

RESEARCH

Open Access



# Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs

Jenny Chen<sup>1,2</sup>, Alexander A. Shishkin<sup>3</sup>, Xiaopeng Zhu<sup>4</sup>, Sabah Kadri<sup>1</sup>, Itay Maza<sup>5</sup>, Mitchell Guttman<sup>3</sup>, Jacob H. Hanna<sup>5</sup>, Aviv Regev<sup>1,6</sup> and Manuel Garber<sup>4,7\*</sup>

## Abstract

**Background:** Recent advances in transcriptome sequencing have enabled the discovery of thousands of long non-coding RNAs (lncRNAs) across many species. Though several lncRNAs have been shown to play important roles in diverse biological processes, the functions and mechanisms of most lncRNAs remain unknown. Two significant obstacles lie between transcriptome sequencing and functional characterization of lncRNAs: identifying truly non-coding genes from *de novo* reconstructed transcriptomes, and prioritizing the hundreds of resulting putative lncRNAs for downstream experimental interrogation.

**Results:** We present *slnky*, a lncRNA discovery tool that produces a high-quality set of lncRNAs from RNA-sequencing data and further uses evolutionary constraint to prioritize lncRNAs that are likely to be functionally important. Our automated filtering pipeline is comparable to manual curation efforts and more sensitive than previously published computational approaches. Furthermore, we developed a sensitive alignment pipeline for aligning lncRNA loci and propose new evolutionary metrics relevant for analyzing sequence and transcript evolution. Our analysis reveals that evolutionary selection acts in several distinct patterns, and uncovers two notable classes of intergenic lncRNAs: one showing strong purifying selection on RNA sequence and another where constraint is restricted to the regulation but not the sequence of the transcript.

**Conclusion:** Our results highlight that lncRNAs are not a homogenous class of molecules but rather a mixture of multiple functional classes with distinct biological mechanism and/or roles. Our novel comparative methods for lncRNAs reveals 233 constrained lncRNAs out of tens of thousands of currently annotated transcripts, which we make available through the *slnky* Evolution Browser.

**Keywords:** Long non-coding RNAs, Evolution, Comparative genomics, Molecular evolution, Annotation, lncRNA, RNA-seq, Transcriptome

## Background

Recent advances in transcriptome sequencing have led to the discovery of thousands of long non-coding RNAs (lncRNAs), many of which have been shown to play important roles in diverse biological processes from development to immunity and their misregulation has been associated with numerous cancers [1–10]. Given the

importance of lncRNAs in biology and disease, there is great interest in defining lncRNAs in new experimental systems, disease models, and even primary cancer samples. Yet, despite important progress in RNA-Sequencing (RNA-Seq), the annotation and computational characterization of lncRNAs from RNA-Seq data remains a major challenge, with no easily accessible software available to accomplish either task.

We previously described a widely adopted computational framework for filtering lncRNAs from RNA-Seq transcript assemblies based on the presence of evolutionarily conserved protein-coding potential [11–14].

\* Correspondence: Manuel.Garber@umassmed.edu

<sup>4</sup>Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA

<sup>7</sup>Program in Molecular Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA

Full list of author information is available at the end of the article

Yet, this approach is limited in both sensitivity and specificity: (1) it incorrectly classifies *bona fide* lncRNAs as protein-coding simply because they are conserved; and (2) it incorrectly classifies transcripts as lncRNAs when they are actually extended untranslated regions (UTRs) of coding genes, pseudogenes, or members of lineage-specific protein-coding gene family expansions, such as zinc finger proteins or olfactory genes. Previous lncRNA cataloging efforts have addressed these issues by incorporating additional filtering criteria along with extensive manual curation to define meaningful lncRNA catalogs [12, 13, 15] or by including specialized libraries that better capture transcript boundaries [14, 16]. While these approaches have proven to be extremely valuable, they remain extremely labor-intensive and time-consuming, even for experienced users.

To address this challenge, we developed *slnky*, a method and accessible software package that enables robust and rapid identification of high-confidence lncRNA catalogs directly from RNA-Seq transcript assemblies without reliance on evolutionary measures of coding potential. *slnky* goes through several key steps to accurately separate lncRNAs from coding genes, pseudogenes, and assembly artifacts, while also identifying novel proteins including small peptides. This approach yields a high confidence lncRNA catalog. Indeed, when applied to mouse embryonic stem cells, *slnky* accurately identifies virtually all well-characterized lncRNAs and performs as well as previous manually curated catalogs.

Comparative analysis remains an important approach to assess potential function of a lncRNA without requiring additional experimental efforts. Despite its importance, identifying conservation of lncRNAs remains a challenge. To address this need, *slnky* incorporates a comparative analysis pipeline specially designed for the study of RNA evolution.

Here we demonstrate the utility of *slnky* by applying it to a comparative study of the embryonic stem (ES) cell transcriptome across human, mouse, rat, chimpanzee, and bonobo, and to previously defined datasets consisting of >700 RNA-Seq experiments across human and mouse. When applying *slnky* to these datasets, we discover hundreds of conserved lncRNAs. Furthermore, our metrics for evaluating transcript evolution show that there are clear evolutionary properties that divide lncRNAs into separate classes that display distinct patterns of selective pressure. In particular, we identify two notable classes of 'intergenic' ancestral lncRNAs ('lincRNAs'): one showing strong purifying selection on the RNA sequence and another showing only conservation of the act of transcription but with little conservation on the transcript produced. These results highlight that lncRNAs are not a homogenous class of molecules but are likely a

mixture of multiple functional classes that may reflect distinct biological mechanism and/or roles.

## Results and Discussion

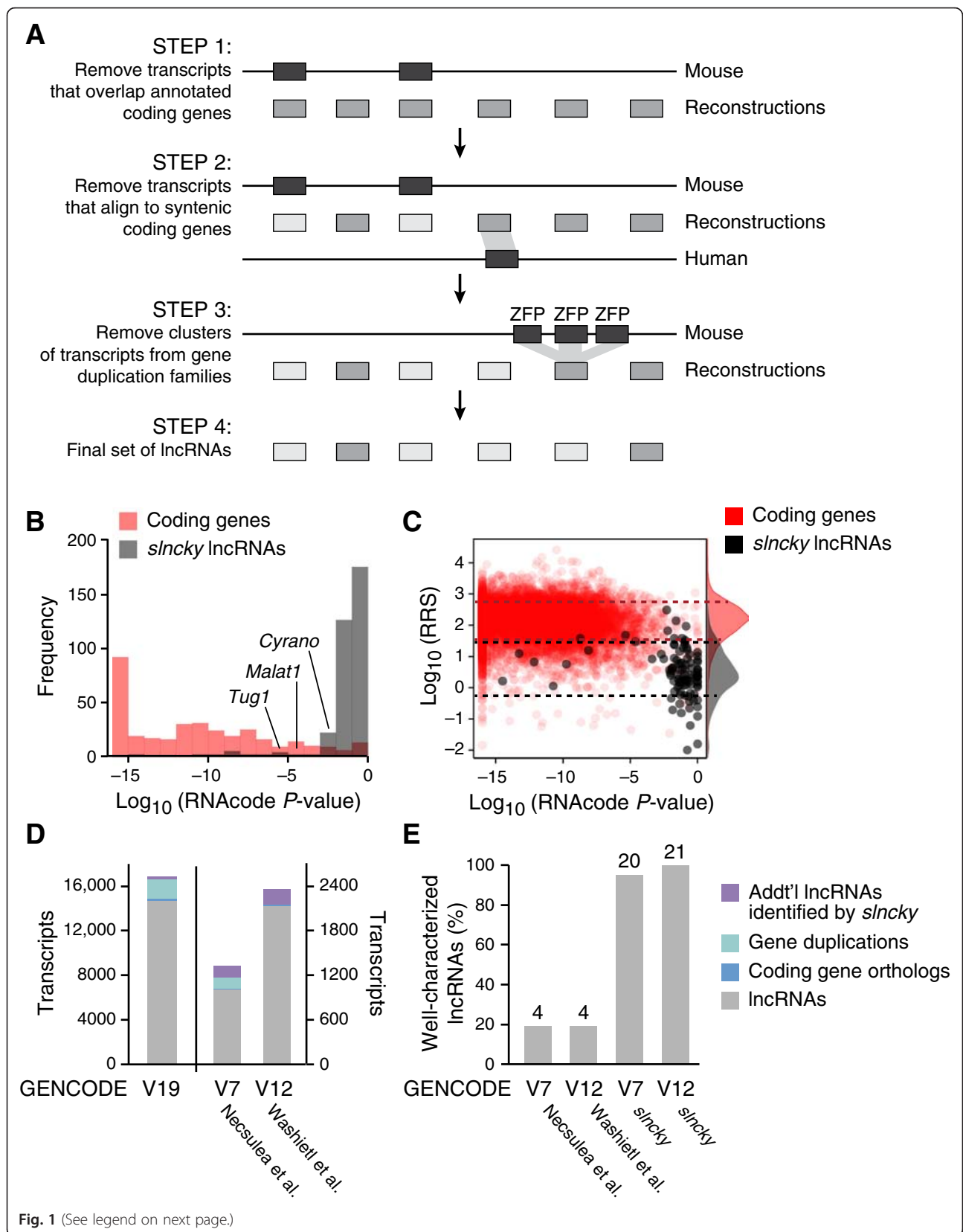
### *slnky* a software package to identify long non-coding RNAs

To develop a simple and accessible method to identify lncRNAs directly from RNA-Seq transcript assemblies, we created *slnky*, a method that enables rapid identification of high-confidence lncRNA catalogs directly from an RNA-Seq dataset.

Determining a set of lncRNAs from reconstructed annotations involves several steps to ensure that transcripts represent complete transcriptional units and that they are unlikely to encode for a protein. Current methods for defining coding potential rely on codon substitution models, such as PhyloCSF [17] and RNACode [18], which fail in three important cases: (1) they often incorrectly classify non-coding RNAs as protein-coding – including *TUG1*, *MALAT1*, and *XIST* – merely because they are conserved; (2) they fail to identify lineage specific proteins as coding; and (3) they erroneously identify non-coding elements (for example, UTR fragments, intronic reads) as lncRNAs. Rather than using codon substitution models, *slnky* implements a set of sensitive filtering steps to exclude fragment assemblies, UTR extensions, gene duplications, and pseudogenes, which are often mischaracterized as lncRNAs, while also avoiding the exclusion of *bona fide* lncRNA transcripts that are excluded simply because they have high evolutionary conservation.

To achieve this goal, *slnky* carries out the following steps (Fig. 1a): (1) *slnky* removes any transcript that overlaps (on the same strand) any portion of an annotated protein-coding gene in the same species; (2) *slnky* leverages the conservation of coding genes and uses annotations in related species to further exclude unannotated protein-coding genes, or incomplete transcripts that align to UTR sequences (Methods); and (3) to remove poorly annotated members of species-specific protein-coding gene expansions, *slnky* aligns all identified transcripts to each other and removes any transcript that shares significant homology with another non-coding transcript (Methods). The result is a filtered set of transcripts that retains conserved, non-coding transcripts that may score highly for coding potential, while excluding up to approximately 25 % of coding or pseudogenic transcripts normally identified as lncRNAs by traditional approaches.

After removing reconstructions that are likely gene fragments, pseudogenes, or members of gene family expansions, *slnky* searches for novel or previously unannotated coding genes, using a method that is less confounded by evolutionary conservation than codon substitution models. Specifically, *slnky* uses a sensitive



**Fig. 1** (See legend on next page.)

(See figure on previous page.)

**Fig. 1** *slncky* sensitively filters lncRNAs from reconstructed RNA-Seq data. **a** Schematic of *slncky*'s filtering pipeline. Annotated coding genes are shown in dark gray, reconstructed transcripts in medium gray, and filtered transcripts in light gray. **b** Histogram of  $\log_{10}(P)$  values of coding potential as evaluated by RNCCode (Washietl et al. [18]) for *slncky*-identified lncRNAs (gray) and coding genes (red). **c** Scatterplot of  $\log_{10}(P)$ -values of coding potential (x-axis) and  $\log_{10}(\text{ribosomal-release scores})$  (y-axis) of *slncky*-identified lncRNAs (gray) and coding genes (red). Distributions of ribosomal-release scores (RRS) are displayed along right side of y-axis. Dotted lines denote one standard deviation above and below the mean of RRS distributions. *slncky*-identified lncRNAs have significantly higher coding potential  $P$ -values and lower RRS than coding genes. **d** Comparison of previously published sets of lncRNAs to *slncky* results. Number of transcripts also annotated as a lncRNA by *slncky* (gray), number removed by *slncky* as gene duplication or coding (light and dark blue), and number of additional transcripts annotated as a lncRNA by *slncky* but not the previous pipeline (purple). **e** Percentage of well-characterized lncRNAs identified in previously published sets compared to *slncky* results. Numbers above bars denote absolute number of lncRNAs

alignment pipeline to find orthologous transcripts (Methods) and analyzes all possible open reading frames (ORFs) (that is, sequences containing both a start codon, a stop codon and containing at least 10 amino acids) that are present in both species. For each ORF, *slncky* computes the ratio of non-synonymous to synonymous mutations (dN/dS) and excludes all annotations with a significant dN/dS ratio (Methods). By requiring the presence of a conserved ORF that is transcribed in multiple species, and by computing the dN/dS ratio across the entire ORF alignment, *slncky* is more specific than conventional coding-potential scoring software, which report all high-scoring segments within an alignment.

Having developed a method to identify lncRNAs directly from RNA-Seq data, we sought to characterize its sensitivity and specificity by comparing lncRNAs identified by *slncky* to the well-studied set of lncRNAs expressed in mouse embryonic stem (ES) cells [11]. To do this, we generated RNA-Seq libraries from pluripotent cells obtained from three different mouse strains cultured using previously described growing conditions [19, 20] and used *de novo* reconstruction to build transcript models (Methods, Additional file 1: Table S1). We then applied *slncky* to define a set of 408 lncRNAs (Methods, Additional file 1: Figure S1). Our analysis also identified four transcripts – *Apela*, *Tunar*, *1500011K16Rik* (*LINC00116*), and *BC094334* (*LINC00094*) – that contain conserved ORFs with high coding potential (Additional file 1: Figure S2A and 2B).

Several lines of evidence indicate that our identified set represents *bona fide* lncRNAs: (1) *slncky* recovered all of the 20 functionally characterized lncRNAs that are expressed in the pluripotent state (Additional file 2), demonstrating that our stringent approach is still sensitive; (2) Our identified lncRNAs contain chromatin modifications of active RNA Polymerase II transcription (K4-K36), exhibiting similar levels as our previous ES catalogs (approximately 70 %) [11, 21]; (3) lncRNAs identified by *slncky* have significantly lower evolutionary coding potential scores than protein-coding genes ( $P = 1.3 \times 10^{-6}$ ,  $t$ -test) (Fig. 1b); (4) *slncky* does not filter out known conserved lncRNAs, such as *Malat1*, *Tug1*, *Miat*, that are often excluded due to significant coding-potential scores (Additional file 1: Figure S2C);

and (5) our set of lncRNAs have a significantly reduced ribosome release score (RRS) [22], a measure that accurately predicts coding potential from ribosome profiling data, than protein-coding genes (73-fold,  $P < 2.2 \times 10^{-16}$ ,  $t$ -test) (Fig. 1c).

Together, these results demonstrate that *slncky* provides a simple and robust strategy for identifying lncRNAs from a *de novo* transcriptome. Rather than requiring many user-defined parameters, *slncky* learns filtering parameters directly from the data making it useful across many different species, including non-model organisms (Methods).

#### ***slncky* provides greater sensitivity and specificity than previous lncRNA catalogs**

To verify the scalability and overall utility of *slncky* for defining lncRNAs across multiple datasets in different species, we ran *slncky* on GENCODE's latest comprehensive gene annotation set (V19) totaling 189,020 transcripts, of which 16,482 are annotated as lncRNAs that do not overlap a coding gene [15]. GENCODE is an ideal test case because it represents the current gold standard lncRNA-annotation set, primarily because much of its content undergoes extensive manual curation. Applying *slncky*, we identified 14,722 human lncRNA genes. Importantly, these include >90 % of the lncRNAs identified by GENCODE, with only 136 human (0.9 %) annotated protein coding gene, and 83 (0.6 %) annotated pseudogenes identified as lncRNAs. Transcripts that are annotated as lncRNAs by GENCODE but not by *slncky* include 1,735 (12 %) transcripts that are part of a cluster of duplicated genes, of which 123 (1 %) aligned to a known zinc finger protein or olfactory gene. An additional 181 (1 %) transcripts were excluded because they aligned significantly to an orthologous protein coding gene in mouse (Fig. 1d).

We then compared our filtering strategy with two previously published large-scale comparative studies that were based on GENCODE annotations [23, 24]. For the set of lncRNAs defined by Washietl et al. [24], *slncky* was able to remove 9.6 % (156) of the annotations that were likely a result of gene duplications and 1.2 % (19) that aligned significantly to a mouse coding transcript. In contrast, *slncky* only removed a handful of transcripts (<0.1 %) from the Necsulea et al. dataset [23]. Importantly,



*slnky* was much more sensitive as it identified virtually all well-characterized lncRNAs (20/21, Methods) compared to only 20 % (4/21) by these previous reports (Fig. 1e). Finally, we compared *slnky* to a recently published pipeline for filtering reconstructed transcripts from RNA-Seq data, called PLAR (Hezroni *et al.* [14]). We found that *slnky* and PLAR performed comparably in removing coding gene orthologs and gene duplications, but *slnky* remained more sensitive in recovering well-characterized transcripts (33/36 recovered by *slnky* compared to 27/36 by PLAR) (Additional file 1: Figure S3).

Together, our results highlight the power of *slnky* for identifying a high-confidence set of lncRNAs by excluding known artifacts that are often mistaken for lncRNAs. Furthermore, our results demonstrate that *slnky* performs as well as manual curation for defining *bona fide* lncRNAs and can even identify the challenging cases that are often missed by curation efforts.

#### ***slnky* enables detailed studies of lncRNA evolution**

Having developed a method to define high-quality lncRNAs, we sought to study the evolutionary properties of lncRNAs. While comparative genomics has provided important insights for studying proteins, enhancers, and promoters [25–30], relatively little has been done to study the evolution of lncRNAs. One of the main challenges is that lncRNAs diverge rapidly, accumulating both base nucleotide substitutions and insertion/deletion (indel) events. Both of these properties render lncRNAs difficult to align with conventional aligners and phylogenetic approaches.

To enable evolutionary analysis of lncRNAs, we implemented a computationally efficient and sensitive strategy to align lncRNAs and characterize their sequence and transcript evolution (Fig. 2a, Methods). To this end, *slnky* identifies the syntenic genomic region for a lncRNA in the orthologous species. If a transcript exists in a syntenic region, *slnky* aligns the two regions using a sensitive seed-based local pairwise aligner [31]. To avoid the possibility of spurious matches, *slnky* scores each alignment relative to a set of random intergenic regions from the orthologous genome and only keeps alignments that score higher than 95 % of the random intergenic sequences (Methods).

Next, *slnky* characterizes sequence and transcript conservation properties of orthologous lncRNAs. *slnky* calculates four metrics: (1) A ‘transcript-genome identity’ (TGI) score, defined as the percent of lncRNA base pairs that align and are identical to a syntenic genomic locus, to characterize how well the transcript sequence is conserved across the two species; (2) A ‘transcript-transcript identity’ (TTI) score, defined as the percent of identical, aligning base pairs found in the transcribed, exonic regions of both lncRNAs, to characterize how much of the transcript is transcribed in both species; (3) A ‘splice

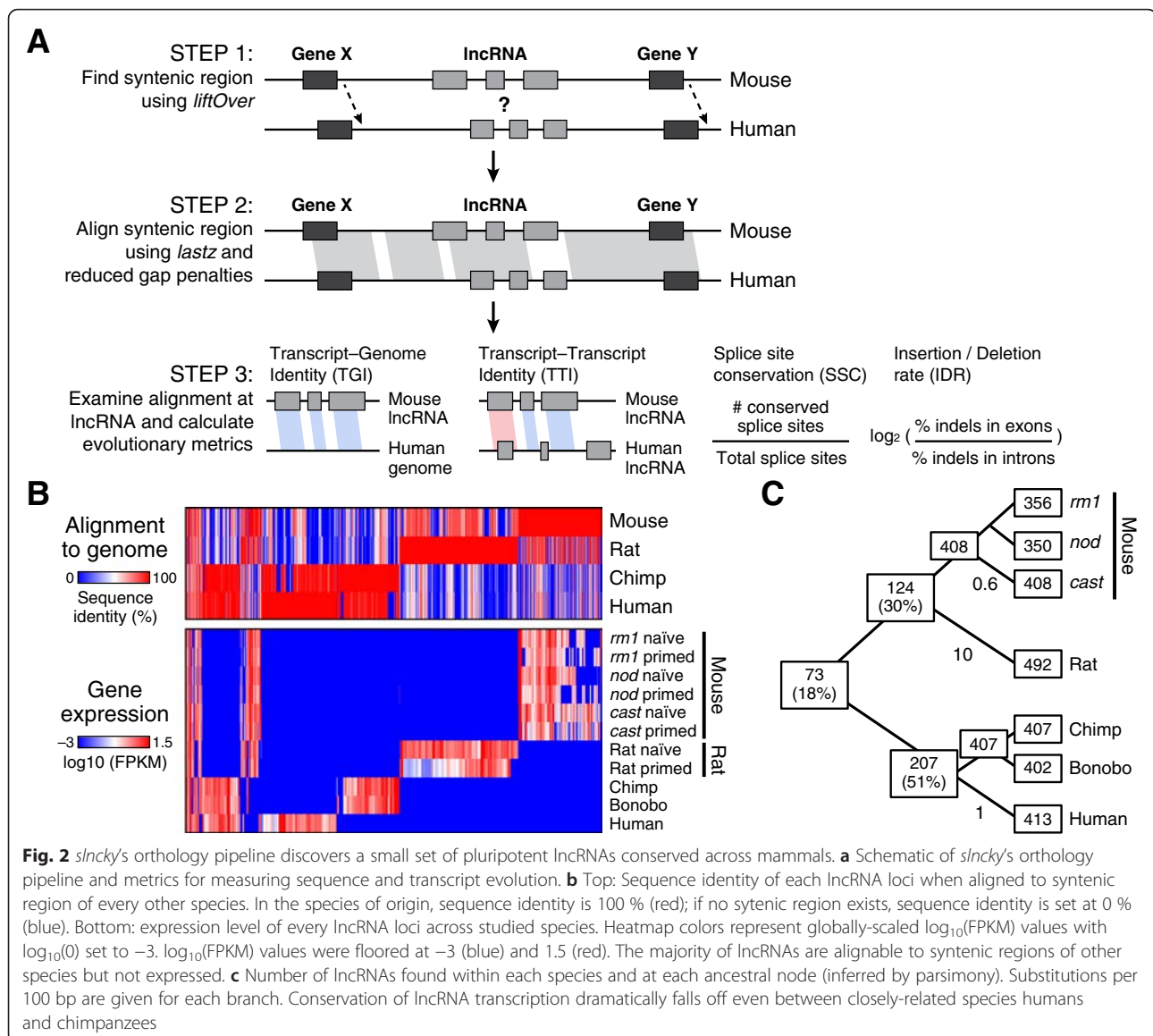
site conservation’ (SSC) score, defined as the percent of splice sites that are conserved across both lncRNAs, to characterize conservation of transcript structure; and (4) An ‘insertion/deletion rate’, defined as the  $\log_2$  rate of insertion/deletion events in exonic regions relative to intronic regions, to provide an alternative measure of sequence conservation (Fig. 2a).

We tested the performance of *slnky*’s orthology finding step by reanalyzing previous studies of lncRNA conservation across mammals [24] and vertebrates [14, 16, 23] (Methods). Our approach of aligning the two syntenic loci rather than just the transcripts increases *slnky* sensitivity with very little drop in specificity. In mammals, *slnky* successfully identified the vast majority (>95 %, 1,466/1,521 lncRNAs) of the previously reported orthologous lncRNAs while also finding an additional 121 pairs (8.0 %) of homologous human-mouse lncRNAs that were previously reported as species-specific (Methods). Similarly, in vertebrates, a four-fold greater evolutionary distance, *slnky* was able to recover 26 of 29 (90 %) of the previously defined ancestral lncRNAs; the alignments for the remaining three, although found, are indistinguishable from alignments that can be randomly found across syntenic loci and do not pass our significance threshold (Methods). Furthermore, *slnky* identified an additional three pairs of vertebrate conserved lncRNAs.

Together, these results demonstrate that *slnky* provides an efficient, sensitive, and accessible method for detecting and characterizing orthologous lncRNAs across any pair of species, providing an important tool for studying lncRNA evolution or for prioritizing lncRNAs based on evolutionary conservation.

#### **Evolutionary analysis reveals multiple lncRNA classes characterized by distinct signatures**

Initial work by us and others incorporating expression data across species showed that the expression of lncRNAs is often poorly conserved – with the rate of transcript expression loss occurring faster than loss of its genomic sequence identity across species [23, 24]. While these results provided important insights into the evolution of lncRNAs, these analyses did not fully explore the properties of the conserved lncRNAs. Having developed a method to comprehensively identify and align lncRNAs across species, we sought to further understand the evolutionary properties of lncRNAs. To do this, we generated RNA-Seq data from ES cells derived from three mouse strains (*I29SvEv*, *NOD*, and *castaneous*), rat, and human (Methods). We added additional published RNA-Seq data for chimpanzee and bonobo iPSC cells [32] (Additional file 1: Table S1). The gene expression between species shows a similarly high correlation to that previously observed for matched tissues across



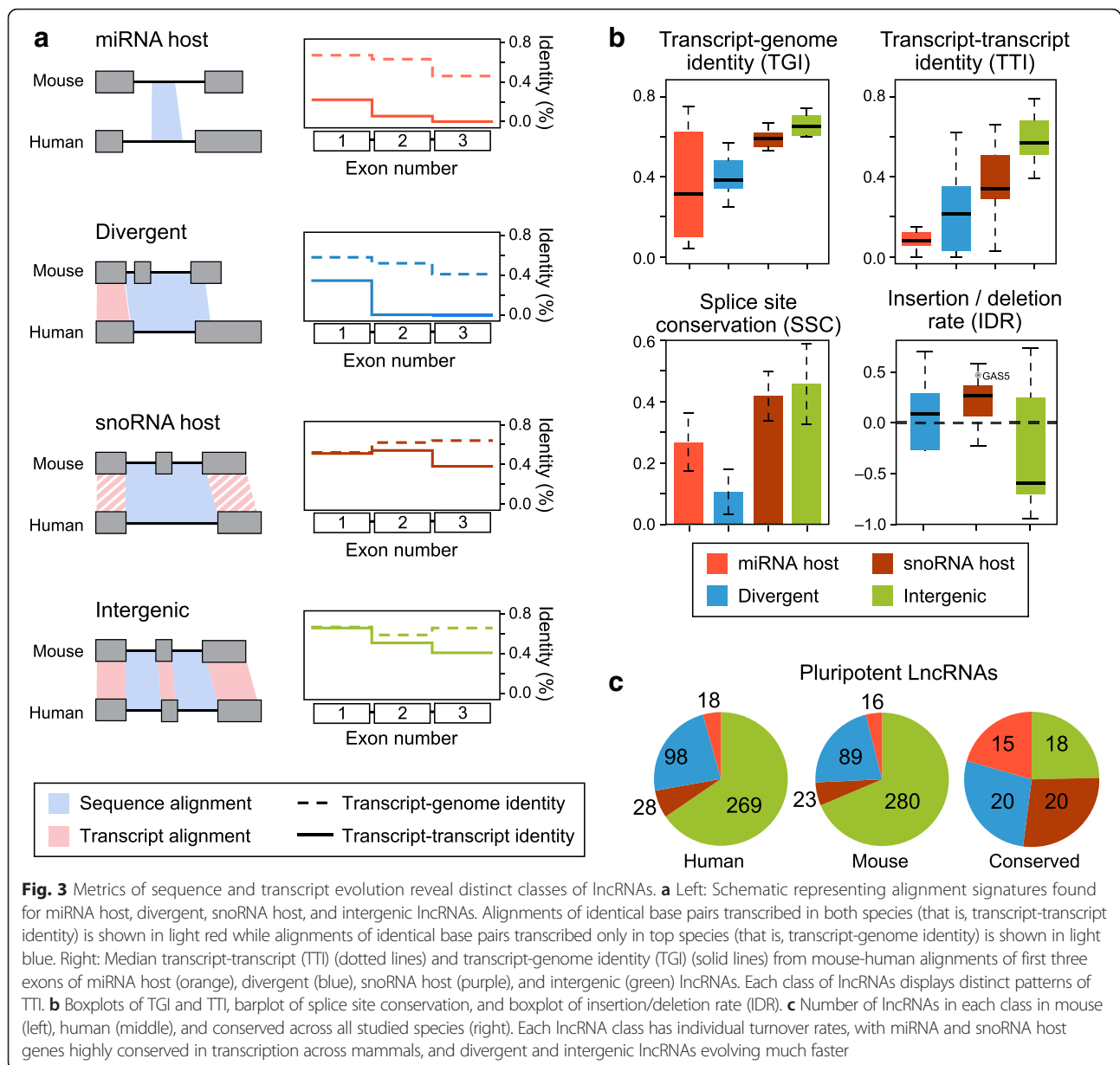
species (Additional file 1: Figure S4), highlighting the suitability of this set for comparative analysis.

Applying *slncky*, we identified 408 mouse, 492 rat, 407 chimpanzee, and 413 human lncRNAs (Additional file 1: Figure S1, Additional file 3). We found that lncRNAs are generally expressed only in a single species, despite the fact that most lncRNA loci can be aligned across species (Fig. 2b). In all, we found 73 (18 %) lncRNAs that are expressed in pluripotent cells across all mammals and are likely to be present prior to the divergence between rodents and primates (Fig. 2c, Additional file 4).

Like previous catalogs, our lncRNAs fall into different classes: miRNA host genes, snoRNA host genes, divergently expressed lncRNAs that are transcribed in the opposite orientation of a coding gene with which they share a promoter (Methods), and a remaining set of

'intergenic' lncRNAs (lincRNAs). Interestingly, we found that these classes have distinct patterns of sequence and transcript evolution.

These classes exhibit modest, but distinct, differences in transcript-genome identity (TGI), and striking differences in transcript-transcript identity (TTI) (Fig. 3a). While the loci of miRNA host genes can readily be aligned between species (that is, have similar TGI identity), their transcript structure have diverged tremendously, with 8.5 % median TTI across humans and mouse. lncRNAs divergently transcribed within 500 base pairs of a coding gene have also diverged rapidly in TTI, except for sequence transcribed near the promoter. For these genes, TTI is generally confined to the first exon. snoRNA host transcripts are very well conserved in both sequence and transcript structure, though we find an excess of indel



events in exons (1.2-fold more) as compared to introns (Fig. 3b). Finally, intergenic lncRNAs (lincRNAs) also have conserved transcript structure but a 1.5-fold reduction in exonic indel events compared to snoRNA hosts (Fig. 3b), despite comparable intronic indel rates (Additional file 1: Figure S5), suggesting that they undergo different selective pressure than host genes. Most of the pluripotent-expressed, well-characterized lncRNAs are found in this class of lincRNAs, which displays high TTI and splice site conservation (SSC). Two notable exceptions to the class of lincRNAs are *FIRRE* and *TSIX*, which have very poor TTI (5 % and 0.1 %, respectively). Both lincRNAs have been previously reported as ‘conserved in synteny’ only

[14, 33], possibly indicating that they may belong to a different class of lincRNAs. In addition to distinct differences in conservation of transcript structure, we found that the turnover of transcription differ across lncRNA classes: the majority of miRNA host and snoRNA host genes show conserved transcription across mammals (95 % and 87 %, respectively), whereas only a small percentage of divergent and intergenic genes show conserved transcription (22 % and 7 %, respectively, Fig. 3c).

We note that some lncRNAs have been proposed to have dual functions and our evolutionary metrics allow us to further explore this possibility. For example, *GASS* is a known snoRNA host gene and has also been

reported to function as a RNA gene [34]. Interestingly, we found that *GASS* has the typical signature of a snoRNA host, with higher indel rates at exons relative to its intronic regions (1.4-fold higher) (Fig. 3b, Additional file 4), suggesting that *GASS*, if truly functional as a non-coding gene, likely acts through a different mechanism than other intergenic lncRNAs.

We further note that these distinct signatures of evolution are robust enough to identify incorrectly annotated transcripts. For example, based on current annotations, *LINC-PINT* is an 'intergenic' lncRNA as the closest annotated coding gene, *MKLNI*, begins approximately 184 kb downstream [35]. However, its transcriptional conservation pattern is typical of a divergent transcript, with transcriptional identity confined only to its first exon. Closer inspection of expression data from our and other tissues [36] revealed that in fact, an unannotated, alternative transcriptional start site of *MKLNI* begins less than 200 base pairs downstream, consistent with *LINC-PINT*'s divergent alignment profile (Additional file 1: Figure S6).

We next sought to extend our evolutionary analysis to larger catalogs of mouse and human lncRNAs [15, 23, 24, 37]. Altogether, we searched for candidate orthologs across 251,786 human and 25,335 mouse transcripts corresponding to 56,280 and 15,508 unique lncRNA loci (Fig. 4a) using default parameters of *slacky*. miRNA hosts, divergent lncRNAs, and snoRNA host genes show the same distinct evolutionary patterns that we observed in pluripotent cells (Fig. 4b and c). Additionally, we found that miRNA hosts that harbor miRNAs inside exonic regions (for example, *HI9* [38]) show a distinct conservation pattern reminiscent of lncRNAs (high TTI and SSC), but without indel-constrained exons (Additional file 1: Figure S7), consistent with the functional importance of their exonic sequence.

In contrast to our previous analysis in matched pluripotent cells, we found that the majority of the 1,861 candidate orthologous intergenic lncRNAs identified from syntenic locations in human and mouse have low TTI (<30 %) and no conserved splice sites (approximately 61 %). Several lines of evidence suggest that the majority of these poorly aligning pairs may not be true orthologs but instead may be transcripts at syntenic loci in different cell types or transcriptional noise. First, applying our orthology-finding pipeline to randomly shuffled transcripts resulted in a similar proportion of syntenic transcripts with low TTI and zero conserved splice sites (Fig. 4d). Second, though poor alignment metrics could be the result of incomplete reconstructions of lowly expressed lncRNAs, when we performed a similar analysis on a FPKM-matched set of reconstructed coding transcripts, orthologous pairs have both high TTI and high SSC (Additional file 1: Figure S8A). Third, incorporating human and mouse expression data and limiting the

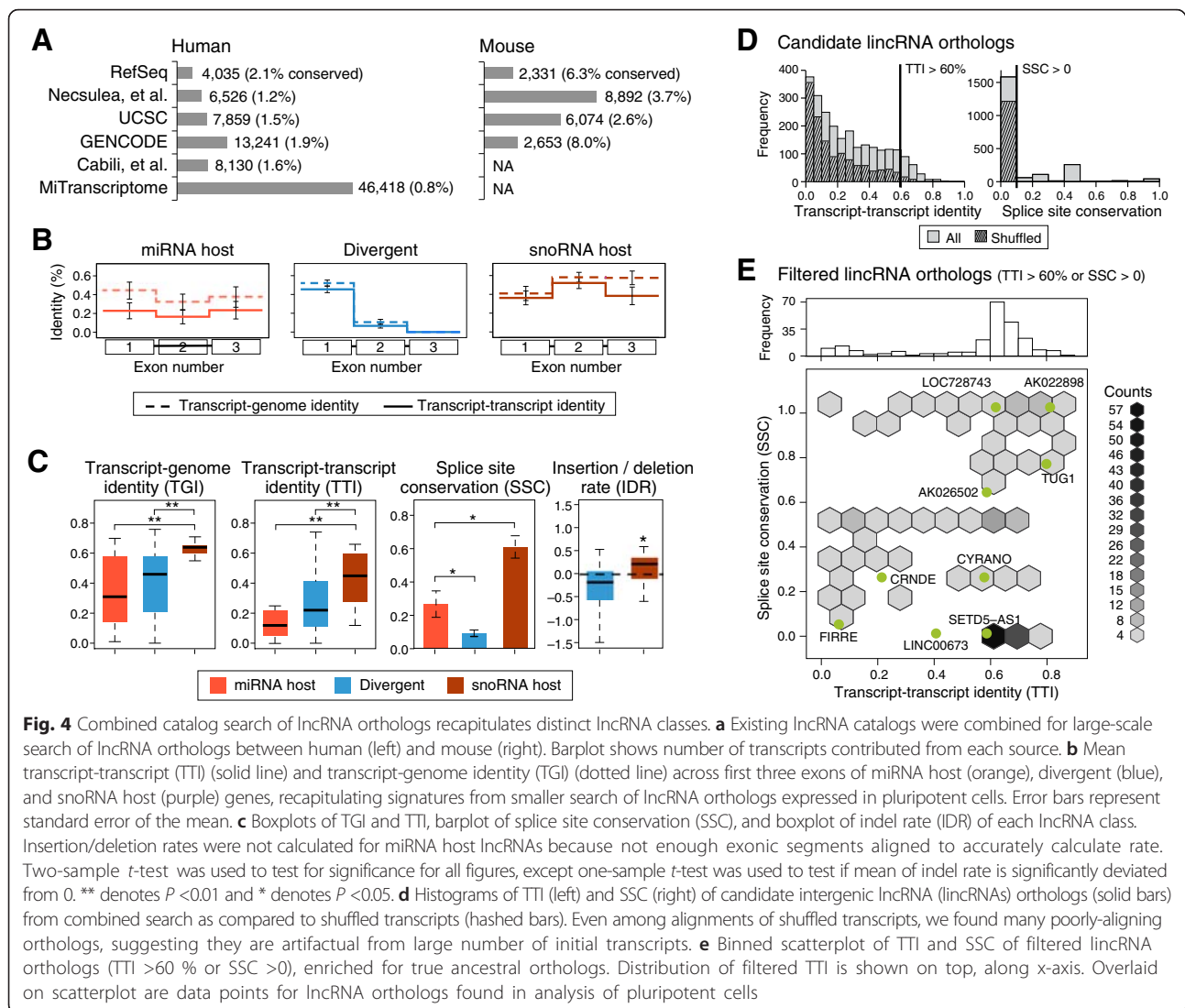
orthology search to only lncRNAs expressed in matched tissues drastically reduced the number of poorly aligning lncRNAs (Additional file 1: Figure S8B).

Taken together, we conclude that the majority of syntenic pairs we find are unrelated transcripts that have been annotated independently in human and mouse, perhaps in very different cell types, and which have no ancestral relationship. It is notable however that we found 39 pairs of human-mouse candidate orthologs that have low TTI, yet have at least one conserved splice site. This is surprising, because under the null hypothesis that these set of orthologs occupy a syntenic loci mostly by chance, we expect no pairs of orthologs to have an orthologous (conserved) donor/acceptor site (Methods). These 39 transcripts are reminiscent of lincRNA *FIRRE*, which has similarly low TTI but has one conserved splice site (out of 12). The fact that a set of lincRNAs are likely ancestral but with exonic sequence that has diverged rapidly points to a different class of lincRNAs with a very low purifying selective pressure on most of transcribed bases.

To investigate whether there are (at least) two distinct classes of lincRNAs, we first sought to reduce the number of possible spurious lincRNA orthologous pairs by either requiring transcript-transcript identity >60 %, which controls the false discovery rate at 10 % (Additional file 1: Figure S8C), or by requiring at least one conserved splice sites. We excluded the eight intergenic transcripts that contain a conserved ORF between human and mouse with a significant dN/dS ratio and significant coding potential score because they may encode for small proteins (Additional file 1: Table S2). Using these criteria, we found 232 pairs of human-mouse lincRNAs orthologs with a conservation profile similar to that found in the pluripotent analysis (Additional file 1: Figure S9), but with a bimodal TTI distribution (Fig. 4e). Modeling the TTI distribution as two Gaussians, we find 186 (80.1 %) lincRNAs with high TTI (mean 65.5 %  $\pm$  7.1 %) and 46 (19.8 %) with low TTI (mean 15.6 %  $\pm$  11.7 %). This further suggests that selection may operate in two distinct ways: for the majority of lincRNAs, it acts on the full RNA transcript, preserving the transcript sequence, while for a small subset of lincRNAs, the lincRNA sequence may be under positive selection, or perhaps only the act of transcription may be under selective constraint. With the goal of aiding in the study of these human-mouse conserved lincRNAs, we built an easily accessible application available at <https://scripts.mit.edu/~jjenny> as a resource for visually exploring the alignment and conservation properties of these lincRNAs.

Finally, we sought to understand properties of lincRNAs that explain their conservation or rapid turnover by investigating promoter conservation (Fig. 5). Within our pluripotent-expressed lincRNAs (Fig. 5a), we found that mammalian-conserved lincRNA promoters have



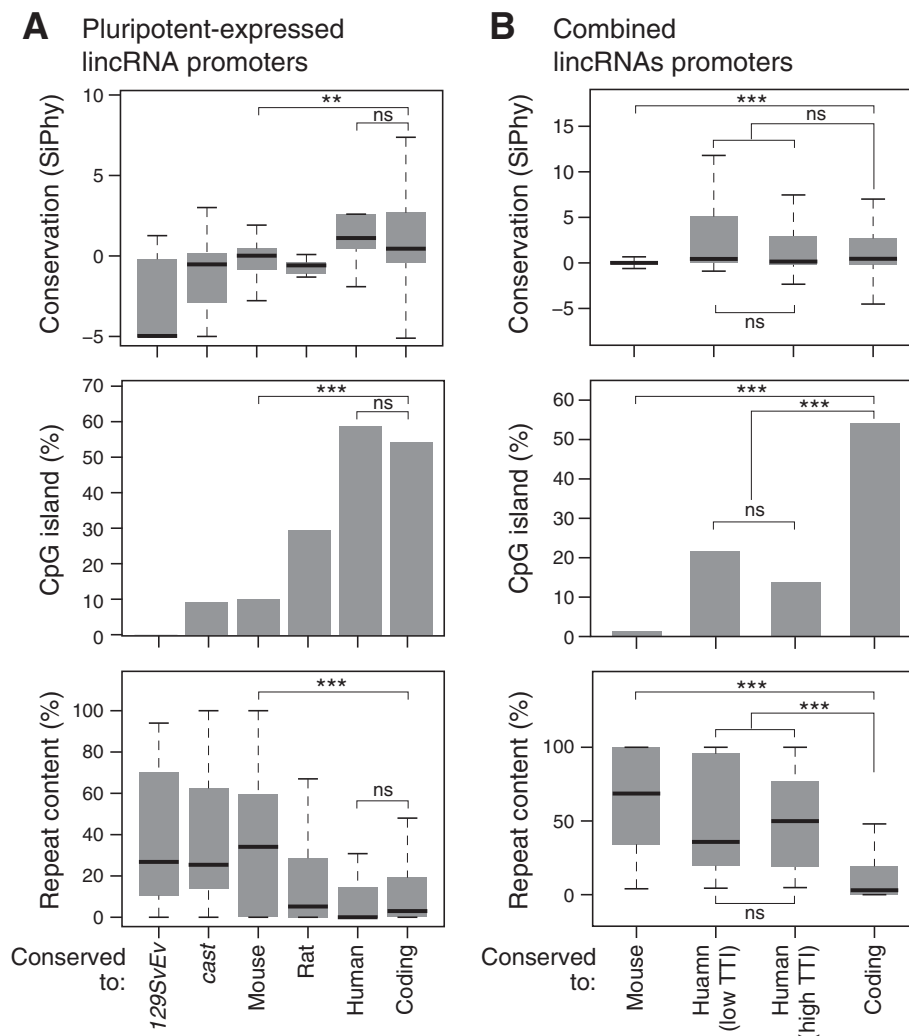


**Fig. 4** Combined catalog search of lincRNA orthologs recapitulates distinct lincRNA classes. **a** Existing lincRNA catalogs were combined for large-scale search of lincRNA orthologs between human (left) and mouse (right). Barplot shows number of transcripts contributed from each source. **b** Mean transcript-transcript (TTI) (solid line) and transcript-genome identity (TGI) (dotted line) across first three exons of miRNA host (orange), divergent (blue), and snoRNA host (purple) genes, recapitulating signatures from smaller search of lincRNA orthologs expressed in pluripotent cells. Error bars represent standard error of the mean. **c** Boxplots of TGI and TTI, barplot of splice site conservation (SSC), and boxplot of indel rate (IDR) of each lincRNA class. Insertion/deletion rates were not calculated for miRNA host lincRNAs because not enough exonic segments aligned to accurately calculate rate. Two-sample *t*-test was used to test for significance for all figures, except one-sample *t*-test was used to test if mean of indel rate is significantly deviated from 0. \*\* denotes  $P < 0.01$  and \* denotes  $P < 0.05$ . **d** Histograms of TTI (left) and SSC (right) of candidate intergenic lincRNA (lincRNAs) orthologs (solid bars) from combined search as compared to shuffled transcripts (hashed bars). Even among alignments of shuffled transcripts, we found many poorly-aligning orthologs, suggesting they are artifactual from large number of initial transcripts. **e** Binned scatterplot of TTI and SSC of filtered lincRNA orthologs (TTI > 60% or SSC > 0), enriched for true ancestral orthologs. Distribution of filtered TTI is shown on top, along x-axis. Overlaid on scatterplot are data points for lincRNA orthologs found in analysis of pluripotent cells

conservation scores comparable to protein coding genes, consistent with previous reports [11, 12], while species-specific lincRNA promoters are indistinguishable from neutral evolution of random intergenic genomic sequence. Conservation also extends to the promoter structure, as we found clear enrichment for CpG islands in conserved lincRNAs, despite comparable CG content (approximately 48 %) to that of species-specific lincRNA promoters, further suggesting strong selection on their transcriptional control. In contrast, we found that conservation is negatively correlated with repeat content in lincRNA promoters, and that a significant fraction (30.6 %,  $P = 1.65 \times 10^{-3}$ , Fisher's exact test) of species-specific lincRNA promoters contain species-specific endoretroviral K (ERV) repeat element that appear to be driving transcription. This repeat element is enriched only in promoters of lincRNAs expressed in pluripotent and testis cells (Additional file 1: Table S3), consistent with previous observations that

repeat elements are transcribed in ES and germline tissues and silenced in differentiated tissues. We observe that for 60.7 % of rodent-specific lincRNAs (that is, mouse or mouse and rat expressed lincRNAs), the time of ERVK integration on the evolutionary tree corresponds exactly with the evolutionary pattern of lincRNA transcription, providing strong evidence that the ERVK element is a primary driver for the origin of the lincRNA. We found corroborating trends of promoter conservation when examining the larger set of lincRNAs from our combined set of annotations (Fig. 5b). Importantly, we found no statistical difference in promoter conservation between high and low TTI lincRNA orthologs, suggesting selection for transcription even with poorly aligning orthologs.

Together, these results highlight the power of evolutionary analysis to identify distinct functional classes of lincRNAs and to reveal distinct features of these classes. In particular, we found 232 intergenic lincRNAs that



**Fig. 5** Conserved lincRNA promoters display strong selection for transcriptional control. **a** In each plot, each bar from left to right represents lincRNAs from pluripotent analysis that increase in conservation: *129SvEv*-specific, *cast*-specific, expressed across all mouse subspecies, expressed in mouse and rat, expressed in all mammalian species, and finally, expressed coding genes. Top: Promoter conservation in SiPhy scores (0 represents neutral evolution). Middle: Percent of promoters harboring CpG island. Bottom: Percent of promoter base pairs that belong to repeat element. **b** Same promoter metrics as A for mouse-specific and human-mouse conserved lincRNAs from combined lincRNA catalogs, and coding genes. Human-mouse conserved orthologs are split between those with low TTI and high TTI. \*\*\* denotes  $P < 0.001$ ; \*\* denotes  $P < 0.01$ ; \* denotes  $P < 0.05$  (t-test)

appear to be under selective constraint for and may play important roles in biology. We note that the majority of lincRNAs appear to be species-specific, raising questions about whether most of these transcripts are simply byproducts of transcription, with no important biological function. Alternatively, these lincRNA functions may be highly redundant or easily replaceable, in which case evolutionary turnover could be explained by a stochastic evolutionary process where redundant lincRNAs are fixed randomly along the evolutionary tree.

## Conclusion

While interest in lincRNAs has exploded, there is still relatively little known about the functions of lincRNAs

and much skepticism about what these large number of transcripts mean. The main challenge is that the number of functionally characterized lincRNAs remains a tiny fraction of the total number of lincRNAs that have been annotated. The significant effort required for functional characterization of a single lincRNA compared to its annotation has impeded the functional characterization of the large catalogs of lincRNAs. Accordingly, liberal cataloging efforts have led to a plethora of transcripts defined as lincRNAs that are rarely transcribed or artifacts of transcript assembly, thereby preventing experimental progress. *slmcky* provides an important and conservative approach for defining lincRNAs that enriches for *bona fide* lincRNAs. While *slmcky* will not necessarily capture every

single lncRNA nor will it provide the longest list of possible lncRNAs, it provides a method to define high confidence annotation of lncRNAs from any RNA-Seq dataset. This approach will enable meaningful experimental characterization of lncRNAs, making it easier to reconcile the large numbers of defined lncRNAs with the functional roles of these lncRNAs, and providing a consistent standard for evaluating *bona fide* lncRNAs.

Evolutionary conservation has long been a confusing feature of lncRNAs. While it is clear that lncRNAs are enriched for conserved sequences, their high levels of sequence divergence make them a challenge to study. While most lncRNAs do not appear to be conserved across mammals, it is currently unclear whether these lineage-specific lncRNA play important roles in lineage-specific biology. It is possible that many lncRNAs have 'functional orthologs': genes with similar function but no ancestral relationship. Importantly, evidence of functional orthology was recently reported for *XIST*. Although *XIST* is not found in marsupials, an opossum lncRNA called *RSX* was shown to have similar function. While *RSX* is capable of silencing the X chromosome in mouse, it shares no ancestral relationship with *XIST* [39]. We note that functional orthology cannot be studied with the methods presented here and future work will be needed to explore how many lncRNAs might play such lineage-specific roles or to what extent non-homologous lncRNAs carry similar function.

We demonstrated that lncRNAs can be categorized into distinct sets based on their evolutionary properties. Most notably, we found two sets of conserved intergenic lncRNAs: one that shows signs of purifying selection at the sequence level, and one that shows selection only for transcription. It will be fascinating to determine whether these two sets of lincRNA also correlate with functional differences. While we defined classes based on conservation, there are likely many other classes of lncRNAs that cannot be defined by conservation alone. We anticipate that as more cell types and tissues are explored, these annotation and evolutionary approaches will be even more valuable and enable more detailed studies of lncRNA biology.

## Methods

### *slncky*

#### **A stringent pipeline for filtering for lncRNAs**

*slncky* filters for lncRNAs in three simple steps. First, *slncky* filters out reconstructed transcripts that overlap coding genes or 'mapped-coding' genes on the same strand, in any amount.

After this step, *slncky* chooses a canonical isoform to represent overlapping transcripts. To do this, *slncky* clusters all transcripts with any amount of exonic overlap into one cluster, and chooses the longest transcript as the canonical isoform.

Next, *slncky* searches for gene duplication events (for example, zinc finger protein or olfactory gene expansions) by aligning each transcript to every other putative lncRNA transcript using *lastz* with default parameters [31]. *slncky* then aligns each transcript to shuffled intergenic regions to find a null distribution of alignment scores, repeating this procedure 200 times in order to estimate an empirical *P*-value. Any alignment with a *P*-value lower than 0.05 is considered significant. Sets of putative lncRNAs transcript that share significant homology are then merged, creating larger "duplication clusters". These transcripts do not necessarily share similarity to a protein-coding gene, though *slncky* will check and report homology to known ZFPs and olfactory genes. *slncky's* default parameters, which we used in all analyses reported (`--min_cluster_size 2`), notes and removes any duplication cluster containing two or more transcripts.

Finally, *slncky* removes any transcript that aligns to a syntenic coding gene in another species. (Human and mouse annotations are provided, though users can define their own). First, *slncky* learns a positive distribution by aligning all the transcripts removed in the first filtering step, which we know overlap coding genes, to their syntenic coding gene. *slncky* builds an empirical positive score distribution from these alignments. To align genes *slncky* first uses *liftOver* (`--minMatch = 0.1`) [40] to determine the syntenic loci in the comparing genome and *lastz* [31] to perform the alignment across the syntenic region. Using the empirical distribution, *slncky* learns an exonic identity threshold that has an empirical *P* value of 0.05. *slncky* repeats the alignment procedure on the putative lncRNAs to syntenic coding genes and filters out any transcripts that align at a higher score than this threshold, even if alignments occur only in UTR or intronic regions. In this way, *slncky* removes unannotated coding genes, pseudogenes, as well use UTR or intronic fragments from incomplete transcript assemblies. To reduce computational cost, whenever more than 250 coding-overlapping genes were filtered out from the first step, only a random subset of 250 transcripts is used to build the positive distribution.

#### **Flagging potentially coding 'lncRNAs'**

To find conserved lncRNAs that potentially harbor novel, unannotated protein, *slncky* aligns putative lncRNAs to syntenic non-coding transcripts in a comparing species, using a sensitive non-coding alignment strategy described below. *slncky* then crawls through each significant alignment and reports back any aligned open reading frame (ORF) longer than 30 base pairs. Only ORFs that do not contain a frame shift inducing indel in either species are reported. The start codon is defined as 'ATG' and stop codons are defined as 'TAA', 'TAG', or 'TGA'. *slncky* further calculates the ratio of non-synonymous to synonymous substitutions

(dN/dS ratio). We calculated an empirical  $P$  value for each dN/dS ratio by aligning 50,000 random intergenic regions and repeating the ORF finding procedure. Because the distribution of dN/dS ratio is dependent on ORF length (Additional file 1 1: Figure S2), we binned ORF lengths by 5 base pair windows and assigned an empirical  $P$  value if we had at least 100 random ORFs within that bin.  $P$  values were corrected for multiple hypothesis testing. For long ORFs, for which less than 100 length-matched random ORFs existed, we kept all alignments with dN/dS ratios  $<1$ .

#### **A sensitive method for aligning orthologous lncRNAs**

In searching for conserved lncRNA orthologs, *slncky* first defines the syntenic region of the comparing genome with *liftOver* ( $-\text{minMatch} = 0.1$   $-\text{multiple} = Y$ ) [40]. If a non-coding transcript exists in the syntenic region, *slncky* then aligns the area 150,000 base pairs upstream to 150,000 base pairs downstream of two syntenic regions. We choose 150,000 base pairs as a general heuristic that is likely to include an easily-alignable coding transcript up- and downstream of the lncRNA, which helps *lastz* to find a positively scoring alignment. Importantly, we also found that lncRNAs could only be aligned with a reduced gap-open penalty ( $-\text{gap} = 25,040$ ) because of many small insertions that appear to be well-tolerated by lncRNA transcripts.

To ensure we are not reporting alignments that may occur at random (driven mostly by repetitive elements), we align each lncRNA to shuffled intergenic regions to establish a null distribution and determine the empirical 5 % threshold for determining significant alignment scores. Because of our inclusion of flanking regions, it is possible to have a significant alignment in which only the flanking regions align but not the lncRNA transcripts. *slncky* reports these transcripts since it is possible that they are 'syntologs' and carry out orthologous functions but have evolved to a point where they no longer align.

#### **Data collection**

##### **Pluripotent cell lines and growth conditions**

Naïve 2i/LIF media for mouse and rat (rodent) naïve pluripotent cells was assembled as follows: 500 mL of N2B27 media was generated by including: 240 mL DMEM/F12 (Biological Industries – custom-made), 240 mL Neurobasal (Invitrogen; 21103), 5 mL N2 supplement (Invitrogen; 17502048), 5 mL B27 supplement (Invitrogen; 17504044), 1 mM glutamine (Invitrogen), 1 % non-essential amino acids (Invitrogen), 0.1 mM  $\beta$ -mercaptoethanol (Sigma), penicillin-streptomycin (Invitrogen), and 5 mg/mL BSA (Sigma). Naïve conditions for murine embryonic stem cells (ESCs) included 10  $\mu\text{g}$  recombinant human LIF (Peprotech) and small-molecule inhibitors CHIR99021 (CH, 1  $\mu\text{M}$ -Axon Medchem) and PD0325901 (PD, 0.75  $\mu\text{M}$  - TOCRIS) referred to as naïve 2i/LIF conditions. Naïve rodent cells were expanded on fibronectin coated plates (Sigma

Aldrich). Primed (EpiSC) N2B27 media for murine and rat cells (EpiSCs) contained 8 ng/mL recombinant human bFGF (Peprotech Asia), 20 ng/mL recombinant human Activin (Peprotech), and 1 % Knockout serum replacement (KSR- Invitrogen). Primed rodent cells were expanded on matrigel (BD Biosciences).

*129SvEv* (Taconic farms) male primed epiblast stem cell (EpiSC) line was derived from E6.5 embryos previously described in [41]. *129SvEv* naïve ESCs were derived from E3.5 blastocysts. *NOD* naïve ESC and primed EpiSC lines were previously embryo-derived generated and described in [42]. *castaneous* ESC line was derived from E3.5 in naïve 2i/LIF conditions and rendered into a primed cell line by passaging over eight times into primed conditions [43, 44]. Rat naïve iPSC lines were previously described in [44]. Briefly, rat tail tip derived fibroblasts were infected with a DOX inducible STEMCA-OKSM lentiviral reprogramming vector and M2rtTa lentivirus in 2i/LIF conditions. Established cell lines were maintained on irradiated MEF cells in 2i/LIF independent of DOX. Simultaneously, primed rat pluripotent cells were generated by transferring the rat naïve iPSC cells into primed EpiSC medium for more than eight passages before analysis was conducted. Naïve human C1 iPSC lines were derived and expanded on irradiated DR4 feeder cells as previously described [19].

#### **RNA-Sequencing**

RNA-Seq libraries were prepared as described in [45]. Briefly, 10  $\mu\text{g}$  of total RNA was polyA selected twice using Oligo(dT)25 beads (Life Technologies) and NEB oligo(dT) binding buffer. PolyA-selected RNA was fragmented, repaired, and cleaned using Zymo RNA concentrator-5 kit. A total of 30 ng of polyA-selected RNA per sample were used to make RNA-Seq libraries. An adapter was ligated to RNA, RNA was reverse transcribed, and a second adapter was ligated on cDNA. Illumina indexes were introduced during nine cycles of PCR using NEB Q5 Master Mix. Samples were sequenced 100-index-100 on HiSeq2500.

#### **Filtering**

##### **Filtering pluripotent lncRNAs from four mammalian species**

Transcripts were reconstructed from RNA-Sequencing data using Scripture (v3.1,  $-\text{coverage} = 0.2$ ) [11] and multi-exonic transcripts were filtered using *slncky* with default parameters. Annotations of coding genes were downloaded from UCSC ('coding' genes from track UCSC Genes, table kgTxInfo) [46] and RefSeq [47]. Mapped coding genes were downloaded from UCSC Transmap database (track UCSC Genes, table transMapAlnUcscGenes) [46]. For the mouse genome, we also included any blat-aligned human coding gene (track UCSC Genes, table blastHg18KG) [46]. As expected, the majority of reconstructed transcripts overlapped an



annotated coding or mapped coding gene at >95 % (Additional file 1: Figure S2). In the next step, *slnky* aligned each putative lncRNA to every other putative lncRNA to detect duplications of species-specific gene families. Across mouse, rat, and human transcriptomes, we found large clusters (15+ genes) of transcripts sharing significant sequence similarity with each other that also aligned to either zinc finger proteins or olfactory proteins. For unclear reasons, but likely due to the draft status of the assembly which results in collapsed repetitive sequence, we did not find any large clusters of duplicated genes in the chimpanzee genome, and instead found five small clusters of paralogs (Additional file 1: Figure S1).

Finally, *slnky* aligned the remaining transcripts to syntenic coding genes. For mouse and chimp transcripts, we aligned to syntenic human coding genes and for rat and human transcripts, we aligned to syntenic mouse coding genes. The learned transcript similarity threshold for each pair of comparing species varied as a function of distance between species: the empirical threshold for calling a significant human-chimp alignment was 29.8 % sequence similarity while for human-mouse alignments it was approximately 14 % (Additional file 1: Figure S1).

#### Single exon lncRNAs

Transcript reconstruction software tends to report thousands of single exon transcripts existing in a RNA-Seq library. Previous work suggests that the vast majority of these transcripts are results from incomplete UTR reconstruction, processed pseudogenes, very low expressed regions, and DNA contamination [14]). Although *slnky* filters a great number of these artifacts, we find that especially for single exon transcripts, many spurious reconstructions remain. For this reason, when analyzing single exon genes, we only focused on single-exon lncRNAs that are conserved across species.

#### Verification of filtered lncRNAs

We first verified *slnky's* lncRNA annotations by applying the filtering pipeline to our own generated RNA-Seq data and comparing the resulting lncRNA set with other computational and experimental methods, detailed below.

#### Chromatin modifications

Raw reads from ChIP-Sequencing experiments for H3K4me3 and H3K4me36 histone modifications in mouse embryonic stem cells (E14) were downloaded from [48] (GSE36114). Reads were mapped to mouse genome (mm9) using Bowtie (v0.12.7) [49] with default parameters. Peaks were called as previously described [50].

#### Coding potential

We scored coding potential of mouse lncRNAs using RNACODE (v0.3) [18] with default parameters using

multiple sequence alignments of 29 vertebrate genomes from the mouse perspective [29].

#### Ribosome release scores

Ribosome profiling data of mouse ES cells (E14) was downloaded from [51] (GSE30839). Ribosome release scores (RRS) were calculated as described in [22] using the RRS Program provided by the Guttman Lab.

#### Functionally characterized lncRNAs

To test the sensitivity of lncRNA filtering pipelines, we derived a list of well-characterized lncRNAs. To do this, we first took the intersection of annotated non-coding transcripts from UCSC [46], RefSeq [47], and GENCODE [52]. We then removed any lncRNA with a generically assigned name (for example, *LINC00028* or *LOC728716*) as well as generically named snoRNA and miRNA host genes (for example, *SNHG8* or *MIR4697HG*). Finally, we performed a literature search on the remaining lncRNAs, and kept only those that were specifically experimentally interrogated rather than reported from a large-scale screen. This list of well-characterized lncRNAs is available in Additional file 2.

#### Reanalysis of previously published lncRNA sets

We compared *slnky's* annotation of lncRNAs to three different human lncRNA sets: GENCODE V19 'Long non-coding RNA gene' set [52], a set reported by [23] based, in part, on GENCODE V7 annotations, and a set reported by [24] based on GENCODE V12 annotations. For all three comparisons, we first downloaded the appropriate version of GENCODE's 'Comprehensive gene' annotations and applied *slnky* using default parameters. For comparison to [23] and [24] we further scored expression of GENCODE annotations on the original RNA-Seq data used [53] using Cufflinks v2.1.1 [54] with default parameters and only compared robustly expressed (FPKM >10) lncRNAs.

#### Evolutionary study of lncRNAs

##### Reanalysis of previous studies of lncRNA conservation

We downloaded lncRNA annotations and ortholog tables derived from [23] and applied *slnky's* orthology pipeline to mouse and human lncRNAs using default parameters. We compared the human-mouse orthologs discovered by *slnky* to the list of transcripts that were defined by [23] to be ancestral to all Eutherians. We used downloaded FPKM tables to filter the additional orthologs discovered by *slnky* for pairs in which both transcripts are expressed in corresponding tissues.

To assess the ability of *slnky* to discover lncRNAs of a further evolutionary distance than mouse and human, we downloaded lncRNA and ortholog annotations from [16] and applied *slnky* using more relaxed parameters



(--minMatch 0.01, --pad 500000) to search for human-zebrafish and mouse-zebrafish lncRNA orthologs. Note that in both analyses, lncRNA annotations were not filtered by *slnky*'s filtering pipeline prior to the ortholog search so that our results could be directly comparable with the original publication.

#### **Annotating orthologous lncRNAs in pluripotent mammalian cells**

We applied *slnky* to our pluripotent RNA-Seq data to conduct an evolutionary analysis of lncRNAs across multiple mammalian species. We first searched for orthologous lncRNAs in a pairwise manner between every possible pair of species. Because the reconstruction software we used does not report lowly expressed transcripts that do not pass a significance threshold, and because we removed single-exons from our filtering step, we devised a method to rescue orthologous transcripts that may have been removed in those steps. For each lncRNA, if no orthologous lncRNA was detected by *slnky*, we went back to the original RNA-Seq data and forced reconstruction of lowly-expressed and/or single-exon transcripts in the syntenic region. We then re-aligned the lncRNA with these newly reconstructed transcripts and added the transcript to our lncRNA set when a significant alignment was found. We kept only pairs of conserved lncRNAs where a significant alignment was found in both reciprocal searches (for example, mouse-to-human and human-to-mouse).

Next, given pairs of lncRNA orthologs across all species, we created ortholog groups by greedily linking ortholog pairs. For example, given pairs {A,B} and {B,C}, we assigned {A,B,C} to one orthologous group, even if pairing {A,C} did not exist. Finally, we used Fitch's algorithm [55] to recursively reconstruct the most parsimonious presence/absence phylogenetic tree for each lncRNA and determine the last common ancestor (LCA) in which each lncRNA appeared. In the event a single LCA could not be determined by parsimony, we chose the most recent ancestor as the LCA in order to have conservative conservation estimates. For example, if a lncRNA was found in mouse and rat, but missing in human and chimp, we assigned the LCA to be at the rodent root, rather than at the mammalian root with a loss event at primates.

#### **Annotating matched low expression coding genes**

We tested our ability to detect conservation of lowly expressed transcripts by using our pipeline to reconstruct lowly-expressed coding genes known to be conserved across our tested species. We binned the set of intergenic lncRNAs by increments of 0.1 log<sub>10</sub>(FPKM), and sampled a set of 162 coding genes that matched in log<sub>10</sub>(FPKM) distribution in mouse ES cells. We then applied *slnky*'s orthology-finding module to the *de novo* reconstructions

of these coding genes from our generated RNA-Seq data. Repeating the same analysis as described above., we assigned the last common ancestor (LCA) of each coding gene. We were able to correctly assign the human-mouse ancestor as the LCA for 134 of 162 (83 %) coding genes, providing confidence that we are able to sensitively detect orthologs of lncRNAs, even though they are lowly expressed.

#### **Combined catalog analysis**

We downloaded human and mouse lncRNA annotations, where they existed, from RefSeq [47, 23], UCSC [46], GENCODE (v19 and vM1) [52, 12], and MiTranscriptome [36]. We filtered lncRNAs and searched for orthologs using *slnky* with default parameters. For overlapping isoforms that belong to the same gene, we chose one canonical ortholog pair that had the highest number of conserved splice sites and/or highest transcript-transcript identity. miRNA host and snoRNA host genes were annotated using Ensembl annotations of miRNAs and snoRNAs [56]. Divergent genes were annotated based on distance and orientation of closest UCSC or RefSeq-annotated coding gene. Orthologous lncRNAs were classified as a miRNA host, divergent, or snoRNA host if the transcript was annotated as such in both species. All other lncRNAs were classified as intergenic.

An orthology search was conducted on shuffled transcripts by collapsing overlapping isoforms to a canonical gene as described above, and shuffling to an intergenic location (that is, not overlapping an annotated coding gene) using *shuffleBed* utility [57]. We then carried out the orthology search and alignment exactly as described for lncRNAs. To empirically estimate the expected number of conserved splice sites across shuffled orthologs, we took each pair of true lncRNA orthologs and reshuffled splice sites within the loci such that it was correctly located at donor/acceptor sites (GT, AG), and re-evaluated number of conserved splice sites.

We used distributions resulting from our shuffled orthology search to filter and remove spurious hits from our set of candidate lincRNA orthologs. We then fitted two Gaussians to the resulting transcript-transcript identity using the *mixtools* package for R and default parameters [58]. Convergence was reached after 31 iterations of EM and final log-likelihood was 146.64. Each ortholog pair was assigned to a Gaussian based on posterior probability cutoff of 50 %.

#### **Promoter properties**

We defined promoters to be the 500 base pairs upstream of the lincRNA's transcription start site (TSS). We calculated several genomic properties of this region as follows:

### SiPhy

We calculated average SiPhy score across promoter region as previously described [59] using 29 mammals' alignment from mouse perspective [29].

### CpG islands

For the analysis of CpG islands, we used annotations provided by the UCSC Genome Browser (assembly mm9, track CpG Islands, table cpgIslandExt).

### Repeat elements

We intersected promoter regions with annotations from RepeatMasker [60] and calculated the number of base pairs of a lincRNA promoter belonging to a repeat element as well as percentage of lincRNA promoters harboring each class of repeat element. We then repeated this analysis with random intergenic regions, matched in size and GC content. To find statistically significant deviations in repeat content, we used Fisher's exact test to compare the proportion of species-specific lincRNA promoters containing each repeat element to the proportion of random, GC-matched intergenic regions containing the same element. We reported any repeat element that deviated from random, intergenic regions with a *P* value <0.005 (corrected for number of repeat types we tested).

### Data availability

- Raw and processed RNA-Seq data are available under GEO accession GSE64818: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE64818>
- A database of conserved lincRNAs discovered in this analysis is available at <https://scripts.mit.edu/~jjenny>

### Software availability

*slncky* (<http://slncky.github.io>) was developed in Python 2.0 and is freely available as source code distributed under the MIT License. *slncky* was tested on Linux and Mac OS X. The version used in this manuscript is available from DOI: 10.5281/zenodo.44628 (<https://zenodo.org/badge/latestdoi/19958/slncky/slncky>).

### Additional files

**Additional File 1: Supplementary figures and tables.** (PDF 3.85 MB)

**Additional file 2: Curated list of "well-characterized lincRNAs".** (XLSX 52 kb)

**Additional file 3: Bed file of lincRNAs discovered from mouse (mm9), human (hg19), chimp/bonobo (panTro4), and rat (rn5).** (XLSX 229 kb)

**Additional file 4: Excel file of evolutionary metrics of all lincRNAs found to be conserved to the human/chimp/rat/mouse ancestor.** (XLSX 19 kb)

### Abbreviations

ES: embryonic stem; ESC: embryonic stem cell; FPKM: fragments per kilobase of transcript per million reads mapped; ORF: open reading frame; RNA-Seq: RNA-Sequencing; RRS: ribosomal release score; lincRNA: long intergenic non-coding RNA; lincRNA: long non-coding RNA; SSC: splice site conservation; TGI: transcript-genome identity; TTI: transcript-transcript identity; UTR: untranslated region.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

JC participated in the design and coordination of the study, carried out all computational analysis and software development of *slncky* and *slncky Evolutionary Browser*, and wrote the manuscript. AS carried out RNA-Sequencing. XZ and SK participated in development of supporting software. IM and JH participated in deriving cell lines. M Guttman and AR participated in writing the manuscript. MG conceived of the study, participated in its design and coordination, and wrote the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

We thank Leslie Gaffney for artwork and advise on figures. JC was supported by an NHGRI training grant and by the Jan and Ruby Krouwer Fellowship Fund. MG was supported by DARPA grants D12AP00004 and D13AP00074. AR and MG were also supported by the CEGS 1P50HG006193. AR is supported by the Howard Hughes Medical Institute. JHH is supported by Ilana and Pascal Mantoux; the New York Stem Cell Foundation and is a New York Stem Cell Foundation - Robertson Investigator. We thank the Garber, Lander, and Regev laboratory members for helpful discussions.

### Author details

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. <sup>2</sup>Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA. <sup>3</sup>Division of Biology and Biological Engineering, California Institute of Technology, Cambridge, MA 02140, USA. <sup>4</sup>Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA. <sup>5</sup>Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. <sup>6</sup>Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02140, USA. <sup>7</sup>Program in Molecular Biology, University of Massachusetts Medical School, Worcester, MA 01655, USA.

Received: 26 October 2015 Accepted: 14 January 2016

### References

1. Ballabio A, Sebastio G, Carozzo R, Parenti G, Piccirillo A, Persico MG, et al. Deletions of the steroid sulphatase gene in "classical" X-linked ichthyosis and in X-linked ichthyosis associated with Kallmann syndrome. *Hum Genet.* 1987;77:338–41.
2. Greider CW, Blackburn EH. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. *Nature.* 1989;337:331–7.
3. Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, et al. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet.* 2010;42:1113–7.
4. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, et al. A long noncoding RNA mediates both activation and repression of immune response genes. *Science.* 2013;341:789–92.
5. Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science.* 2005;309:1570–3.
6. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature.* 2011;477:295–300.
7. Flockhart RJ, Webster DE, Qu K, Mascarenhas N, Kovalski J, Kretz M, et al. BRAFV600E remodels the melanocyte transcriptome and induces BANCR to regulate melanoma cell migration. *Genome Res.* 2012;22:1006–14.

8. Guan Y, Kuo W-L, Stilwell JL, Takano H, Lapuk AV, Fridlyand J, et al. Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin Cancer Res*. 2007;13:5745–55.
9. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol*. 2011;29:742–9.
10. Ellis BC, Molloy PL, Graham LD. CRNDE: A long non-coding RNA involved in Cancer, Neurobiology, and Development. *Front Genet*. 2012;3:270.
11. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol*. 2010;28:503–10.
12. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;25:1915–27.
13. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*. 2012;22:577–91.
14. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*. 2015;11:1110–22.
15. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res*. 2012;22:1775–89.
16. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*. 2011;147:1537–50.
17. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*. 2011;27:i275–82.
18. Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*. 2011;17:578–94.
19. Hanna J, Cheng AW, Saha K, Kim J, Lengner CJ, Soldner F, et al. Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs. *Proc Natl Acad Sci*. 2010;107:9222–7.
20. Gafni O, Weinberger L, Mansour AA, Manor YS, Chomsky E, Ben-Yosef D, et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature*. 2013;504:282–6.
21. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009;458:223–7.
22. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*. 2013;154:240–51.
23. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lincRNA repertoires and expression patterns in tetrapods. *Nature*. 2014;505(7485):635–40.
24. Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res*. 2014;24(4):616–28.
25. Bafna V, Huson DH. The conserved exon method for gene finding. *Proc Int Conf Intell Syst Mol Biol*. 2000;8:3–12.
26. Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*. 2000;10:950–8.
27. Korf I, Flicek P, Duan D, Brent MR. Integrating genomic homology into gene structure prediction. *Bioinformatics*. 2001;17 Suppl 1:S140–8.
28. Pachter L, Alexandersson M, Cawley S. Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J Comput Biol*. 2002;9:389–99.
29. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 2011;478:476–82.
30. Wenger AM, Clarke SL, Guturu H, Chen J, Schaar BT, McLean CY, et al. PRISM offers a comprehensive genomic approach to transcription factor function prediction. *Genome Res*. 2013;23:889–904.
31. Harris RS. Improved Pairwise Alignment of Genomic DNA. Ph.D. Thesis. The Pennsylvania State University; 2007. Retrieved from [http://www.bx.psu.edu/~rsharris/rsharris\\_phd\\_thesis\\_2007.pdf](http://www.bx.psu.edu/~rsharris/rsharris_phd_thesis_2007.pdf).
32. Marchetto MCN, Narvaiza I, Denli AM, Benner C, Lazzarini TA, Nathanson JL, et al. Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature*. 2013;503:525–9.
33. Hacisuleyman E, Goff LA, Trapnell C, Williams A, Henao-Mejia J, Sun L, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature Publishing Group*. 2014;21:198–206.
34. Smith CM, Steitz JA. Classification of gas5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol*. 1998;18:6897–909.
35. Sauvageau M, Goff LA, Lodato S, Bonev B, Groff AF, Gerhardinger C, et al. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife*. 2013;2, e01749.
36. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*. 2012;338:1593–99.
37. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*. 2015;47:199–208.
38. Brannan CI, Dees EC, Ingram RS, Tilghman SM. The product of the H19 gene may function as an RNA. *Mol Cell Biol*. 1990;10:28–36.
39. Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, et al. Rxs is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature*. 2012;487:254–8.
40. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*. 2006;34(Database issue):D590–8.
41. Tesar PJ, Chenoweth JG, Brook FA, Davies TJ, Evans EP, Mack DL, et al. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature*. 2007;448:196–9.
42. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, et al. Dissecting direct reprogramming through integrative genomic analysis. *Nature*. 2008;454:49–55.
43. Guo G, Yang J, Nichols J, Hall JS, Eyres I, Mansfield W, et al. Klf4 reverts developmentally programmed restriction of ground state pluripotency. *Development*. 2009;136:1063–9.
44. Hanna J, Markoulaki S, Mitalipova M, Cheng AW, Cassidy JP, Staerk J, et al. Metastable pluripotent states in NOD-mouse-derived ESCs. *Cell Stem Cell*. 2009;4:513–24.
45. Shishkin AA, Giannoukos G, Kucukural A, Ciulla D, Busby M, Surka C, et al. Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat Methods*. 2015;12:323–5.
46. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*. 2014;42(Database issue):D764–70.
47. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*. 2014;42(Database issue):D756–63.
48. Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, et al. Comparative epigenomic annotation of regulatory DNA. *Cell*. 2012;149:1381–92.
49. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10:R25.
50. Garber M, Yosef N, Goren A, Raychowdhury R, Thielke A, Guttman M, et al. A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol Cell*. 2012;47:810–22.
51. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011;147:789–802.
52. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*. 2012;22:1760–74.
53. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. *Nature*. 2011;478:343–8.
54. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012; 7:562–578.

55. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Syst Zool.* 1971;20:406–16.
56. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. *Nucleic Acids Res.* 2014;42(Database issue):D749–55.
57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–2.
58. Benaglia T, Chauveau D, Hunter D, Young D. mixtools: An r package for analyzing finite mixture models. *J Stat Softw.* 2009;32:1–29.
59. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics.* 2009;25:i54–62.
60. Smit AFA, Hubley R, Green P: RepeatMasker. Available at: <http://www.repeatmasker.org>. [Accessed 9 April 2013].

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

