BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2480–2485
doi:10.1093/bioinformatics/btu325

Databases and ontologies

Advance Access publication May 9, 2014

lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs

Charny Park†, Namhee Yu†, Ikjung Choi, Wankyu Kim and Sanghyuk Lee*

Ewha Research Center for Systems Biology (ERCSB), Department of Life Science, Ewha Womans University, Seoul 120-750, Republic of Korea

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: A number of long non-coding RNAs (lncRNAs) have been identified by deep sequencing methods, but their molecular and cellular functions are known only for a limited number of lncRNAs. Current databases on lncRNAs are mostly for cataloging purpose without providing in-depth information required to infer functions. A comprehensive resource on lncRNA function is an immediate need.

Results: We present a database for functional investigation of lncRNAs that encompasses annotation, sequence analysis, gene expression, protein binding and phylogenetic conservation. We have compiled lncRNAs for six species (human, mouse, zebrafish, fruit fly, worm and yeast) from ENSEMBL, HGNC, MGI and lncRNAdb. Each lncRNA was analyzed for coding potential and phylogenetic conservation in different lineages. Gene expression data of 208 RNA-Seq studies (4995 samples), collected from GEO, ENCODE, modENCODE and TCGA databases, were used to provide expression profiles in various tissues, diseases and developmental stages. Importantly, we analyzed RNA-Seq data to identify coexpressed mRNAs that would provide ample insights on lncRNA functions. The resulting gene list can be subject to enrichment analysis such as Gene Ontology or KEGG pathways. Furthermore, we compiled protein–lncRNA interactions by collecting and analyzing publicly available CLIP-seq or PAR-CLIP sequencing data. Finally, we explored evolutionarily conserved lncRNAs with correlated expression between human and six other organisms to identify functional lncRNAs. The whole contents are provided in a user-friendly web interface.

Availability and implementation: lncRNAtor is available at http://lncrnator.ewha.ac.kr/.

Contact: sanghyuk@ewha.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 3, 2013; revised on April 26, 2014; accepted on May 2, 2014

1 INTRODUCTION

Long non-coding RNAs (lncRNAs) are non–protein-coding transcripts >200 nt. Many lncRNAs have been reported to play important molecular and cellular functions despite of not producing protein products. They typically recruit proteins to help or disturb formation of ribonucleoprotein (RNP) complexes. Depending on the role of interacting proteins, lncRNAs play central roles in a wide range of cellular processes. An illustrative example is RMST, a brain-specific lncRNA, which physically interacts with SOX2 and plays a key role in transcriptional regulation of neurogenic transcription factors (Ng *et al.*, 2013).

Several lncRNAs were implicated in various types of cancers (Maruyama and Suzuki, 2012). MALAT1 was reported to be a critical regulator of the metastasis phenotype of lung cancer cells (Gutschner *et al.*, 2013), and CCAT2 was shown to promote tumor growth, metastasis and chromosomal instability in WNT-dependent fashion in colon cancer (Ling *et al.*, 2013). In fact, lncRNAs are expected to be involved in every stage of hallmarks of cancer (Gutschner and Diederichs, 2012), thereby serving as critical regulators and therapeutic targets in many cases.

Owing to the development of high-throughput sequencing technologies, thousands of lncRNAs have been recently discovered from RNA-Seq data. However, the biological and molecular characteristics of the large majority of lncRNAs remain unknown. Databases or tools that facilitate functional investigation of lncRNAs would be of great value to identify important lncRNAs based on understanding of biological roles.

Several databases on lncRNAs have already been developed for various purposes. LNCipedia (Volders *et al.*, 2013), lncRNAdb (Amaral *et al.*, 2011) and lncRNome (Bhartiya *et al.*, 2013) are mostly for annotation databases based on literature evidence. NRED (Dinger *et al.*, 2009) and NONCODE v3.0 (Bu *et al.*, 2012) provide microarray expression profiles in various tissues for human and mouse. Some of these databases contain additional useful information on function such as the RNA secondary structure, microRNA binding and protein-lncRNA interactions (Bhartiya *et al.*, 2013; Volders *et al.*, 2013), but they are mostly limited for human. Furthermore, RNA-Seq data, the most useful source of lncRNA properties, have never been systematically collected and mined for functional investigation yet. Thus, we still lack a comprehensive resource to cover sequence characteristics and function-related data organized in a systematic way.

Here, we introduce lncRNAtor to perform in-depth functional analysis of lncRNAs by combining sequence characteristics, gene expression data and protein–lncRNA interactions. To the best of our knowledge, this is the first attempt to compile massive RNA-Seq data with the support of downstream analysis such as

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

coexpression and gene set enrichment, covering six important model organisms.

2 DATA AND METHODS

- **lncRNA compendium:** lncRNAs of length >200 bp in ENSEMBL (version 70), HGNC, MGI and lncRNAdb were collected for six model organisms (human, mouse, zebrafish, fruit fly, worm and yeast). Statistics of compiled lncRNAs are provided in Supplementary Table S1.
- **Sequence analysis and phylogenetic conservation:** Each lncRNA was analyzed for protein-coding potential with CPC, an SVM-based classifier using sequence features (Kong *et al*., 2007). We also calculated the evolutionary conservation score using the phastCons track (multiz alignments of 46 vertebrates) from UCSC genome database (Pollard *et al*., 2010). We used the reduced representaion of genes by collaping all splice variants into a hypothetical gene. Lineage-specific scores were obtained by analyzing multiple alignment of organisms in specific lineages such as primates, mammals and vertebrates. Genomic regions overlapping with protein-coding genes were eliminated in calculating average conservation score to avoid the misleading hyper-conservation.
- **RNA-Seq data:** We have collected RNA-Seq data in GEO, ENCODE, modENCODE and TCGA databases. Each dataset was manually curated and classified into tissue types, cancer types, drugs and developmental stages according to the experimental design. In total, we have compiled 208 datasets of 4995 samples. Raw RNA-Seq data were subject to in-house pipeline of mapping by TopHat 2.0.8 (Kim *et al*., 2013) and quantification by Cufflinks 2.1.0 (Trapnell *et al*., 2010). Multi-hits was set to 20 as the default option. Differentially expressed genes (DEGs) were identified by Cuffdiff2 (Trapnell *et al*., 2013). TCGA datasets were downloaded at level 3 of read counts and expression profiles obtained from RSEM quantification (Li and Dewey, 2011), and reannotated from hg18 to ENSEMBL GRCh37. DEG test was performed by DESeq 2 1.0.19 (Anders *et al*., 2013).
- **Protein–lncRNA binding data:** Deep sequencing data of CLIP-Seq, RIP-Seq and PAR-CLIP were amassed from GEO and modENCODE. It included 319 samples covering 96 RNA-binding proteins. RIPSeeker was used to identify the associated transcripts and binding sites (Li *et al*., 2013). We have adjusted the binding *P*-value taking the number of binding regions into consideration.
- **Coexpression and gene set analysis:** Coexpression of genes and lncRNAs was calculated for RNA-Seq datasets with the number of samples >10. The most abundant transcript was selected to be the representative of gene expression, thus ignoring the isoform difference. All pairwise correlations were pre-calculated using the Spearman's rank correlation to reduce the influence of outlier genes, and top 1000 correlations were stored for each coding or lncRNA gene for speed and efficiency in web implementation. Gene set analysis of overrepresentation were supported for Gene Ontology (GO) terms and KEGG pathways. We used the upper-tail hypergeometric test with the *P*-value cutoff of 0.01 and the multiple test corrections (Bonferroni and Benjamini–Hochberg methods) using all genes with GO or KEGG annotation as the background distribution. Gene sets of size ≤5 or hits ≤2 were filtered out from the overrepresented terms.

3 RESULTS

3.1 Database overview

The scope of database, category of analyses and characteristics of data are briefly summarized in the schematic overview of lncRNAtor in Figure 1.

LncRNAtor is based on three main types of data—annotation and conservation, gene expression from RNA-Seq and protein-lncRNA interactions. It includes 21 575 lncRNAs from six model organisms of human, mouse, zebrafish, fruit fly, worm and yeast. Our compendium dataset includes 14 051 human, 4030 mouse, 1666 zebrafish, 501 fruit fly, 1312 worm and 15 yeast lncRNA genes. Importantly, lncRNAtor features the most extensive compilation of RNA-Seq data, covering 208 datasets of 4995 samples, and protein–lncRNA interactions for 96 proteins of 319 samples.

These data were organized into several modules of analyzing (i) sequence features such as protein-coding potential and cross-species conservation, (ii) expression profiles and differential expression based on RNA-Seq data, (iii) protein–lncRNA binding obtained from deep sequencing data and (iv) functional inference of GO and KEGG pathways based on coexpressed protein-coding genes.

3.2 Basic features

The Web site is composed of the basic search and several advanced analyses of coexpression, differential expression and binding proteins. Figure 2a shows the output screenshot from the basic search, which includes the brief summary of each lncRNA with linkouts to other databases. The non-coding nature of lncRNAs can be confirmed by examining the protein-coding potentials, ORF prediction and the blast hits provided by the CPC program (Kong *et al*., 2007).

Next, we show the conservation score obtained from the multiple alignment data of phastCons track in the UCSC genome database. Lineage-specific scores were calculated for promoter (–500 bp), 5′-end of transcript (100 bp), exon, intron and 3′-end of transcript (100 bp) regions on various lineages including primate, mammal and vertebrate. We also provide a link to the UCSC genome browser, showing phylogenetic conservation at base-pair resolution.

Gene expression profiles are provided in a separate tab menu for representative tissues, cancers and developmental stages according to the characteristics of RNA-Seq data. We show the normalized expression value (FPKM) in box plots as shown in Figure 2b. For user convenience, we provide additional tab menus of differential expression and binding proteins for lncRNA of interest, with details explained in the following section.

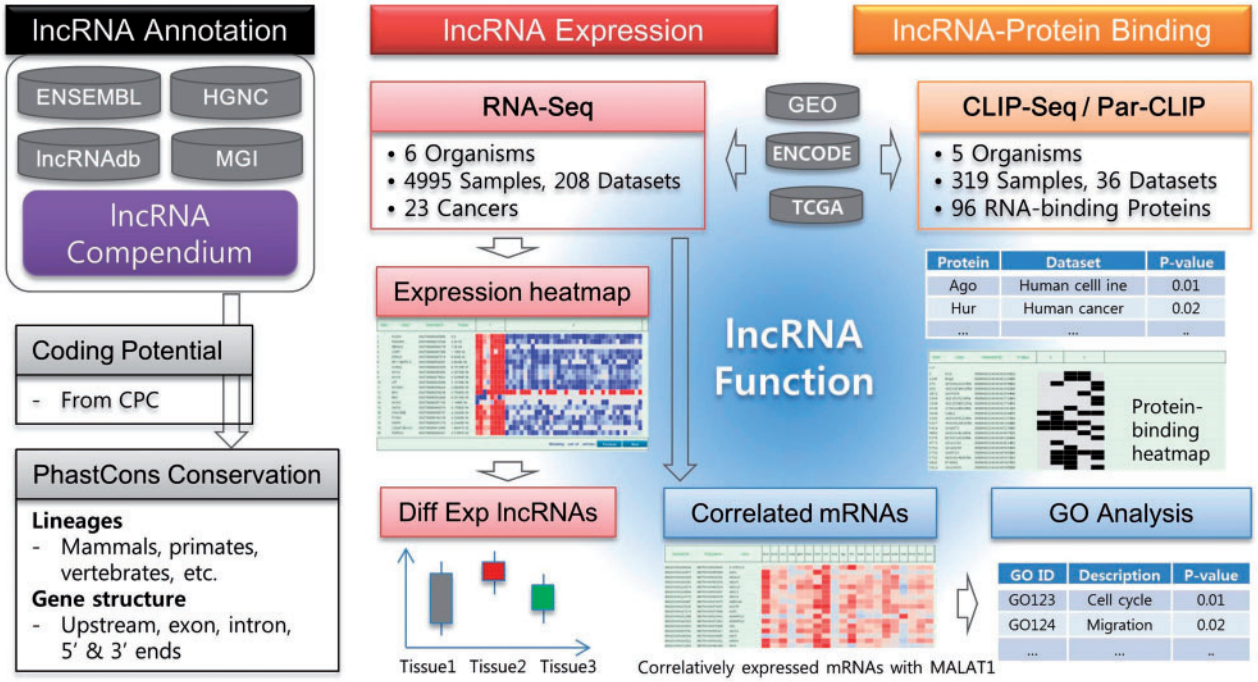

Fig. 1. Overview of lncRNAtor database

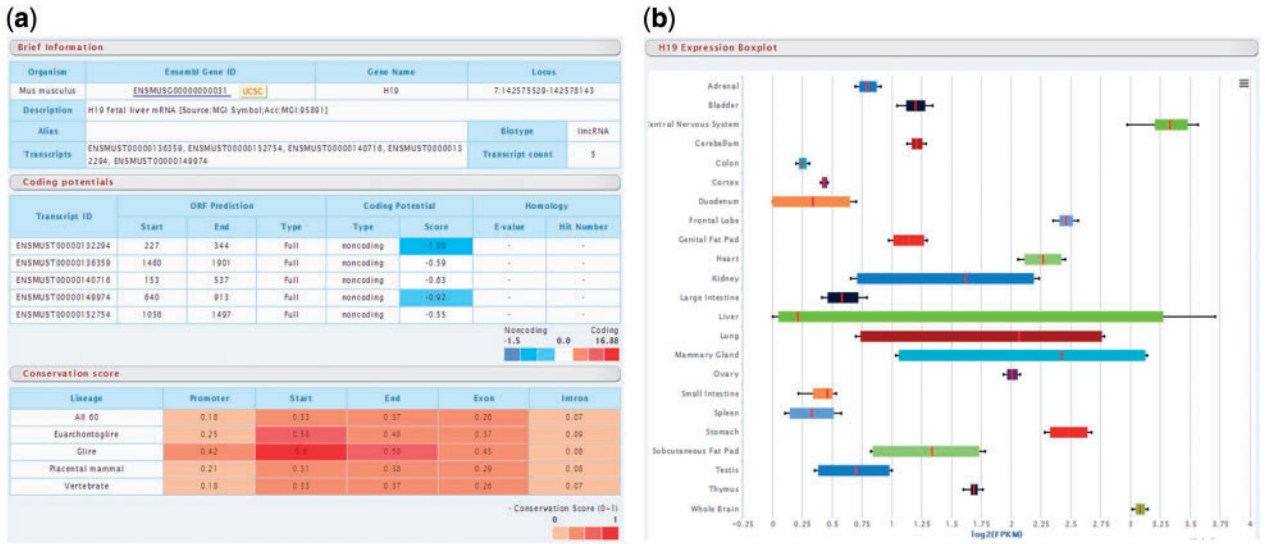

Fig. 2. Sample output from searching mouse H19 lncRNA. (**a**) Brief information of ID annotation, coding potential and conservation scores by lineage. (**b**) Bar plot of gene expression profile in various tissues

We have examined the gene expression levels in nine RNA-Seq datasets from TCGA, GEO and ENCODE databases including 352 samples from human, mouse, worm, fruit fly and zebrafish (Supplementary Table S2). The distribution curves in Supplementary Figure S1 show that the cutoff value of FPKM = 1.0 seemed a reasonable choice to differentiate transcripts within noise level for all datasets. Transcripts of average FPKM < 1.0 were removed from further analyses of differential expression and coexpression. We have also compared the expression of protein-coding genes and lncRNAs in each dataset. The distribution curves of lncRNAs and protein-coding genes were shown in Supplementary Figure S2. The proportion of expressed (i.e. FPKM > 1.0) lncRNAs was < 10% of total transcripts in most datasets, much smaller values compared with that of

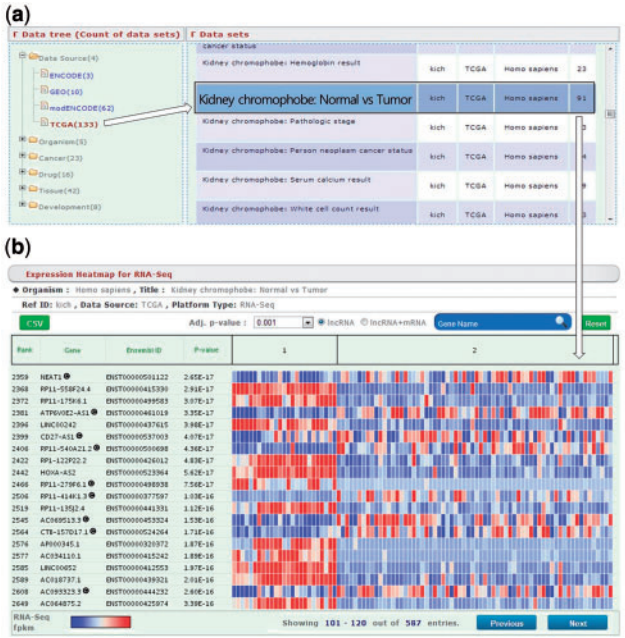

Fig. 3. Exploring differential expression. (**a**) Dataset browser to search for experimental conditions. (**b**) Expression heatmap for differentially expressed lncRNAs

protein-coding genes (Supplementary Table S3). This is consistent with the previous result that reported tissue-specific expression pattern for lncRNAs (Derrien *et al.*, 2012).

3.3 Differential expression and binding proteins

Differential expression of lncRNAs is valuable information to identify lncRNAs of functional roles in specific contexts. In an effort to provide a comprehensive summary of RNA-Seq datasets, we devised a dataset browser of tree structure, classifying RNA-Seq datasets into distinct groups of organism, cancer, drug, tissue and developmental stage. Figure 3a shows the example of 133 datasets in the TCGA category (the kidney part).

All RNA-Seq datasets were pre-processed for differential expression of lncRNAs. For each dataset, we manually classified samples into the different groups and performed the statistical test to identify differentially expressed lncRNAs and mRNAs (see Section 2 for details).

Selecting the dataset and comparison of interest in the browser tree ('Kidney chromophobe: Normal versus Tumor' in this example) shows the expression heatmap of differentially expressed lncRNAs (and mRNAs if needed) as shown in Figure 3b. The list can be sorted, searched and exported into a separate file for further analysis. Abundant datasets of TCGA and GEO allow users to explore differential expression of lncRNAs in diverse contexts such as tumor stages and drug treatment. Model organism data provide valuable information on lncRNAs of developmental roles.

Proteins binding to lncRNAs are valuable source of information on lncRNA function. We obtained 35 867 lncRNA–protein bindings (adj. $P<0.01$) by analyzing the IP-based deep sequencing data of 96 RNA-binding proteins and 319 samples. These RNA-binding proteins were classified into different functional groups of transcription (hnRNPs, CPSFs), binding (ELAV1, RBM4), microRNA (AGO, LIN28), epigenetic regulation (Ezh2, Polycomb complex) and so on. Although the current dataset is rather small, many proteins of importance in RNA processing and regulation are already covered to provide lncRNA candidates involved in these processes. Statistics of lncRNA–protein binding is available in online documentation. Selecting a specific RNA-binding protein shows the binding heatmap of lncRNAs (and mRNAs) in a similar manner to the expression heatmap.

3.4 Coexpression and functional gene set analysis

Guilt-by-association is the principal method of predicting functions of genes with unknown function. Thus, protein-coding mRNAs that are coexpressed with lncRNA of interest often provide ample insights into molecular functions (Liao *et al.*, 2011). Because RNA-Seq data provide the gene expression profile of coding as well as non-coding RNAs in an unbiased manner, it is an ideal source data to explore coexpression between coding mRNAs and lncRNAs. A systematic study on correlated expression of lncRNAs was carried out by the GENCODE consortium (Derrien *et al.*, 2012).

We developed a module to explore the coexpressed protein-coding mRNAs in a context-specific manner. Supported datasets include RNA-Seq data in the TCGA, GEO and ENCODE databases. Human datasets are mostly for cancer from TCGA and GEO. Mouse (GSE36025 of CSHL, GSE36026 of LICR) and zebrafish (GSE30608, ERP000016 from Sanger) datasets from ENCODE cover various tissues. Worm and fruit fly datasets from modENCODE cover developmental stages.

Selecting a dataset and lncRNA displays the expression heatmap of lncRNA and mRNAs of correlated expression as shown in Figure 4a. Here, we show the coexpression of lncRNA NEAT1 using the dataset of 'Kidney chromophobe: Normal versus Tumor' from TCGA. The search can be performed for genes as well to identify lncRNAs of correlated expression. The list can be exported for further analysis.

In an effort to support functional interpretation of coexpressed genes, we implemented the gene set overrepresentation analysis for GO terms and KEGG pathways. Coexpressed genes whose correlation coefficients are above or below the cutoff value can be automatically subject to statistical enrichment test in specific GO terms or KEGG pathways. Figure 4b shows the sample output from the gene set analysis using top 200 protein-coding genes coexpressed with NEAT1. The result indicates that the complement activation and Rho signal transduction pathways are significant processes. The role of NEAT1 for the kidney tumor development in the context of associated pathways warrants further experimental works.

It is often the case that coexpressed genes are not necessarily connected to homogeneous molecular functions or processes, yielding presumably false terms in gene set overrepresentation analysis. We implemented a filtering scheme to discard the isolated genes using the protein–protein interactions (PPIs) in the REACTOME (version 47) database that included 1.827 million PPIs for human. Users may activate this PPI filtering procedure before applying the enrichment analysis of functional

Fig. 4. Coexpression analysis for NEAT1 in the dataset of 'Kidney chromophobe: Normal versus Tumor'. (**a**) Expression heatmap of lncRNA and coexpressed mRNAs. NEAT1 expression is shown on the top line. (**b**) Gene set analysis of GO terms for biological processes

annotation terms. It is shown that ~20% of coexpressed genes pass the REACTOME filtering condition (Supplementary Table S4).

3.5 Conserved lncRNAs with correlated expression

Cross-species conservation and correlated expression pattern are strong evidences for functional lncRNAs. To suggest the candidates of functional lncRNAs, we searched for lncRNAs that were conserved between human and orthologous genomes and whose expression patterns were highly correlated in orthologs. These lncRNAs are evolutionarily conserved in terms of sequences as well as expression patterns, thus being expected to play important biological roles.

Cabili *et al*. reported identification of orthologous lncRNAs by assembling and comparing transcript sequences from RNA-Seq (Cabili *et al*., 2011). In general, lncRNAs are known to be less conserved than protein-coding genes. We have investigated the sequence conservation across species using the UCSC phastCons track (Supplementary Fig. S3). For protein-coding genes, exons and 5′-UTR regions were better conserved than lncRNAs in both human and mouse. However, the sequence conservation of upstream (–500 bp of transcription start site) and intronic regions was comparable with the exonic region in lncRNAs. Thus, we used the genomewide multiple alignment of the UCSC phastCons track to identify orthologous lncRNAs instead of comparing transcriptome

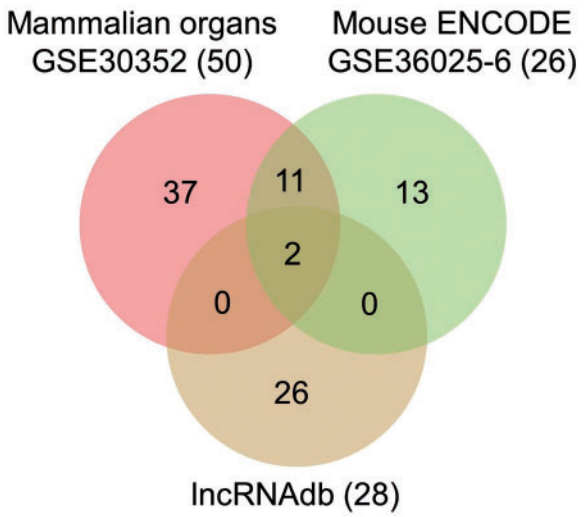

Fig. 5. Conserved lncRNAs with correlated gene expression between human and mouse. Entries from lncRNAdb are conserved ones with no information on expression pattern

sequences. Orthologous relations with many-to-many or antisense–sense correspondence were discarded. We obtained 628 non-human lncRNAs corresponding to 507 human lncRNAs.

Next, we searched for correlated expression pattern between orthologous lncRNAs. RNA-Seq data (GSE30352) for the polyadenylated RNA fraction of six organs from 10 species were used to estimate the expression correlation (Brawand *et al*., 2011). To remove biases from tissue-specific lncRNAs, we filtered out lncRNAs that were expressed in one or two tissues only. Using the cutoff of Pearson correlation coefficient >0.5, we have identified 72 conserved lncRNAs with correlated expression between human and orthologous non-human species, including 50 human–mouse cases. Among lncRNAs expressed in one or two tissues, we have identified 27 conserved lncRNAs additionally.

We applied the same strategy to independent expression datasets of multiple tissues in human and mouse. Using the Illumina human BodyMap 2.0 dataset that profiled 16 normal tissues (E-MTAB-513 in ArrayExpress) and the mouse ENCODE transcriptome data from CSHL and LICR (GSE36025, GSE36026) (Stamatoyannopoulos *et al*., 2012), we calculated the expresssion correlation in 11 common tissues. From these datasets, we found 26 conserved lncRNAs with correlation coefficient of expression >0.5.

Investigating lncRNAdb that collected literature-based information yielded only 28 conserved lncRNAs between human and mouse. Results from two independent datasets yielded 13 common members, only two of which were included in the annotated database of lncRNAdb (Fig. 5). These lncRNAs would serve as highly reliable candidates of functional lncRNAs, which warrant further investigation.

The 'Cons + Corr lncRNAs' menu in the web site shows our list of conserved lncRNAs with correlated expression. The scatterplot of expression correlation is available for detailed analysis.

4 CONCLUSION

Even if thousands of lncRNAs have been identified so far, their functional roles are known only for a limited number of lncRNAs. Thus, we need an efficient tool for inferring molecular expression and functions of lncRNAs to serve increasing number of scientists interested in this important class of non-coding RNAs.

Recently, three other groups released update databases on lncRNAs. NONCODEv4 (Xie *et al*., 2014) provides the gene expression pattern from RNA-Seq data. NPInter v2.0 (Yuan *et al*., 2014) and starBase v2.0 (Li *et al*., 2014) are databases of protein–RNA interactions based on CLIP-Seq data. Those high-throughput gene expression and protein–RNA interaction data were integrated together in the lncRNAtor database, enabling users to investigate diverse properties related to molecular functions. With the support of diverse features such as coexpression, differential expression and binding proteins, lncRNAtor would become a valuable resource on the role of lncRNAs to diverse groups of bench biologists. The coverage and predictive power of lncRNAtor are expected to increase as the regular update of the deep sequencing data in public. We plan to update the database on annual basis.

Funding: The research was supported by grants from the National Research Foundation of Korea [(NRF-2012M3A9D1054744, NRF-2012M3A9B9036673), Gwangju Institute of Science and Technology Systems Biology Infrastructure Establishment Grant through ERCSB and Ewha Global Top5 Grant of Ewha Womans University.

Conflict of Interest: none declared.

REFERENCES

Amaral,P.P. *et al.* (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res*., **39**, D146–D151.

Anders,S. *et al.* (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc*., **8**, 1765–1786.

Bhartiya,D. *et al.* (2013) lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. *Database*, **2013**, bat034.

Brawand,D. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Bu,D. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res*., **40**, D210–D215.

Cabili,M.N. *et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*., **25**, 1915–1927.

Derrien,T. *et al.* (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*., **22**, 1775–1789.

Dinger,M.E. *et al.* (2009) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res*., **37**, D122–D126.

Gutschner,T. and Diederichs,S. (2012) The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol*., **9**, 703–719.

Gutschner,T. *et al.* (2013) The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res*., **73**, 1180–1189.

Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*., **14**, R36.

Kong,L. *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*., **35**, W345–W349.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,Y. *et al.* (2013) RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments. *Nucleic Acids Res*., **41**, e94.

Li,J.H. *et al.* (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*., **42**, D92–D97.

Liao,Q. *et al.* (2011) Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene coexpression network. *Nucleic Acids Res*., **39**, 3864–3878.

Ling,H. *et al.* (2013) CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome Res*., **23**, 1446–1461.

Maruyama,R. and Suzuki,H. (2012) Long noncoding RNA involvement in cancer. *BMB Rep*., **45**, 604–611.

Ng,S.Y. *et al.* (2013) The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol. Cell*, **51**, 349–359.

Pollard,K.S. *et al.* (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*., **20**, 110–121.

Stamatoyannopoulos,J.A. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol*., **13**, 418.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*., **28**, 511–515.

Trapnell,C. *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol*., **31**, 46–53.

Volders,P.J. *et al.* (2013) LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res*., **41**, D246–D251.

Xie,C. *et al.* (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res*., **42**, D98–D103.

Yuan,J. *et al.* (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res*., **42**, D104–D108.