confirmed by spliced reads on the exact orthologous position in the other species (see below).

lincRNAs that were expressed in any given species according to the above criteria were assigned an "expressed" state. These states were then used to build a simple phylogenetic model, whose tree topology is shown in Figure 2C, where observed states were assigned to the tips of the tree. We assigned ancestral states to the internal nodes by considering the evolutionary scenario that required the fewest gain/loss events along the phylogeny and only allowing one gain event.

Splice sites

To assess conservation of actively used splice sites, we extracted all reads in windows of 50 nucleotides around all annotated splice sites in human and the orthologous sites in the other species. Read counts shown in Figure 4B,C were normalized between 0 and 1 in this window. Figure 4B shows the count of all reads, whereas in Figure 4C we only considered "split" reads that map to two different regions in the genome. We considered a splice site as detected in a species if the mean of the normalized split read count was higher in the exonic part of the window than in the intronic part. We used this simple metric because we found it gave essentially the same results as more complex statistical approaches to evaluate the difference in read density between exon and intron.

To compare exon boundaries and variation of exon length (Fig. 4A), we used the exons as predicted by Cufflinks (Trapnell et al. 2010). For each annotated exon in GENCODE, we tested whether it overlapped a predicted exon in the putatively orthologous region in the other species. If this was the case, we defined an anchor point that represents an orthologous position in human and the other species. We measured the distance from this anchor point to the exon end in both human and the other species and report the absolute value of their difference. If exon length is perfectly conserved, the difference is 0. We ignored all distances longer than 500 nt for the distribution in Figure 4A. If multiple exons were predicted, we took the minimal distance difference, i.e., we report the results for the best matching exon.

Tissue-specific expression

For the analysis shown in Figure 3, we started with average read count per (cross-species) mapped exonic base pair for each lincRNA in each of the nine tissues in each of the four species. We combined all reads from all three individuals. The raw read count was divided by the total number of reads in the respective libraries yielding a normalized expression value comparable to the commonly used FPKM value. Using the same method as described by Cabili et al. (2011), this expression vector was transformed to a normalized density vector with values between 0 and 1.

In addition, we calculated a single value for each lincRNA quantifying the tissue specificity. The tissue specificity score was introduced by Cabili et al. (2011) and is based on an entropy based measure that quantifies the distance of a given transcript's expression vector to a predefined expression vector that represents the extreme case of only being transcribed in one tissue. This value is calculated for each tissue and the tissue specificity score is the maximum value across all tissues (Cabili et al. 2011).

To calculate the tree shown in Figure 3B, we constructed a vector for each tissue in each species holding the normalized expression values for all lincRNAs. We then calculated a distance

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

matrix based on the Euclidian distance of these vectors and constructed a tree using the neighbor-joining algorithm.

To compare the similarity of expression levels of all tissues between species (Fig. 3C), we concatenated the vectors described before yielding one vector per species holding all normalized expression values for all lincRNAs and all tissues in the same order. We then calculated the Pearson correlation coefficient between these vectors. The analysis has been repeated using identical methods on a sample of 300 mRNAs that have found to be expressed in human, cow, mouse, and rat ($P < 0.1$). The complete list of lincRNAs and their expression properties in all species and tissues are available in the Supplemental Material and at http://garberlab.umassmed.edu/data/humanlincRNAEvol.

Sequence conservation

SiPhy (Garber et al. 2009) was run on the 46-way alignment available from UCSC (Karolchik et al. 2014) ignoring the following vertebrate genomes (danRer6, petMar1, oryLat2, gasAcu1, fr2, tetNig2) and using a window of 10 bases as previously described (Lindblad-Toh et al. 2011). We used the "omega" conservation values calculated by SiPhy throughout the paper. Data is available at http://www.broadinstitute.org/mammals/2x/ or upon request from the authors. To assess conservation level within APES, we used SiPhy on the 46-way alignment restricted to only human (hg19), chimpanzee (panTro2), gorilla (gorGor1), orangutan (ponAbe2), and rhesus (rheMac2), to score 20 base windows with a 15-base overlap across the exon of each transcript set: hominid-specific lincRNAs, mammalian-conserved lincRNAs, random set of 400 protein-coding genes, and sized-matched random noncoding genomic sequence. Each annotation was scored using the 0.75 percentile log-odds ratio score of all windows within the annotation. We then compared the distribution of these scores using a Mann-Whitney test.

Annotation enrichment analysis

We studied the enrichment of lincRNAs for common gene ontology terms using GREAT (McLean et al. 2010). Briefly, GREAT performs annotation enrichment analysis on noncoding genomic regions by analyzing the annotations of nearby genes. Noncoding regions (in our case, lincRNA loci) are assigned to putative target genes by association rules and, using gene annotations of the putative target genes, GREAT calculates statistical enrichment for associations between noncoding regions and annotations. For our analysis, we used human lincRNAs that were found in at least one other additional species. We defined tissue-specific lincRNAs as those that had relative RPKMs of at least 70% in a single tissue. A small number of lincRNAs are bidirectionally transcribed from the promoter of a coding gene. To prevent these lincRNAs from biasing our analysis toward enrichment of annotations of expressed coding genes, we removed any bidirectionally transcribed lincRNA within 500 bp of the TSS of a protein-coding gene. For each set of tissue-specific lincRNAs, we preformed GREAT analysis (version 2.0.2) using the "Basal plus extension" association rule and the entire genome as the background. The GREAT analysis can be found in the Supplemental Material.

Acknowledgments

We thank Jason Merkin for sharing prepublication data sets and for useful discussions. We thank Lucas D. Ward for lineage-specific constraint analysis and Jennifer Chen for gene ontology enrichment analysis and manuscript comments. We thank Mitch Guttman for manuscript comments and many discussions and Kristin Reiche for early discussions. This work was supported by the Austrian Science Fund (Erwin Schrödinger Fellowship J2966-B12 to S.W.); NIH U54-HG004555, NIH R01-HG004037, and NSF CAREER 0644282 to M.K.; and DARPA D12AP0004, NHGRI Center for Excellence in Genome Science 1P50HG006193, and the Broad Institute SPARC program to M.G.

References

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

Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008. The eukaryotic genome as an RNA machine. *Science* **319:** 1787–1789.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: A reference database for long noncoding RNAs. *Nucleic Acids Res* **39:** D146–D151.

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338:** 1587–1593.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478:** 343–348.

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.

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.

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.

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.

The FANTOM Consortium, Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559–1563.

Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD. 2006. The *Evf-2* noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* **20:** 1470–1484.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2013. Pfam: The protein families database. *Nucleic Acids Res* **42:** D222–D230.

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.

Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25:** i54–i62.

Green P, Ewing B. 2013. Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions." *Science* **340:** 682.

Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* **482:** 339–346.

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, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, et al. 2010. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28:** 503–510.

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.

Harris RS. 2007. "Improved pairwise alignment of genomic DNA." PhD thesis, The Pennsylvania State University.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42:** D764–D770.

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci* **100:** 11484–11489.
Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, Pääbo S. 2006. Functionality of intergenic transcription: An evolutionary comparison. *PLoS Genet* **2:** e171.
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.
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478:** 476–482.
Marques AC, Ponting CP. 2009. Catalogues of mammalian long noncoding RNAs: Modest conservation and incompleteness. *Genome Biol* **10:** R124.
McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501.
Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338:** 1593–1599.
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.
Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N. 1996. Requirement for *Xist* in X chromosome inactivation. *Nature* **379:** 131–137.
Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17:** 556–565.
Pruitt KD, Tatusova T, Brown GR, Maglott DR. 2012. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res* **40:** D130–D135.
Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* **81:** 145–166.
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.
Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312:** 1614–1620.
Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. 2011. genenames.org: The HGNC resources in 2011. *Nucleic Acids Res* **39:** D514–D519.
Trapnell C, Pachter L, Salzberg S. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.
Tsai M-C, 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.
Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72.
Wang Z, Burge CB. 2008. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. *RNA* **14:** 802–813.
Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337:** 1675–1678.
Ward LD, Kellis M. 2013. Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions." *Science* **340:** 682.
Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17:** 578–594.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst)* **15:** 496–503.
Young TL, Matsuda T, Cepko CL. 2005. The noncoding RNA *taurine upregulated gene 1* is required for differentiation of the murine retina. *Curr Biol* **15:** 501–512.

Received August 10, 2013; accepted in revised form January 14, 2014.

Downloaded from genome.cshlp.org on January 13, 2015 - Published by Cold Spring Harbor Laboratory Press



Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals

Stefan Washietl, Manolis Kellis and Manuel Garber

Genome Res. 2014 24: 616-628 originally published online January 15, 2014
Access the most recent version at doi:10.1101/gr.165035.113

Supplemental Material http://genome.cshlp.org/content/suppl/2014/02/10/gr.165035.113.DC1.html

References This article cites 46 articles, 24 of which can be accessed free at:
http://genome.cshlp.org/content/24/4/616.full.html#ref-list-1

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported), as described at http://creativecommons.org/licenses/by-nc/3.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2014 Washietl et al.; Published by Cold Spring Harbor Laboratory Press