SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

August 2015 Vol.58 No.8: 787–798
doi: 10.1007/s11427-015-4881-9

Evolutionary annotation of conserved long non-coding RNAs in major mammalian species

BU DeChao[1,2†], LUO HaiTao[1,2†], JIAO Fei[3], FANG ShuangSang[1,2], TAN ChengFu[1,2], LIU ZhiYong[1] & ZHAO Yi[1*]

[1]*Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China;*
[3]*Department of Biochemistry and Molecular Biology, Binzhou Medical College, Yantai 264003, China*

Received October 8, 2014; accepted April 17, 2015; published online June 26, 2015

Mammalian genomes contain tens of thousands of long non-coding RNAs (lncRNAs) that have been implicated in diverse biological processes. However, the lncRNA transcriptomes of most mammalian species have not been established, limiting the evolutionary annotation of these novel transcripts. Based on RNA sequencing data from six tissues of nine species, we built comprehensive lncRNA catalogs (4,142–42,558 lncRNAs) covering the major mammalian species. Compared to protein-coding RNAs, expression of lncRNAs exhibits striking lineage specificity. Notably, although 30%–99% human lncRNAs are conserved across different species on DNA locus level, only 20%–27% of these conserved lncRNA loci are detected to transcription, which represents a stark contrast to the proportion of conserved protein-coding genes (48%–80%). This finding provides a valuable resource for experimental scientists to study the mechanisms of lncRNAs. Moreover, we constructed lncRNA expression phylogenetic trees across nine mammals and demonstrated that lncRNA expression profiles can reliably determine phylogenic placement in a manner similar to their coding counterparts. Our data also reveal that the evolutionary rate of lncRNA expression varies among tissues and is significantly higher than those for protein-coding genes. To streamline the processes of browsing lncRNAs and detecting their evolutionary statuses, we integrate all the data produced in this study into a database named PhyloNONCODE (http://www.bioinfo.org/phyloNoncode). Our work starts to place mammalian lncRNAs in an evolutionary context and represent a rich resource for comparative and functional analyses of this critical layer of genome.

lncRNA, conservation, evolution

Citation: Bu DC, Luo HT, Jiao F, Fang SS, Tan CF, Liu ZY, Zhao Y. Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Sci China Life Sci, 2015, 58: 787−798, doi: 10.1007/s11427-015-4881-9

A large proportion of functional sequence within mammalian genomes can be transcribed into long non-coding RNAs (lncRNAs), which have a length ranging from 200 nt to 100 kb and do not show any evidence of being translated to protein [1]. LncRNAs have been implicated in a multitude of biological processes such as transcriptional regulation, cell growth, differentiation and senescence [2–5] and associated with some diseases [6–10]. The expression levels of lncRNA are in average lower than those for protein-coding genes [1,11] and they often exhibit stronger tissue specificities than coding transcripts [11]. Compared with protein-coding genes and small RNAs (e.g., miRNA and snoRNA), most of lncRNAs are less conserved in sequence [12]. Despite limited overall conservation, many lncRNAs contain local regions that are preserved across multiple spe-

†Contributed equally to this work
*Corresponding author (email: biozy@ict.ac.cn)

cies attesting to their functional potential [13–16]. In general, genomic sequences of lncRNAs show reduced substitution and insertion/deletion rates compared with expected random rates [12,17]. In addition, lncRNA transcripts also exhibit clear tissue-specific expression and show lower mutation rate indicating that they are subject to considerable purifying selection. Rapid transcriptional turnover of lncRNAs is found to affect their lineage-specific emergence or disappearance [18], and the lower expression level of lncRNAs may be associated with their faster evolutionary rate [19]. These findings suggest that variations in lncRNA expression levels might contribute to phenotypic differences between species and in some occasions might be equally critical as protein-coding genes in species or lineage determination. However, questions regarding the overall conservation of lncRNAs in comparison to those of coding genes in major mammalian lineages are still unanswered and the key evolutionary characteristics: tempo and mode of lncRNA transcriptomes have not been fully assessed. These questions remain partly due to the lack of comprehensive lncRNA catalogs of major mammalian lineages and thorough analyses of such datasets from an evolutionary perspective.

In this report, we applied a computational approach to the RNA-Seq data of polyadenylated RNA [20] from six tissues (brain after removing cerebellum or cerebrum brain, cerebellum, heart, kidney, liver, testis) across nine species (human, chimpanzee, gorilla, orangutan, rhesus macaque, mouse, opossum, platypus, and chicken) that represented three major mammalian lineages (placentals, marsupials and monotremes) and birds (the evolutionary outgroup), and identified all transcripts that are expressed and have a negligible potential to encode proteins thus classified as lncRNAs. In total, between 4,141–13,709 lncRNAs were identified from the RNA-Seq data in nine species. By combining Ensembl database [21] and other well-known lncRNA databases [22–25], we compiled comprehensive lncRNA catalogues of mammals (4,141–42,558 lncRNAs). Consistent with previous studies, the expression of lncRNAs tends to be lower and more tissue-specific compared with protein-coding genes [11,22,26]. To evaluate the overall conservation of lncRNAs in mammalian lineages, we performed a comprehensive whole-genome conservation and transcription analysis on both lncRNAs and coding transcripts. The results demonstrated that although 30%–99% human lncRNAs are conserved across different species on DNA locus level, a much smaller portion of lncRNA genomic loci are transcribed. We subsequently built lncRNA expression phylogenetic trees across nine species and showed that lncRNA expression phylogenies for most tissues are in line with the known mammalian phylogeny. Evolutionary patterns of lncRNA expression across all other tissues closely resemble the previous published results derived from protein-coding genes [20]. The evolutionary rate of lncRNA expression varies among tissues and is significantly higher than those for protein-coding genes, which suggests that lncRNAs may have experienced stronger positive selection.

Since the comprehensive lncRNA data collection with conservation and evolution annotation generated by our study represents an informative resource for the scientific community, we deposit our data and analyses into a new database we named PhyloNONCODE—a resourceful database of lncRNAs covering their genomic and transcriptional conservation in major species based on molecular evolutionary analysis. PhyloNONCODE allows users to browse and search genomic and transcriptomic conservation status of lncRNA across multiple species, and also provide intuitive visualization tool presenting lncRNA information in the widely used UCSC Genome Browser format.

1 Materials and methods

1.1 Data collection

To obtain a comparative transcriptome for mammals, we performed in-depth analyses of multi-species RNA-Seq data [20] (under accession code GSE30352) from the Gene Expression Omnibus. Those data contains 2.9 billion Illumina Genome Analyser IIx reads of 76 base pairs for the polyadenylated RNA fraction of brain, cerebellum, heart, kidney, liver and testis collected from eight mammalian species. The species were as follows: placental mammals (humans, chimpanzee, gorilla, orangutan, rhesus macaque, and mouse), marsupials (opossum), and monotremes (platypus). Corresponding data were collected for a bird (chicken), as an evolutionary outgroup. RNA-Seq data of five rat tissues (brain, heart, kidney, liver and testis) are used as validation dataset and applied the same process line. These RNA-Seq data are retrieved from GEO dataset with accession code GSE41637 [27].

The Reference Genome in this work was as follows: *hg19* for human, *pantro3* for chimpanzee, *gorGor3* for gorilla, *ponAbe2* for orangutan, *rheMac3* for macaque, *mm9* for mouse, *rn4* for rat, *monDom5* for opossum, *ornAna1* for platypus, and *galGal3* for chicken, which were downloaded from the assemblies featured in the UCSC Genome Browser [28]. Corresponding gene annotations for these assemblies were also downloaded from the UCSC Genome Browser.

1.2 LncRNA identification

Based on these RNA-Seq datasets, short-reads were mapped to the genome of corresponding species using the spliced read aligner Tophat (version V1.4.1) [29] with the following parameters: min-anchor=5, min-isoform-fraction=0, and the rest set as default. Mapped reads of biological replicates from the same tissue were merged into a single BAM file in order to facilitate assembly and quantification of these data. Subsequently *ab initio* assemble software Cufflinks (version

V1.3.0) [30] was used to reconstruct transcriptomes for each tissue based on the reads with default parameters.

In order to obtain comprehensive lncRNA catalogs for each species, we developed a computational approach that consisted of the following five steps: (i) For each species, all transcripts constructed from the tissues were combined into an initial catalog of whole transcriptome using cuffcompare program. (ii) The combined transcripts were compared with Ensembl genes to eliminate known protein-coding genes, pseudogenes, microRNA, tRNAs, snoRNAs, rRNAs, snRNAs, scoRNAs; The pseudogenes here we use are those called "polymorphic pseudogene", which are annotated with an Ensembl protein ID in the Ensembl database. We excluded this class of pseudogenes to avoid the likelihood that they may have coding potential in the cell lines and tissues studied by other ENCODE groups. (iii) The coding potential of each transcript using the Coding Noncoding Index (CNCI) software was calculated to recover the transcripts which could be categorized as non-coding [31]. CNCI is a powerful tool to effectively distinguish protein-coding and non-coding sequences independent of known annotations. CNCI software is available at http://www.bioinfo.org/software/cnci. (iv) The data of Brawand et al. [20] is not strand-specific. It is the case for the single-exon transcripts that their direction cannot be decided during transcripts-building through cufflinks. Transcripts which were single exon or less than 200 nt were then discarded in our approach. (v) The transcripts in intergenic, intronic and antisense regions of protein-coding genes were retained, which are annotated with class code "i", "u" or "x" by cuffcompare. These remaining transcripts were combined with known Ensembl annotations (GENCODE [32], NONCODE [23–25] and Human Body Map lncRNAs [11] were included for human beings; NONCODE were also included for mouse). These five steps yielded a total of 4,141–42,558 lncRNAs, from nine species.

1.3 Expression and tissue specificity analysis

The fragments per kilobase of exons per million fragments mapped (FPKM) expression values of lncRNAs and protein-coding genes were calculated by Cufflinks [30]. To investigate the tissue specificity of lncRNA expression, each transcript was assigned a tissue specificity score derived from an entropy-based metric that relies on Jensen-Shannon (JS) divergence [11]. This specificity metric (ranging from 0 to 1) quantifies the similarity of a transcript's expression pattern across tissues, together with a predefined pattern that represents the extreme case in which a transcript is expressed only in one tissue. Thus, a perfect tissue-specific pattern would receive a JS score of 1.

1.4 Conservation analysis based on pair-wise conserved counterparts

To assess gene conservation, we made a comparison between all genes of one species against all the other species in our sample to find their conserved counterparts using UCSC LiftOver tool [28,33]. In brief, the LiftOver utilized BLASTZ [34], an independent implementation of the Gapped BLAST algorithm specifically designed for aligning two long genomic sequences, as a core algorithm to detect homologous regions in other genomes. It firstly split the original genome into smaller fragments and then aligned the abbreviated genomes to the target genome using syntenic BLASTZ alignments. Taking into account the relative low resolution of LiftOver, we focused the conservation of lncRNAs on genomic loci rather than transcript structure. It is worth noting that only if a gene has a homology, can its genomic location in another species be elucidated. The gene was considered as a lineage-conserved gene if it had conserved counterparts across all species of the samples in all putative lineages.

We defined Coverage Pattern of Conserved Counterparts (CPCC) score to measure the conservation degree of lncRNAs. Given the following statement, Gene A in X species has conserved counterpart B in Y species, the conserved counterpart of B in X species is C. If A has no overlap with C, its CPCC score is 0, and A is considered a unilateral conserved lncRNA; if A partially overlaps with C, then CPCC score of A is the ratio of the overlap part between A and C to A, and A is called as bilateral and partial conserved lncRNAs. if A completely overlaps with C, then CPCC score of A is 1, and A is called as bilateral and complete conserved lncRNAs.

Not all genomic-conserved counterparts of genes are transcribed in other species. For genes in one species, their conserved counterparts were defined as expressed in other species when there is at least one transcript completely or partially (>80%) located in the conserved counterpart region. The expression abundance of conserved counterparts was estimated by summing the expression value of all the transcripts located therein, which is based on the gene loci quantitative strategies of FPKM via Cufflinks as mentioned above [35].

1.5 Construction of mammalian gene expression phylogenetic trees

Both lncRNA and protein-coding gene expression trees were constructed using the neighbor joining approach based on pair-wise distance matrices between species [36,37]. The distance between samples (as a measure of divergence) was computed from the gene expression profiles as $1-\rho$, where ρ is Spearman's correlation coefficient. Spearman's correlation coefficient is a nonparametric measure of statistical dependence between two variables. It assesses how well the

relationships between these two variables. This measure was used because it is insensitive to outliers and potential data normalization inaccuracies. Genes with low level of expression at zero in all species were excluded at the first step. The neighbor-joining trees were built using functions in the "ape" package [38] in R. The reliability of branching patterns was assessed with bootstrap analysis, which randomly sampled 1,000 times by replacement randomly.

2 Results

2.1 Construction of comprehensive lncRNA catalogs across nine species

To generate comprehensive lncRNA catalogs, we firstly applied a computational approach to reconstruct lncRNAs using RNA-Seq data from six tissues across nine species that represent all major mammalian lineages as well as birds (see Materials and methods). Totally, we identified 14,294–17,202 protein-coding genes and 4,141–13,709 lncRNAs per species (Figure 1a; Table S1). These lncRNAs were grouped into three classes, namely intergenic, intronic and antisense lncRNAs. This classification information can be easily fetched from our PhyloNONCODE database. By comparing with known protein-coding and lncRNA genes from Ensembl database [21], we could identify 76%–86% of protein-coding genes and 40%–50% of human and mouse lncRNAs respectively. Furthermore, we confirmed that we had successfully constructed 49% (6,426/13,249) of coding and 37% (3,023/8,195) noncoding human transcripts annotated by GENCODE V12 [32] and Human Body Map [11], validating the efficacy of our assemble pipeline.

Secondly, we completed the integration of the lncRNA data with Ensembl annotations [21] of corresponding

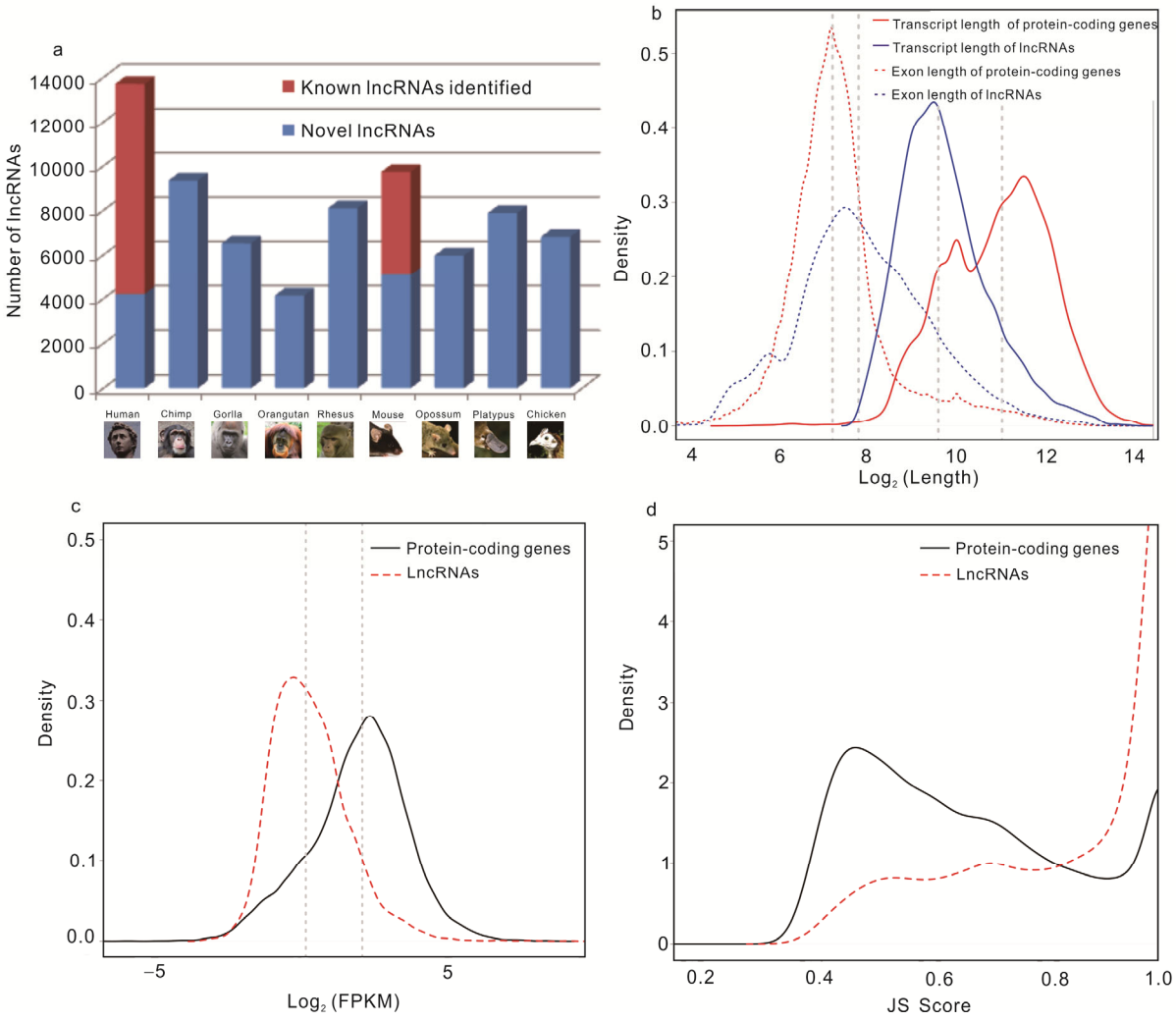

Figure 1 Comprehensive mammalian lncRNA catalogs. a, The statistics of lncRNA catalogs across eight mammalian species and chicken. The *Y*-axis indicates the number of lncRNAs. The known lncRNAs that were identified (red bar) in this study are presented. The novel lncRNAs identified are shown as blue bar. b, Exon and transcript length distributions for mouse lncRNAs and protein-coding genes. The *X*-axis is $\log_2$ of exon or transcript length and the *Y*-axis is the density. c, The expression profile of mouse lncRNAs and protein-coding genes. The *X*-axis indicates $\log_2$-normalized FPKM value estimated by Cufflinks. d, Distributions of maximal tissue specificity scores (JS score) calculated for both mouse lncRNAs (red) and protein-coding genes (black).

species (GENCODE [32], NONCODE [23–25] and Human Body Map [11] lncRNAs were included for human beings, NONCODE [24–26] were included for mouse). The current catalogs include 42,558 lncRNAs for human, 9,347 for chimpanzee, 6,517 for gorilla, 4,141 for orangutan, 8,094 for rhesus, 25,464 for mouse, 5,964 for opossum, 7,872 for platypus and 6,804 for chicken (Table S1). The exon length of lncRNAs is similar to that of protein-coding genes, but the length of whole transcripts of lncRNAs is shorter (Figure 1b; Figure S1).

Thirdly we characterized the expression of lncRNAs for each species based on the FPKM value [35]. Consistent with previous studies [11,18], the average expression level of lncRNAs in mammals are lower than protein-coding genes across all species (Figure 1c; Figure S2). To assess the extent of tissue specificity of lncRNA expression, each transcript was assigned with a tissue specificity score (JS score) [11]. Same as the previous reports [11,26], a markedly higher proportion (21%–50%, JS score>0.9) of lncRNAs are expressed in a tissue-specific manner, compared to only 10%–17% for protein-coding genes (Figure 1d; Figure S3; Tables S2 and S3). Thus, in all species we studied, lncRNAs are clearly more tissue-specific than their protein-coding counterparts, which might reflect their unique functionality in the nexus of complex biological systems.

2.2 Conservation study of lncRNAs in mammals

To systematically assess the conservation of lncRNAs in mammals, we analyzed their conservation levels at different evolutionary distances from each other. For each mammalian species examined, the genomic sequences of both lncRNAs and protein-coding genes were compared respectively to all the other species, and then pair-wise conserved status was evaluated. Using human as an example, we detected an average of 94.4% (40,185/42,558) genes that have conserved counterparts in placental species but the percentages drop to 70.1% (29,849/42,558) and 54.8% (23,303/42,558) for marsupials and monotremes respectively (Figure 2a; Table S4). For protein-coding genes, the similar conservation pattern was observed among three lineages (Figure S4). According to the genomic location of lncRNAs and their conserved counterparts, we defined Coverage Pattern of Conserved Counterparts (CPCC) score (see Material and methods) and classified conserved lncRNAs into three levels: (i) Unilateral conserved lncRNAs (CPCC score=0) represent the lowest conserved level. (ii) Bilateral and partial conserved lncRNAs (0<CPCC score<1) represent the intermediate conserved level. (iii) Bilateral and complete conserved lncRNAs (CPCC score=1) represent the highest conserved level. Based on this classification, a significantly larger portion of bilateral and complete conserved lncRNAs are observed in primates than in other species (Figure 2a). The pair-wise conservation atlases of each chromosome for human lncRNA genes also exhibit the same pattern (Figure 2b; Figure S5). Consistent with the reported studies [39,40], *hotair* were found widely existing in mammals, as shown by the CPCC score in PhyloNONCODE (Figure S11), and the *xist* gene existing from humans to the opossum, absent in the platypus and chicken (Figure S11). 62% and 59% of human lncRNAs were found to be 1∶1 orthologous in mouse and rat with CPCC score in 0 to 1, similar to a study of human long non-coding RNAs in the other six mammals, in which the figures are 58% and 54% [26]. As a means of alternative validation, we checked the conservation status of 18 evolutionarily conserved long intergenic non-coding RNAs (lincRNAs) in the eye identified in a previous study [41], most of which were found to be conserved across mammals in our study (Table S7).

Furthermore, investigation of the lineage-conservation of lncRNAs and protein-coding genes showed similar results with pair-wise conservation comparisons. The overall conservation of lncRNAs is significantly lower than that of protein-coding genes. Specifically, only 47.5% (20,202/42,558) of lncRNAs are conserved across mammals (94.4%, 84.9% and 63.8% are conserved across primates, placentals and both placentals and marsupials respectively), while nearly 81.0% (16,417/20,279) of protein-coding genes are mammal-conserved (Figure 2c; Tables S5 and S6). The conservation status evaluated by the phastcons score further demonstrated that the score of our conserved catalogs for both lncRNAs and protein-coding genes are higher than score of total genes (Figure 2d). Moreover, we checked the conservation status of 18 evolutionarily conserved long intergenic non-coding RNAs (lincRNAs) in the eye identified by previous study [41], 14 of which are conserved across mammals in our study (Table S7).

Determining the extent to which transcription of conserved lncRNA loci is retained or lost across multiple evolutionary lineages is essential if we are to understand their contribution to mammalian biology and to lineage-specific traits. To investigate the proportion of conserved genes that are also transcribed, the transcriptional status of pairwise- and lineage-conserved genes were evaluated based on RNA-Seq data (see Materials and methods). The results showed a distinct pattern between lncRNAs and protein-coding genes. Although a large proportion of lncRNAs exhibited conservation in placentals in a pair-wise manner but only 19.3%–26.3% of these genomic conserved lncRNAs are expressed in corresponding species (Figure 2a; Table S4). On the contrary, the expressed genes contribute to most of the genomic conserved genes for protein-coding genes (70.1%–79.6%) (Figure S4). These observations are also obvious when it comes to lineage-conserved genes. Respectively, 45.9% (18,434/40,160), 51.0% (18,428/36,131) and 68.3% (13,804/20,202) of primate-, placental- and mammalian-conserved lncRNAs are expressed in their lineages (Figure 2c). By comparison, 86.8% (17,050/19,635), 88.6% (17,156/19,371) and 92.8% (15,235/16,417)

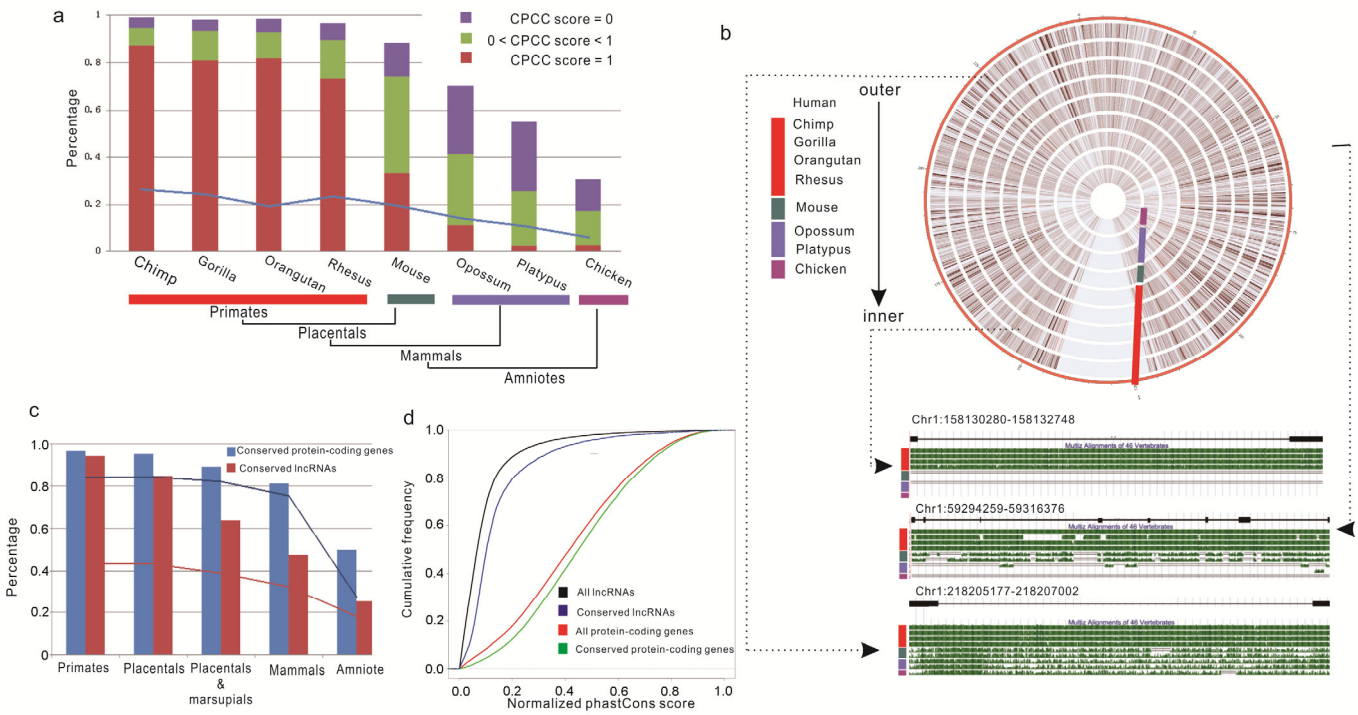

Figure 2 Conservation analysis of lncRNAs and protein-coding genes. a, Number of human lncRNAs found to be conserved through pair-wise conservation search. The bars for each species represent the conserved counterpart number in the target species. The line across the bar represents the number of conserved lncRNAs that are expressed. The known taxonomy is showed on the bottom. CPCC score is defined in material and methods. b, LncRNAs conservation atlas for human chromosome 1. Up panel shows circos plots of conservation atlas. The outer to inner rings represent the human lncRNAs to conserved lncRNAs in chicken respectively. Conservation degree is shown in dark color. Three examples representing primate-, placental mammal- and amniotes-conserved lncRNAs are shown at the bottom. For each example, the lncRNA structure (black) and multiz-alignments of 46 vertebrates (green) are shown. c, Number of lncRNAs (red) and protein-coding genes (blue) found to be conserved across different lineages. The lines represent the number of conserved lncRNAs (red) and protein-coding genes (blue) that expressed in at least one species of each such lineage. d, Cumulative distribution of phastcons score for all protein-coding genes (red), conserved protein-coding genes (green) in mammal, all lncRNAs (black) and conserved lncRNAs (blue) in mammal.

are the figures for protein-coding genes (Figure 2c). These results suggest that lncRNA transcriptome is much more specific (including species-specific and lineage-specific) than their corresponding genomic context and could serve as a more reliable reflection of their evolutionary standing among different species.

2.3 The evolutionary trends of lncRNA expression

A growing body of evidence supports the concept that gene regulation modifications could produce the major phenotypic differences that underlie adaptive changes and our results above have highlighted the contribution of lncRNA expression to the species- and lineage-specificity. To trace the evolutionary pattern of the lncRNA expression, we constructed gene expression trees by building expression distance matrices for each tissue. In order to gain insights into the differences in evolutionary modes between protein-coding gene and lncRNA expressions, separate trees were built for 10,061 protein-coding genes and 10,770 lncRNAs conserved in all nine amniotic species that allows side-by-side comparison. Consistent with previous study and current protein-coding gene expression trees [20], lncRNA expression trees were found to be highly consistent with known mammalian phylogeny (Figure 3a and 3b; Figures S6 and S7). The lncRNA expression trees correctly resolved the three major mammalian lineages (placentals, marsupials, and monotremes). Primates were separated from rodent, while human and the other great apes were grouped together with exclusion of the rhesus macaque. This observation suggests that expression changes of lncRNAs accumulate over evolutionary time leading to similar expression levels among closely related species. To further validate the robustness and scalability of our approach, we carried out similar analysis on an independent RNA-Seq data from five tissues of rat [27]. Overall, 18,330 lncRNAs genes were identified, of which 16,163 are novel discovered. These 18,330 rat lncRNA genes were analyzed by the above downstream procedure. The final results show that when rat lncRNA expression information from five tissues was incorporated into phylogenetic trees, all species maintain their correct evolutional placement (Figure S8). This excise validates that our molecular evolutionary analysis using RNA-Seq data is very robust and could be applied to RNA-Seq data of diverse sources.

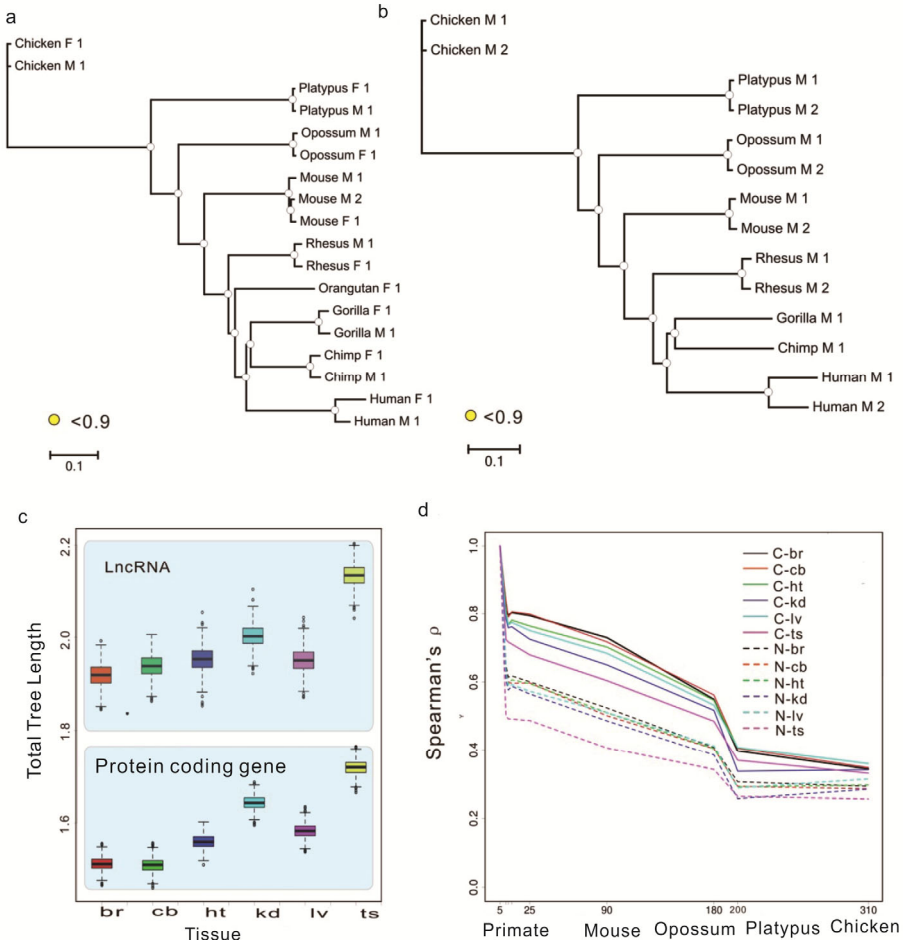

Figure 3 Mammalian lncRNA expression phylogenies. (a) and (b) are mammalian lncRNA expression phylogenetic trees for cerebellum and testis. Bootstrap values (10,770 lncRNA genes randomly sampled with replacement 1,000 times) are indicated by circles: white≥0.9; yellow<0.9. c, Comparisons of total branch lengths of expression trees between the six tissues examined (br: brain; cb: cerebellum; ht: heart; kd: kidney; lv: liver; ts: testis), for protein-coding genes (box with solid line) and lncRNAs (box with dotted line). Errors: 95% confidence intervals based on bootstrap analysis (1,000 replicates, with one individual per species sampled in each replicate). d, Spearman's correlations between gene expression of human and the other species. C: protein-coding genes; N: lncRNAs. br: brain; cb: cerebellum; ht: heart; kd: kidney; lv: liver; ts: testis. For example, C-br denotes for Spearman's correlation of protein-coding gene expression in the brain.

2.4 Rates of expression change for lncRNAs in lineages and tissues

We next investigated the rates of expression changes for lncRNAs in lineages and tissues. The total branch lengths of expression trees for lncRNA were markedly greater than those for protein-coding genes, suggesting that the evolution of lncRNA expression may proceed more rapidly (Figure 3c), which is consistent with previous study [17]. However, the total branch lengths of the expression trees among different tissues exhibited similar trends for both protein-coding genes and lncRNAs: expression trees of both of them are found to vary widely among tissues (Figure 3c). Consistent with previous study derived from protein-coding genes [20], the lncRNA expression trees showed that the two neural tissues apparently evolved significantly more slowly than the other tissues. This indicates that the lncRNAs in these neural tissues may also have experienced stronger selection pressure during mammalian evolution. Similarly, lncRNAs in the testis tissue appear to have evolved significantly more quickly than in the other tissues, suggesting weaker selection pressure in force upon lncRNAs in the tissue (Figure 3c). Furthermore, pair-wise species comparisons conducted for both protein-coding genes and lncRNAs generated consistent results as expression phylogeny analyses described above and provided addition support that the lncRNA expression divergence increased more than protein-coding gene expression (Figure 3d).

2.5 PhyloNONCODE is a comprehensive data source to study the conservative and evolutionary status of lncRNAs

Based on lncRNA data collected and analyzed in this study, we have developed PhyloNONCODE database. Phylo-

NONCODE is included in our analysis platform for noncoding RNAs, which consists of ncRNA resources such as NONCODE [23–25], NPInter [42], antiCODE [43] and PhyloNONCODE as well as online tools and web servers for analysis of ncRNAs [44]. As a member of the union of specific databases and tools for noncoding RNAs, PhyloNONCODE is the first open access resource for large-scale identification and evolutionary annotation of mammalian lncRNAs and a one-stop knowledge gateway for lncRNA evolution studies. Currently, PhyloNONCODE cover the most comprehensive dataset of lncRNAs in 10 organisms, including nine mammals and one bird. User queries can be customized based on the conservation statuses of user-defined species and conservation degree for each lncRNA at search page. PhyloNONCODE provides "Browser" tool to display all genes in all species and their conservation status in other species. Links to the detailed information of each lncRNA are included in the browse or blast search result page to allow users to easily obtain lncRNA basic information as well as visualization for further analysis (Figure 4). Since its visualization platform is based on a local UCSC genome browser, both existing and

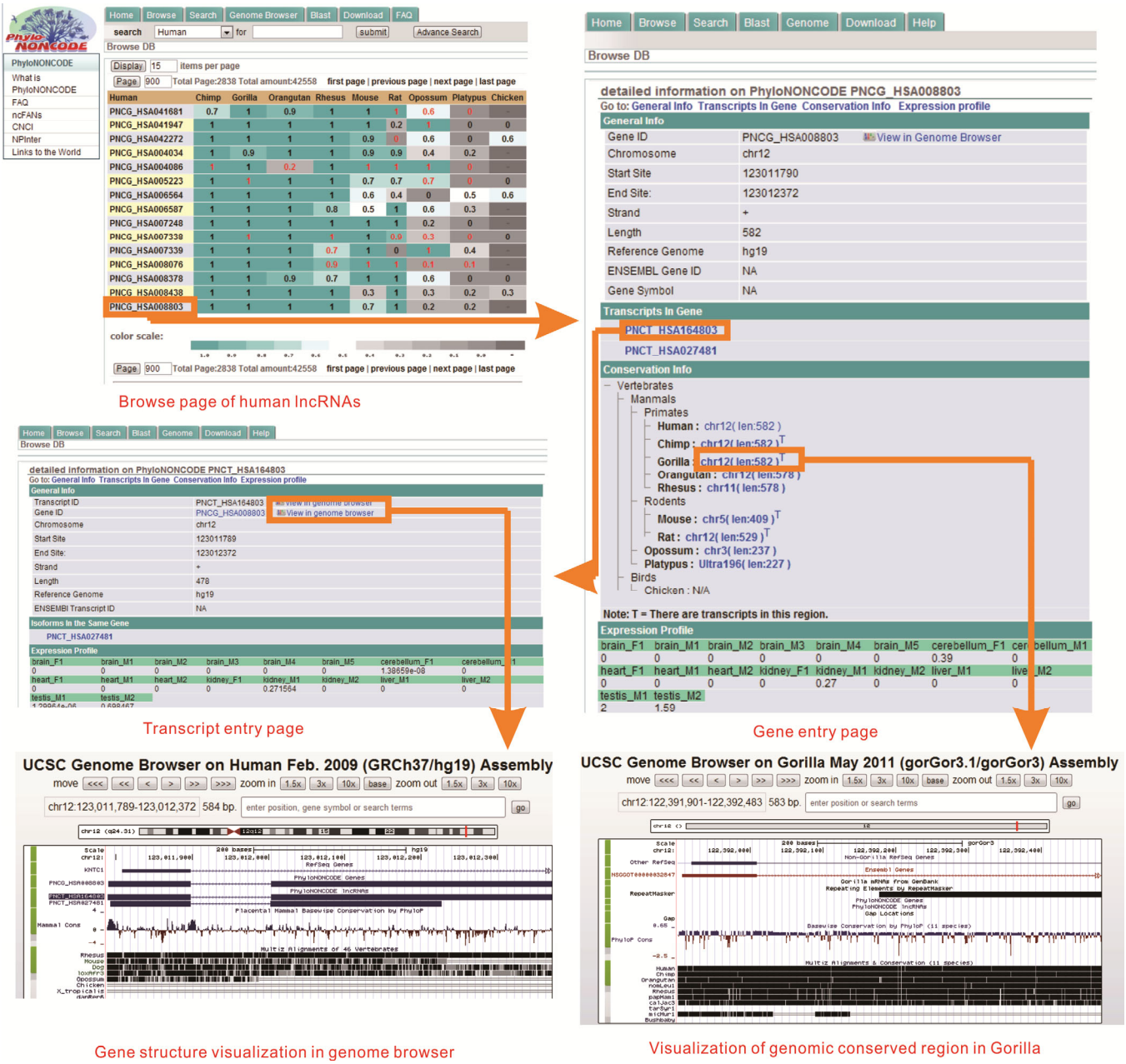

Figure 4 Browse, and visualization of human lncRNAs in PhyloNONCODE. In browse page, the degree of conservation for each lncRNA in the other nine species is colored and marked as the CPCC score. Links to the transcript/gene entry page of each lncRNA are included in the browse page, that allow users to easily view its conserved information as well as visualization for further analysis.

newly updated data are available. In addition, users could also add their own tracks to display the genome context information of lncRNAs of their interest. The analysis of the independent rat RNA-Seq dataset described above has proved that our approach is universally effective for most RNA-Seq data.

3 Discussion

The existing annotations of lncRNAs for different organisms vary significantly in their coverage and depth. Human and mouse have the largest numbers of annotated lncRNAs in the published database, while lncRNAs in others species, even within mammalian lineage, have rarely been documented and analyzed in a systemic manner. Our current work provides comprehensive lncRNA catalogs including some of the poorly-annotated mammalian species. Unlike the previous approach, which uses Codon Substitution Frequencie (CSF) score or the presence of sequence similarity with known protein to distinguish lncRNA from protein-coding genes [11,17], here we used CNCI [31], a powerful signature tool, by profiling adjoining nucleotide triplets to effectively distinguish protein-coding and non-coding sequences independent of known annotations. CNCI is effective for classifying incomplete transcripts and sense-antisense pairs. The implementation of CNCI offered highly accurate classification of transcripts assembled from whole-transcriptome sequencing data in a cross-species manner. Our previous work has proved that CNCI is more suitable for RNA assembled from cross-species RNA-Seq data, especially for those non-model organisms.

For human and mouse dataset, it is interesting to note that only a small fraction of the known lncRNA genes can be reconstructed using RNA-Seq data for six tissues, yet most of the known protein-coding genes can be easily assembled. This observation probably reflects the different expression patterns between lncRNAs and protein-coding genes, with lncRNAs much more tissue and cell-type specific [11]. Certainly the proportion of known human lncRNAs identified from the six tissues is consistent with a previous investigation [11] which utilized RNA-Seq data of a total of 24 normal tissues and cell lines. Compared with the lncRNAs reported by Necsulea et al. [17] we found the overlap between two studies is limited. In the case of homo sapiens, the consistency of two studies is about 50%. Taken it into account that the lncRNA we report are mainly from NONCODE and prediction from sequencing data, it is conceivable that differences of lncRNA sources and prediction tool may have significant impact on results. As the reliability of lncRNA data is still unknown to date, we are trying to extract one subset of lncRNAs, which can be regarded as the golden set, which would be available on our next version of PhyloNONCODE.

The poor genomic conservation of lncRNAs made it difficult to assess their functions in species evolution but their unique tissue-specific expression pattern does suggest they might contribute to this process. The emergence of next-generation sequencing technology made it possible to conduct evolutionary investigations based on the expression level of lncRNAs. Here we were able to integrate gene conservation and transcriptional profiles to establish a global view of mammalian lncRNA evolution. As the existing studies illustrate, protein-coding genes are often widely expressed and nearly always very deeply conserved, while lncRNAs have high spatio-temporal specificity and rapid turnover [17,18,22,26]. We have assumpted that the difference may suggest the different roles that these two types of genes might play: with protein-coding genes as the direct encoder of the phenotype and lncRNAs as the functional regulator in the specific organisms [45–48].

It is noteworthy that, the conservation status of some lncRNAs whose genomic locations overlap with protein-coding genes may not be able to be defined as accurately as those of long intergenic non-coding RNA (lincRNAs) [14]. However, in order to provide an inclusive lncRNA conservation data to the broad scientific community, all lncRNAs are included in our current analysis. In fact, when using the same pipeline to lincRNAs, only minor difference was observed (the position of platypus and opossum in the tree are reversed compared with known taxonomy) (Figures S9 and S10).

Both levels of conservation were integrated into the database PhyloNONCODE for the first time, in which the CPCC score represents genomic conservation and FPKM represents the transcriptional information. PhyloNONCODE provides biologists with an informative resource to further explore the general features and unique characteristics of lncRNAs which might eventually shed light on their functional and evolutionary significance. The decreasing cost and improved depth of the RNA-sequencing technology have allowed transcriptomes of diverse species to be studied in their entirety. As a result, it is expected that novel lncRNAs will be discovered on a continuous basis with an unprecedented pace. The lncRNA analysis pipeline we established represents a powerful tool to curate these enormous datasets, and our PhyloNONCODE database offers a resourceful platform for assessing the evolutionary standings of individual or collection of lncRNAs, which could provide critical insights into their functions.

The authors declare that they have no conflict of interest. This research has fully complied with research ethics.

This work was supported by Training Program of the Major Research Plan of the National Natural Science Foundation of China (91229120).

1 Mattick JS, Makunin IV. Non-coding RNA. Hum Mol Genet, 2006, 15: R17–R29

2 Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell, 2009, 136: 629–641
3 Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev, 2009, 23: 1494–1504
4 Moran VA, Perera RJ, Khalil AM. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. Nucleic Acids Res, 2012, 40: 6391–6400
5 Puvvula PK, Desetty RD, Pineau P, Marchio A, Moon A, Dejean A, Bischof O. Long noncoding RNA PANDA and scaffold-attachment-factor SAFA control senescence entry and exit. Nat Commun, 2014, 5: 5323
6 Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res, 2013, 41: D983–D986
7 Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. Genetics, 2013, 193: 651–669
8 Liu Q, Huang J, Zhou N, Zhang Z, Zhang A, Lu Z, Wu F, Mo YY. LncRNA loc285194 is a p53-regulated tumor suppressor. Nucleic Acids Res, 2013, 41: 4976–4978
9 Xue Y, Ma G, Gu D, Zhu L, Hua Q, Du M, Chu H, Tong N, Chen J, Zhang Z, Wang M. Genome-wide analysis of long noncoding RNA signature in human colorectal cancer. Gene, 2015, 556: 227–234
10 Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, Mosquera JM, Pauwels J, Park K, Kossai M, MacDonald TY, Fontugne J, Erho N, Vergara IA, Ghadessi M, Davicioni E, Jenkins RB, Palanisamy N, Chen Z, Nakagawa S, Hirose T, Bander NH, Beltran H, Fox AH, Elemento O, Rubin MA. The oestrogen receptor α-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun, 2014, 5: 5383
11 Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011, 25: 1915–1927
12 Marques AC, Ponting CP. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol, 2009, 10: R124
13 Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res, 2007, 17: 556–565
14 Mitchell Guttman IA, Garber M, French C, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, 458: 223–227
15 Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res, 2010, 20: 1313–1326
16 Liz J, Portela A, Soler M, Gómez A, Ling H, Michlewski G, Calin GA, Guil S, Esteller M. Regulation of pri-miRNA processing by a long noncoding RNA transcribed from an ultraconserved region. Mol Cell, 2014, 55: 138–147
17 Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 2014, 505: 635–640
18 Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet, 2012, 8: e1002841
19 Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. Genome Biol Evol, 2011, 3: 1390–1404
20 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H. The evolution of gene expression levels in mammalian organs. Nature, 2011, 478: 343–348
21 Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM. Ensembl 2011. Nucleic Acids Res, 2011, 39: D800–D806
22 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res, 2012, 22: 1775–1789
23 Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res, 2005, 33: D112–D115
24 He S, Liu C, Skogerbø G, Zhao H, Wang J, Liu T, Bai B, Zhao Y, Chen R. NONCODE v2. 0: decoding the non-coding. Nucleic Acids Res, 2008, 36: D170–D172
25 Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, Zhao H, Liu Z, Liu C, Chen R, Zhao Y. NONCODE v3. 0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012, 40: D210–D215
26 Washietl S, Kellis M, Garber M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome Res, 2014, 24: 616–628
27 Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 2012, 338: 1593–1599
28 Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ. The UCSC genome browser database: update 2011. Nucleic Acids Res, 2011, 39: D876–D882
29 Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
30 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
31 Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res, 2013, 41: e166
32 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res, 2012, 22: 1760–1774
33 Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci USA, 2003, 100: 11484–11489
34 Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. Human-mouse alignments with BLASTZ. Genome Res, 2003, 13: 103–107
35 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and

quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515
36 Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, 1987, 4: 406–425
37 Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol, 1988, 5: 729–731
38 Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 2004, 20: 289–290
39 Rens W, O'Brien PC, Grutzner F, Clarke O, Graphodatskaya D, Tsend-Ayush E, Trifonov VA, Skelton H, Wallis MC, Johnston S, Veyrunes F, Graves JA, Ferguson-Smith MA. The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. Genome Biol, 2007, 8: R243
40 He S, Liu S, Zhu H. The sequence, structure and evolutionary features of hotair in mammals. BMC Evol Biol, 2011, 11: 102
41 Mustafi D, Kevany BM, Bai X, Maeda T, Sears JE, Khalil AM, Palczewski K. Evolutionarily conserved long intergenic non-coding RNAs in the eye. Hum Mol Genet, 2013, 22: 2992–3002
42 Wu T, Wang J, Liu C, Zhang Y, Shi B, Zhu X, Zhang Z, Skogerbø G, Chen L, Lu H, Zhao Y, Chen R. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res, 2006, 34: D150–D152
43 Yin Y, Zhao Y, Wang J, Liu C, Chen S, Chen R, Zhao H. antiCODE: a natural sense-antisense transcripts database. BMC Bioinformatics, 2007, 8: 319
44 Liao Q, Xiao H, Bu D, Xie C, Miao R, Luo H, Zhao G, Yu K, Zhao H, Skogerbø G, Chen R, Wu Z, Liu C, Zhao Y. ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res, 2011, 39: W118–W124
45 Roberts TC, Morris KV, Weinberg MS. Perspectives on the mechanism of transcriptional regulation by long non-coding RNAs. Epigenetics, 2014, 9: 13–20
46 Pennisi E. Lengthy RNAs earn respect as cellular players. Science, 2014, 344: 1072–1072
47 Lau E. Non-coding RNA: zooming in on lncRNA functions. Nat Rev Genet, 2014, 15: 574–575
48 Alvarez-Dominguez JR, Hu W, Yuan B, Shi J, Park SS, Gromatzky AA, van Oudenaarden A, Lodish HF. Global discovery of erythroid long noncoding RNAs reveals novel regulators of red cell maturation. Blood, 2014, 123: 570–581

Supporting Information

Figure S1 On and transcript length distributions for lncRNAs and protein-coding genes.

Figure S2 Gene expression distributions for lncRNAs and protein-coding genes.

Figure S3 Distributions of tissue specificity scores for lncRNAs and protein-coding genes.

Figure S4 Number of human protein-coding genes found to be conserved through pair-wise conservation search.

Figure S5 Circos plots of lncRNAs conservation atlas for all human chromosomes.

Figure S6 Mammalian protein-coding gene expression phylogenies.

Figure S7 Mammalian lncRNA expression phylogenies(without rat).

Figure S8 Mammalian lncRNA expression phylogenies (including rat).

Figure S9 Mammalian lincRNA expression phylogenies(without rat).

Figure S10 Mammalian lincRNA expression phylogenies (including rat).

Figure S11 Evolutionary conservation of hotair and xist.

Table S1 Statistics of lncRNA catalogues across mammals

Table S2 Statistics of tissue specific protein-coding genes

Table S3 Statistics of tissue specific lncRNAs

Table S4 Number of conserved lncRNA gene loci from pair-wise conservation search

Table S5 Number of protein-coding genes conserved in different lineages

Table S6 Number of lncRNAs conserved in different lineages

Table S7 Conservation status of 14 evolutionarily conserved eye lincRNAs

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.