Bioinformatics tools for lncRNA research

Junichi Iwakiri, Michiaki Hamada, Kiyoshi Asai

PII: S1874-9399(15)00167-4
DOI: doi: 10.1016/j.bbagrm.2015.07.014
Reference: BBAGRM 915

To appear in: *BBA - Gene Regulatory Mechanisms*

Received date: 30 March 2015
Revised date: 7 July 2015
Accepted date: 14 July 2015

Please cite this article as: Junichi Iwakiri, Michiaki Hamada, Kiyoshi Asai, Bioinformatics tools for lncRNA research, *BBA - Gene Regulatory Mechanisms* (2015), doi: 10.1016/j.bbagrm.2015.07.014

Bioinformatics tools for lncRNA research

Junichi Iwakiri[a,b,*], Michiaki Hamada[c,b,*], Kiyoshi Asai[a,b,**]

[a] *Graduate School of Frontier Sciences, University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa 277–8562, Japan.*
[b] *Computational Biology Research Consortium (CBRC), in National Institute of Advanced Industrial Science and Technology (AIST), 2–41–6, Aomi, Koto-ku, Tokyo 135–0064, Japan.*
[c] *Department of Electrical Engineering and Bioscience, Faculty of Science and Engineering, Waseda University, 55N–06–10, 3–4–1, Okubo Shinjuku-ku, Tokyo 169–8555, Japan.*

Abstract

Current experimental methods to identify the functions of a large number of the candidates of long non-coding RNAs (lncRNAs) are limited in their throughput. Therefore, it is essential to know which tools are effective for understanding lncRNAs so that reasonable speed and accuracy can be achieved. In this paper, we review the currently available bioinformatics tools and databases that are useful for finding non-coding RNAs and analyzing their structures, conservation, interactions, co-expressions and localization.

Keywords: lncRNA, expression, mapping, conservation, secondary structure

1. Introduction

Recent high throughput sequencing technologies have enabled us to obtain a number of candidates of long non-coding RNAs (lncRNAs). However,

*Joint first authors
**Corresponding author: asai@k.u-tokyo.ac.jp

because the current experimental identification methods are still limited in their throughput, fast bioinformatics tools to identify and characterize lncRNAs with reasonable accuracy are required.

Because *non-coding* RNA is an exclusive category of RNAs that do not code for functional polypeptides, the first task for bioinformatics is to identify lncRNAs by screening long transcripts that do not seem to code for proteins. The objective of lncRNA research, however, is not only to find long *non-coding* RNAs but also to identify their functions. There are various bioinformatics tools for predicting the structures and functions of RNA sequences, including several tools that incorporate other experimental data in the analysis, but it is not obvious which tools are most useful for any particular objective.

It is known that structures, especially secondary structures, are important determinants of the functions of non-coding RNAs. It is also observed that genomic elements sharing similar functions are conserved between species. Therefore, secondary structures and their conservation are examined using bioinformatics tools to try to determine their functional categories. The predictions of secondary structures, however, are not always accurate. Nevertheless, although it is not always easy to extract concrete structural motifs related to functions, functional domains still may have structural features.

Important clues for the functions of lncRNAs, including *when*, *where* and *with what* they are used, can be extracted from experimental data. Spatiotemporal expression patterns (in tissues, subcellular compartments, and differentiation/developmental stages) by RNA-seq or microarray indicate *when* and *where* functions are activated. Co-expression analysis with

protein coding genes is useful for predicting *with what*, but a more direct way is to detect the interactions with proteins and other RNAs. Interactions with proteins may indicate the type of the function; furthermore, complementary bases in two RNA molecules often form base-pairs, giving high sequence specificity for the target RNAs of the functional RNAs. RNA–RNA interactions can be screened by searching reverse complementary subsequences, but precise analysis of structures both within and between RNA molecules is necessary for accurate prediction.

In this paper, we review available bioinformatics tools for research into lncRNAs, including their discovery, analyses and predictions of the secondary structures, conservation, interactions with other RNAs and proteins, co-expression with protein-coding genes, tissue-specificities, and subcellular localizations. We also consider useful databases.

2. Finding long non-coding RNAs

There are two steps in the identification of lncRNAs. In the first step, the transcribed units of the lncRNAs are identified. The fragments of the transcribed RNA sequences, observed by using next-generation sequencing (NGS) technologies or tiling microarrays, are mapped to the reference genome and summarized to obtain the transcribed units of the RNAs. The second step classifies the transcribed units as coding or non-coding: the sequences of transcribed units are evaluated on the basis of codon statistics and similarity to known protein sequences.

Before NGS technologies became available, however, it was common to *predict* candidates of (functional) non-coding RNAs on the basis of their

sequences and to experimentally verify their expression. For this prediction, conserved features (including secondary structures) of candidate sequences are considered. These analyses are still important for characterization of the functions of lncRNAs.

2.1. Identifying transcribed units of lncRNAs

Recent progress in NGS technology has enabled high throughput analysis of transcription in various types of cells. The obtained sequences are partial segments of the full-length transcripts. Those reads are mapped to a reference genome by using tools such as TOPHAT [1], LAST [2], and STAR [3]. The transcribed units of RNAs are determined from the mapped reads by using tools such as CUFFLINKS [4] and SCRIPTURE [5].

2.2. Evaluating coding potential of the transcripts

The approaches for discriminating between non-coding and coding sequences resemble the methods for gene discovery in genomic sequences. They are based on either similarity to known coding sequences or the statistics of codon frequencies specific to each organism. Similarity to known coding sequences is detected by using tools for homology search, typically BLASTX [6]. It should be noted that the existence of even a short functional segment of peptides (motifs) supports the identification of the transcript as coding. For novel lncRNAs, however, we cannot always expect conservation between species. The simplest way that does not rely on phylogenetic conservation is to check whether there is a long open reading frame (ORF) where no stop codon appears. However, it is necessary to evaluate any candidate segments

more carefully because there are short ORFs that encode functional peptides [7].

Among the tools for evaluating coding potential, CPC(Cording-Potential Calculator) [8] and PORTRAIT [9] use pairwise comparisons for the evaluation of coding potential; in contrast, PhyloCSF [10] and RNAcode [11] use multiple alignments to determine phylogenetically conserved features.

It is also possible to evaluate coding potential by using statistical features of the sequences of the transcripts, without using homology information from alignment to other sequences. Length of the ORF and codon usage bias are commonly used features. In its logistic regression, CPAT [8] uses ORF length, ORF coverage in the transcripts, Fickett TESTCODE score [12], and hexamer frequencies (reflecting codon usage bias and di-amino frequencies).

Igor Ulitsky reviews this important topic, including bioinformatics tools and experimental approaches, in detail in another article in this issue.

2.3. Conservation of sequences and structures

There are several tools for finding structurally conserved RNAs from multiple genome sequences, but simple multiple-alignment tools based on standard dynamic programming (DP) misalign most structured RNA sequences. More sophisticated tools include RNAz [13] and QRNA [14], which predict structurally conserved stable RNA secondary structures in multiple sequence alignments, both in non-coding RNAs and in cis-acting regulatory elements of mRNAs. EvoFold [15] also finds functional RNA structures in multiple sequence alignments by using a sophisticated probabilistic model (phylo-SCFG) for the substitution process in stem-pairing and unpaired regions. However, calculating multiple structural alignments of RNA sequences is computa-

tionally expensive when conservation of secondary structures is considered. $O(L^6)$ computational time is required, even for two sequences of length L [16], but there are a few fast tools that are described in Section 5.2, where consensus-based structure prediction using multiple structural alignments is also discussed.

3. Expression profiles of lncRNAs derived from RNA-seq data

Spatiotemporal expression patterns (tissues, subcellular compartments, and differentiation/developmental stages) of lncRNAs are fundamental information for understanding the biological functions of the lncRNAs in cells. There have been several RNA-seq studies obtaining the expression profiles of coding/non-coding genes across various tissues in many species [17, 18, 19, 20, 21, 22, 23, 24, 25]. But extracting expression patterns from RNA-seq data requires large-scale computational resources for quality filtering, reads mapping, and quantifying expressions.

Expression Atlas [26] processed various RNA-seq data using their original computational pipeline and have provided expression profiles of various tissues and cell lines derived from 16 species. Expression profiles of lncRNAs are provided for three of these species (human, mouse, and rat) in their database (Table 1). In particular, for humans, Expression Atlas provides expression profiles of lncRNAs derived from not only 32 normal tissues [17] but also 675 cancer cell lines [20]. Expression profiles of protein-coding genes derived from several tissues of other vertebrates—such as opossum, rhesus monkey, olive baboon, chicken, cattle, pufferfish, frog, and lizards—barley, and rice are also available in Expression Atlas, but expression profiles of

their lncRNAs are not available because of the lack of annotations. Recently, The GTEx consortium has provided an extensive collection of RNA-seq data derived from 43 human tissues (including 11 brain subregions) for 53,934 genes in which the expression profiles of lncRNAs were included. [27, 28]

3.1. Tissue-specificity

Tissue-specificity of lncRNA expressions are also important features for characterizing lncRNAs. Recent analysis of RNA-seq data derived from 24 human tissues [29] revealed that the majority of lincRNAs (approx. 80%) exhibit tissue-specific expression patterns, whereas such expression patterns are observed in a much smaller fraction of protein-coding genes (approx. 20%). Washietl et al. [30] analyzed the evolutionary dynamics of lncRNA expression patterns and tissue-specificity in nine tissues across six mammals, and showed that the tissue-specificity of mammalian lncRNAs is highly conserved. This tissue-specificity of lncRNA expression patterns is informative for identifying their tissue-specific functions.

For investigating the tissue-specificity of expression patterns of coding or non-coding genes from microarray or RNA-seq data, several measures based on Shannon entropy have been proposed [29, 31, 32, 33]. Those measures, however, do not directly specify a small number of tissues with high expression levels because the Shannon entropy of each gene gives the degree of distortion from the ubiquitous expression pattern over all tissues. ROKU [31] is a useful tool for detecting actual tissues with high or low expression levels as outliers, and simultaneously evaluating overall tissue-specificity by calculating the entropy.

3.2. Subcellular Localization

Subcellular localization of lncRNAs is another important factor controlling macromolecular interactions, such as lncRNA–RNA, lncRNA–chromatin (DNA) and lncRNA–protein interactions, in a cell. As resources for investigations of the subcellular localization of human lncRNAs, the ENCODE project has provided the RNA-seq data obtained from each of two subcellular compartments (nucleus and cytosol) in 15 human cell lines [21]. In addition, RNA-seq data obtained from three subnuclear components (nucleoplasm, nucleolus and chromatin) in the K562 cell line were also provided. Their analysis of several human cell lines revealed that more lncRNAs are found in the nucleus than in the cytosol. Expression Atlas has also processed these RNA-seq data to provide the expression profiles of all human genes, including lncRNA genes (Table 1). In addition, these ENCODE RNA-seq data are also available in the UCSC genome browser, which provides convenient access [34, 35]. These datasets offer basic information for accurate predictions of macromolecular interactions involving lncRNAs.

4. Macromolecular interactions involving lncRNAs

Identification of interaction targets (RNAs or proteins) of lncRNAs is a popular approach to determining their functions. Several high-throughput experimental methods, such as RIA-seq [36], RAP-RNA [37], PAR-CLIP [38] and HITS-CLIP [39], have been proposed for investigating RNA–RNA or RNA–protein interactions. In these methods, antisense-probing or antibody-based immunoprecipitation are the key procedures for purifying a *bait* RNA or protein of interest. However, these necessary steps narrow down the scale

of the investigation from all-to-all to one-to-all. Thus, applying these methods to each of a large number of lncRNAs for comprehensive identification of the all-to-all interactome is labor-intensive. This limitation demands the development of computational prediction or screening of lncRNA–RNA and lncRNA–protein interactions.

4.1. Prediction methods for lncRNA–RNA interactions

Computational predictions of RNA–RNA interactions are based on the *interaction energy* that is estimated from the inter-molecular and intra-molecular base-pairing interactions of two RNA molecules. IntaRNA [40] is a tool for predicting RNA–RNA interactions using only the primary sequences of two RNAs. For each input pair of RNAs, IntaRNA provides an *interaction energy* estimated by subtracting unfolding energies (based on intra-molecular base-pairs within each of the two RNAs) from a hybridization energy (based on inter-molecular base-pairs between the two RNAs). An improvement of predictions of RNA–RNA interactions over those of IntaRNA was achieved by CopraRNA [41], which incorporates a comparative genomics approach using at least three genomic sequences of distinct species. However, these methods have not been applied to the prediction of lncRNA–RNA interactions, because these methods focus on the prediction of bacterial small RNA (sRNA)–mRNA interactions.

Recently, the authors developed a computational pipeline including various computational sequence analysis tools (Raccess [42], TanTan [43], LAST [2, 44], IntaRNA [40], and RactIP [45]) for predicting human lncRNA–RNA interactions [46], and implemented this pipeline on the K computer, which is one of the fastest super-computers in the world. The database

of all the predicted human lncRNA–RNA interactions contains the lncRNA–mRNA and lncRNA–lncRNA interactions for 23,898 lncRNAs and 20,185 mRNAs (available at http://rtools.cbrc.jp/cgi-bin/RNARNA/index.pl).

4.2. Prediction methods for lncRNA–protein interactions

Several computational methods have been developed for predicting RNA–protein interactions [47, 48, 49, 50, 51]. In these methods, machine learning approaches, such as Fisher's linear discriminant analysis (LDA), support vector machine (SVM), and random forest (RF), were used to discriminate the interacting RNA–protein pairs from non-interacting pairs. In terms of input data, these methods are categorized into three groups: sequence-based methods, sequence and structure-based methods, and experimental-data-based methods.

RPI-SEQ [47], CATRAPID [48], and LNCPRO [49] are sequence-based methods that require only the primary sequences of the RNA and the protein for the input data. Among these three methods, CATRAPID and LNCPRO use physicochemical properties of amino acids and nucleotides, and predict the secondary structures of proteins and RNAs as the features of interacting or non-interacting RNA–protein pairs. These two methods focus on the prediction of lncRNA–protein interactions, and were benchmarked by using a few experimentally validated lncRNA–protein interactions, including HOTAIR and XIST lncRNA. The third method, RPI-PRED [50], is a tool for predicting RNA–protein interactions using not only primary sequences but also the three dimensional structures of an RNA and a protein for the input data. In this method, structural motifs of proteins (called Protein Blocks) and RNA secondary structures were extracted from their three dimensional

structures, and were used as the features for discriminating interacting RNA–protein pairs from non-interacting pairs. However, using this method for predicting lncRNA–protein interactions would be difficult because the structural information is currently available for only a few lncRNAs.

Pancaldi and Bahler developed a prediction method that uses various experimental data, such as protein localization, RNA half-life, ribosome-profiling, and PARS analysis, for predicting mRNA–protein interactions in yeast [51]. However, applying this method to the prediction of lncRNA–protein interactions is not easy because experimental data is available for only a few lncRNAs.

In all of these prediction methods, pair-input data comprising several features of the RNA and the protein are required for the machine learning algorithms. To evaluate their prediction performances by using cross-validation methods, the contents of entire datasets were randomly divided into training sets and test sets for benchmarking. However, Park and Marcotte indicated that the reported prediction performances are significantly dependent on inappropriate cross-validations [52]; better prediction performances were achieved only when the test set shared the same components (RNAs or proteins) as the training set. In conclusion, predicting lncRNA–protein interactions by using bioinformatics tools is still a challenge.

5. Tools for analyzing structures of lncRNAs

It is known that not only the primary sequences of functional non-coding RNAs but also their structures are closely related to their functions. While the relations between structure and function have been well-studied for short

non-coding RNAs (e.g., [53]), a few studies have suggested that secondary structures are also important for the functions of lncRNAs [54, 55, 56]. In this section, tools for the prediction of lncRNA structures (cf. Figure 1) are reviewed. *Secondary* structures, *consensus* secondary structures, *tertiary* structures and *joint* secondary structures are mainly considered.

5.1. RNA secondary structure of single RNA sequence

RNA *secondary* structures are an abstract form of RNA structure; they are represented by sets of interacting base pairs. In this subsection, we review the tools for RNA secondary structure predictions that are applicable to lncRNAs.

5.1.1. In-silico RNA structure predictions

In-silico RNA secondary structure predictions are based on the *free energy* of the structures, which are computed by using experimentally determined energy parameters. There are too many tools for predicting RNA secondary structures from a single RNA sequence to list here, but they include CentroidFold [57, 58], Mfold [59], RNAfold [60], and RNAstructure [61, 62]. The time complexities of the methods used by those tools are equal to $O(L^3)$, where L is the length of RNA sequence. Therefore, it might be difficult to apply them to lncRNAs longer than several thousands of bases. Rfold [63] reduces this difficulty by limiting the maximal span of base pairs to w, which leads to an $O(w^2L)$ computational time. Because $O(w^2L)$ is linear with respect to the sequence length L, Rfold is applicable to lncRNAs, such as Xist and NEAT1. Secondary structure predictions of lncRNAs, however, may also lack accuracy. RNA secondary structure predic-

tions achieve reliable accuracy for RNAs with lengths less than 1000, but the accuracy is often unsatisfactory for longer RNAs. Accuracy is improved with the aid of the homologous sequences of the subject lncRNAs because the local secondary structures are often evolutionarily conserved [30]. CentroidHomfold [64, 65] utilizes homologous sequences of the subject (lnc)RNA and shows improved accuracy.

The above tools predict RNA secondary structures without *pseudoknots*. Predictions of RNA secondary structures with pseudoknots using the exact algorithms entail large computational costs; but faster approximate tools, such as IPKnot [66], are practically applicable to lncRNAs.

5.1.2. Probing-directed RNA structure predictions

Recently, several experimental methods for high-throughput RNA structure determination techniques have been proposed (Table 2). These techniques, which rely on probing technology, measuring the strand flexibility of an RNA sequence, include FragSeq [67], SHAPE-seq [68], DMS-seq [69, 70], Mod-seq [71], MAP-seq [72], and PARS [73]. (See also [74] for an extensive review for those methods.) Unfortunately, these experimental techniques provide only partial information about the secondary structures. Specifically, they give us only the preferences for single strandedness for each nucleotide. The tools to construct RNA secondary structures from experimental data, are classified as follows. (i) Sampling-based methods, such as SeqFold [75], where RNA secondary structures are predicted from the set of suboptimal RNA structures given by a sampling procedure using, e.g., Sfold [76]. (ii) Energy-based approaches, where energy calculations are modified using pseudo-energy [77, 78] or energy parameters are re-estimated [79] using the

information from experimental data. (iii) An approach for modifying a base-pairing probability matrix (BPPM) [1] under the constraints of experimental probing data [80].

5.2. Consensus secondary structures of RNA sequences

Because RNA secondary structures related to a specific function are evolutionarily conserved, detecting those conserved secondary structures is useful step toward the functional analysis of lncRNAs. In predictions of conserved secondary structures, co-variation of bases to maintain a base-pair (e.g., `G-C` to `A-U`) is utilized. As shown in a study of the evolutionary dynamics of lncRNAs in six mammals [30], evolutionary information about lncRNAs is useful for predicting structures of lncRNAs. Tools such as CentroidAlifold [81] and RNAAlifold [82] predict the evolutionarily conserved consensus secondary structures from multiple alignments of RNA sequences [83]. Additionally, methods for probing-directed RNA secondary structure prediction (Section 5.1.2) are also applicable to consensus secondary structure predictions: PPfold 3.0 [84] incorporates SHAPE data in predicting common secondary structures.

The input for these tools for consensus secondary structure prediction are multiple alignments of RNA sequences, which can be provided by using conventional alignment tools, such as ProbCons [85], or state-of-the-art fast tools that consider secondary structures during the alignment process, such as MAFFT [86], LocARNA [87], LARA [88], MXSCARNA [89] and CentroidAlign [90]. $O(L^3)$ computations are still required for these tools,

[1] The BPPM provides probabilities for forming each base-pairs in a given RNA sequence.

which might be too expensive for lncRNAs; however, a recent update of CentroidAlign, which internally utilizes Rfold described in Section 5.1.1, has enabled us to apply CentroidAlign to longer RNA sequences, including lncRNAs [91].

Note that multiple alignment is also important in lncRNA gene discovery, as described in Section 2.

5.3. Joint secondary structures between two RNA sequences

As described in Section 4, many lncRNAs interact with the other molecules (DNA/chromatin, RNA and protein) in the cells, and predictions of complexes involving lncRNAs are useful for functional analyses of the lncRNAs. Predictions of joint secondary structures of two RNA sequences consist of predictions of inter- and intra-molecular base-pairs. However, simultaneous prediction of both inter- and intra-molecular base-pairs generally entails huge computational costs. Nevertheless, several useful tools exist. RactIP [45] rapidly predicts joint secondary structures (i.e., interactions between two RNA sequences) including both inter- and intra-molecular base-pairs, while IntaRNA [40] discards intra-molecular base-pairs in its prediction. Additionally, PETcofold [92] predicts the *conserved* joint secondary structures of pairs of multiple alignments of RNA sequences.

5.4. Tertiary structure predictions

Tertiary structure predictions of RNA sequences require much larger computational resources than RNA secondary structure predictions. See e.g. [93] for a review of tools for tertiary structure predictions. Unfortunately, it is computationally infeasible to predict tertiary structures for entire lncRNA

sequences, and more efficient tools are necessary for applications involving lncRNAs.

5.5. Other tools or methods related to RNA secondary structures

5.5.1. Mutation analyses of RNA secondary structures for lncRNAs

Recent studies have suggested the importance of single nucleotide polymorphisms (SNPs) that alter RNA secondary structures, called "RiboSNitches" [94], which might be the causes of diseases. RCHANGE [95] and other tools [96] are promising in-silico tools for mutational analyses of lncRNAs. In particular, RCHANGE simultaneously analyzes the mutational effect of SNPs at every position in a lncRNA, which is difficult to do using experimental methods.

5.5.2. Analysis of RNA structural motifs

In order to clarify the functions of lncRNAs with respect to secondary structures, the discovery of secondary structural *motifs* that correlate with functions is important. CAPR [97] computes detailed structural profiles, including not only the base-pairs but also loops, bulges, multi-loops and external loops. CAPR revealed that the binding site in ncRNAs, with which RNA-binding proteins interact, have specific patterns of structural profile depending on the binding protein. RNACONTEXT [98] and MEMERIS [99] (an extension of MEME that is used for the discovery of sequence motifs) find short secondary structural motifs among several RNA sequences expected to have similar functions.

Additionally, the local *accessibility* of lncRNA sequences, which can be computed by RACCESS [42], might be useful in predictions of targets of miR-

NAs and molecules interacting with lncRNAs (cf. [100]).

5.6. Summary of this section

In this section, we briefly reviewed tools that will be useful for prediction of the structure of lncRNAs. Due to space limitations, we did not give detailed methods for each tool. A recent review by Eddy [101] is a good source for probing directed RNA secondary structure predictions. Moreover, two recently-published books on Methods in Molecular Biology [102, 103] will be useful for further study of structures of lncRNAs.

6. LncRNA Databases

Emerging evidence of experimentally identified lncRNAs and their biological properties, including genomic features, expression profiles, sequence conservation, macromolecular interactions, epigenomic modifications, and functional annotations, needs to be organized into public databases as a resource for lncRNA research. Currently, several databases specializing in lncRNAs have been developed, and provide a variety of information, as described in Table 3.

LNCipedia [104] and lncRNome [105] provide a large number ($\geq 100,000$) of human lncRNA entries including primary sequences and predicted secondary structures. In LNCipedia the protein-coding potentials of lncRNAs are assessed using bioinformatic tools [9, 106, 10] and ribosome-profiling data [107]. In addition, LNCipedia provides predicted lncRNA–miRNA interactions [108], which are useful for finding competing endogenous RNAs (ceRNAs) [109]. The single nucleotide polymorphisms (SNPs) and the epigenomic modifications of lncRNA genes are uniquely included in lncRNome.

Expression profiles of lncRNAs derived from human tissues and cell lines are included in LNCRNADB [110] and LNCRNATOR [111]. LNCRNATOR provides three unique pieces of information regarding the expression data of lncRNAs: expression levels of lncRNAs in cancer tissues derived from RNA-seq data in The Cancer Genome Atlas (TCGA), co-expression analysis of lncRNA and protein-coding genes, and gene ontology (GO) enrichment analysis for the co-expressed genes. LNCRNADB provides detailed descriptions of the functions of 287 lncRNAs that were manually curated from recent literature and also provides a list of orthologous lncRNAs across several species with links to the UCSC genome browser. For instance, NEAT1 lncRNA orthologues are observed in five mammal (human, rat, dog, mouse, and cattle) genomes. lncRNA–protein interactions obtained from various CLIP-seq studies are provided by LNCRNOME and LNCRNATOR.

Acknowledgement

This work was supported in part by MEXT KAKENHI (Grant-in-Aid for Young Scientists (A) Grant Number 24680031 for MH; Grant-in-Aid for Scientific Research (A) Grant Number 25240044 for MH and KA; Grant-in-Aid for Scientific Research on Innovative Areas Grant Number 221S0002) to KA.

References

[1] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (2013) R36.

[2] S. M. Kiebasa, R. Wan, K. Sato, P. Horton, M. C. Frith, Adaptive seeds tame genomic sequence comparison, Genome Res. 21 (2011) 487–493.

[3] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21.

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

[5] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, A. Regev, Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, Nat. Biotechnol. 28 (2010) 503–510.

[6] W. Gish, D. J. States, Identification of protein coding regions by database similarity search, Nat. Genet. 3 (1993) 266–272.

[7] S. J. Andrews, J. A. Rothnagel, Emerging evidence for functional peptides encoded by short open reading frames, Nat. Rev. Genet. 15 (2014) 193–204.

[8] L. Wang, H. J. Park, S. Dasari, S. Wang, J. P. Kocher, W. Li, CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model, Nucleic Acids Res. 41 (2013) e74.

[9] L. Kong, Y. Zhang, Z. Q. Ye, X. Q. Liu, S. Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res. 35 (2007) W345–349.

[10] M. F. Lin, I. Jungreis, M. Kellis, PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions, Bioinformatics 27 (2011) i275–282.

[11] S. Washietl, S. Findeiss, S. A. Muller, S. Kalkhof, M. von Bergen, I. L. Hofacker, P. F. Stadler, N. Goldman, RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data, RNA 17 (2011) 578–594.

[12] J. W. Fickett, C. S. Tung, Assessment of protein coding measures, Nucleic Acids Res. 20 (1992) 6441–6450.

[13] A. R. Gruber, S. Findeiss, S. Washietl, I. L. Hofacker, P. F. Stadler, RNAz 2.0: improved noncoding RNA detection, Pac Symp Biocomput (2010) 69–79.

[14] E. Rivas, S. R. Eddy, Noncoding RNA gene detection using comparative sequence analysis, BMC Bioinformatics 2 (2001) 8.

[15] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, D. Haussler, Identification and classification of conserved RNA secondary structures in the human genome, PLoS Comput. Biol. 2 (2006) e33.

[16] K. Asai, M. Hamada, RNA structural alignments, part II: non-Sankoff approaches for structural alignments, Methods Mol. Biol. 1097 (2014) 291–301.

[17] M. Uhlen, L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling,

H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten, Proteomics. Tissue-based map of the human proteome, Science 347 (2015) 1260419.

[18] C. M. Farrell, N. A. O'Leary, R. A. Harte, J. E. Loveland, L. G. Wilming, C. Wallin, M. Diekhans, D. Barrell, S. M. Searle, B. Aken, S. M. Hiatt, A. Frankish, M. M. Suner, B. Rajput, C. A. Steward, G. R. Brown, R. Bennett, M. Murphy, W. Wu, M. P. Kay, J. Hart, J. Rajan, J. Weber, C. Snow, L. D. Riddick, T. Hunt, D. Webb, M. Thomas, P. Tamez, S. H. Rangwala, K. M. McGarvey, S. Pujar, A. Shkeda, J. M. Mudge, J. M. Gonzalez, J. G. Gilbert, S. J. Trevanion, R. Baertsch, J. L. Harrow, T. Hubbard, J. M. Ostell, D. Haussler, K. D. Pruitt, Current status and new features of the Consensus Coding Sequence database, Nucleic Acids Res. 42 (2014) D865–872.

[19] D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csardi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grutzner, S. Bergmann, R. Nielsen, S. Paabo, H. Kaessmann, The evolution of gene expression levels in mammalian organs, Nature 478 (2011) 343–348.

[20] C. Klijn, S. Durinck, E. W. Stawiski, P. M. Haverty, Z. Jiang, H. Liu, J. Degenhardt, O. Mayba, F. Gnad, J. Liu, G. Pau, J. Reeder, Y. Cao, K. Mukhyala, S. K. Selvaraj, M. Yu, G. J. Zynda, M. J. Brauer, T. D. Wu, R. C. Gentleman, G. Manning, R. L. Yauch, R. Bourgon, D. Stokoe, Z. Modrusan, R. M. Neve, F. J. de Sauvage, J. Settleman, S. Seshagiri, Z. Zhang, A comprehensive transcriptional portrait of human cancer cell lines, Nat. Biotechnol. 33 (2015) 306–312.

[21] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, Landscape of transcription in human cells, Nature 489 (2012) 101–108.

[22] J. Merkin, C. Russell, P. Chen, C. B. Burge, Evolutionary dynamics of gene and isoform regulation in Mammalian tissues, Science 338 (2012) 1593–1599.

[23] T. M. Keane, L. Goodstadt, P. Danecek, M. A. White, K. Wong, B. Yalcin, A. Heger, A. Agam, G. Slater, M. Goodson, N. A. Furlotte, E. Eskin, C. Nellaker, H. Whitley, J. Cleak, D. Janowitz, P. Hernandez-Pliego, A. Edwards, T. G. Belgard, P. L. Oliver, R. E. McIntyre, A. Bhomra, J. Nicod, X. Gan, W. Yuan, L. van der Weyden, C. A. Steward, S. Bala, J. Stalker, R. Mott, R. Durbin, I. J. Jackson, A. Czechanski, J. A. Guerra-Assuncao, L. R. Donahue, L. G. Reinholdt, B. A. Payseur, C. P. Ponting, E. Birney, J. Flint, D. J. Adams, Mouse genomic variation and its effect on phenotypes and gene regulation, Nature 477 (2011) 289–294.

[24] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, B. J. Blencowe, The evolutionary landscape of alternative splicing in vertebrate species, Science 338 (2012) 1587–1593.

[25] Y. Yu, J. C. Fuscoe, C. Zhao, C. Guo, M. Jia, T. Qing, D. I. Bannon, L. Lancashire, W. Bao, T. Du, H. Luo, Z. Su, W. D. Jones, C. L. Moland, W. S. Branham, F. Qian, B. Ning, Y. Li, H. Hong, L. Guo, N. Mei, T. Shi, K. Y. Wang, R. D. Wolfinger, Y. Nikolsky, S. J. Walker, P. Duerksen-Hughes, C. E. Mason, W. Tong, J. Thierry-Mieg, D. Thierry-Mieg, L. Shi, C. Wang, A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages, Nat Commun 5 (2014) 3230.

[26] R. Petryszak, T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, J. McMurry, J. C. Marioni, J. Malone, K. Megy, G. Rustici, A. Y. Tang, J. Taubert, E. Williams, O. Mannion, H. E. Parkinson, A. Brazma, Expression Atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments, Nucleic Acids Res. 42 (2014) D926–932.

[27] K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y. H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo,

M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, E. T. Dermitzakis, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, Science 348 (2015) 648–660.

[28] M. Mele, P. G. Ferreira, F. Reverter, D. S. DeLuca, J. Monlong, M. Sammeth, T. R. Young, J. M. Goldmann, D. D. Pervouchine, T. J. Sullivan, R. Johnson, A. V. Segre, S. Djebali, A. Niarchou, F. A. Wright, T. Lappalainen, M. Calvo, G. Getz, E. T. Dermitzakis, K. G. Ardlie, R. Guigo, Human genomics. The human transcriptome across tissues and individuals, Science 348 (2015) 660–665.

[29] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev,

J. L. Rinn, Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses, Genes Dev. 25 (2011) 1915–1927.

[30] S. Washietl, M. Kellis, M. Garber, Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals, Genome Res. 24 (2014) 616–628.

[31] K. Kadota, J. Ye, Y. Nakai, T. Terada, K. Shimizu, ROKU: a novel method for identification of tissue-specific genes, BMC Bioinformatics 7 (2006) 294.

[32] J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, C. J. Stoeckert, Promoter features related to tissue specificity as measured by Shannon entropy, Genome Biol. 6 (2005) R33.

[33] S. Gerstberger, M. Hafner, T. Tuschl, A census of human RNA-binding proteins, Nat. Rev. Genet. 15 (2014) 829–845.

[34] K. R. Rosenbloom, C. A. Sloan, V. S. Malladi, T. R. Dreszer, K. Learned, V. M. Kirkup, M. C. Wong, M. Maddren, R. Fang, S. G. Heitner, B. T. Lee, G. P. Barber, R. A. Harte, M. Diekhans, J. C. Long, S. P. Wilder, A. S. Zweig, D. Karolchik, R. M. Kuhn, D. Haussler, W. J. Kent, ENCODE data in the UCSC Genome Browser: year 5 update, Nucleic Acids Res. 41 (2013) 56–63.

[35] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.

[36] M. Kretz, Z. Siprashvili, C. Chu, D. E. Webster, A. Zehnder, K. Qu, C. S. Lee, R. J. Flockhart, A. F. Groff, J. Chow, D. Johnston, G. E. Kim, R. C.

Spitale, R. A. Flynn, G. X. Zheng, S. Aiyer, A. Raj, J. L. Rinn, H. Y. Chang, P. A. Khavari, Control of somatic tissue differentiation by the long non-coding RNA TINCR, Nature 493 (2013) 231–235.

[37] J. M. Engreitz, K. Sirokman, P. McDonel, A. A. Shishkin, C. Surka, P. Russell, S. R. Grossman, A. Y. Chow, M. Guttman, E. S. Lander, RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites, Cell 159 (2014) 188–199.

[38] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, T. Tuschl, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, Cell 141 (2010) 129–141.

[39] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, R. B. Darnell, HITS-CLIP yields genome-wide insights into brain alternative RNA processing, Nature 456 (2008) 464–469.

[40] A. Busch, A. S. Richter, R. Backofen, IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions, Bioinformatics 24 (2008) 2849–2856.

[41] P. R. Wright, A. S. Richter, K. Papenfort, M. Mann, J. Vogel, W. R. Hess, R. Backofen, J. Georg, Comparative genomics boosts target prediction for bacterial small RNAs, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) E3487–3496.

[42] H. Kiryu, G. Terai, O. Imamura, H. Yoneyama, K. Suzuki, K. Asai, A

detailed investigation of accessibilities around target sites of siRNAs and miRNAs, Bioinformatics 27 (2011) 1788–1797.

[43] M. C. Frith, A new repeat-masking method enables specific detection of homologous sequences, Nucleic Acids Res. 39 (2011) e23.

[44] M. C. Frith, M. Hamada, P. Horton, Parameters for accurate genome alignment, BMC Bioinformatics 11 (2010) 80.

[45] Y. Kato, K. Sato, M. Hamada, Y. Watanabe, K. Asai, T. Akutsu, RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming, Bioinformatics 26 (2010) i460–466.

[46] G. Terai, J. Iwakiri, T. Kameda, M. Hamada, K. Asai, Comprehensive prediction of lncRNA-RNA interactions in human transcriptome, 2015. (submitted).

[47] U. K. Muppirala, V. G. Honavar, D. Dobbs, Predicting RNA-protein interactions using only sequence information, BMC Bioinformatics 12 (2011) 489.

[48] M. Bellucci, F. Agostini, M. Masin, G. G. Tartaglia, Predicting protein associations with long noncoding RNAs, Nat. Methods 8 (2011) 444–445.

[49] Q. Lu, S. Ren, M. Lu, Y. Zhang, D. Zhu, X. Zhang, T. Li, Computational prediction of associations between long non-coding RNAs and proteins, BMC Genomics 14 (2013) 651.

[50] V. Suresh, L. Liu, D. Adjeroh, X. Zhou, RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information, Nucleic Acids Res. 43 (2015) 1370–1379.

[51] V. Pancaldi, J. Bahler, In silico characterization and prediction of global protein-mRNA interactions in yeast, Nucleic Acids Res. 39 (2011) 5826–5836.

[52] Y. Park, E. M. Marcotte, Flaws in evaluation schemes for pair-input computational predictions, Nat. Methods 9 (2012) 1134–1136.

[53] E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, R. D. Finn, Rfam 12.0: updates to the RNA families database, Nucleic Acids Res. 43 (2015) D130–137.

[54] I. V. Novikova, S. P. Hennelly, K. Y. Sanbonmatsu, Structural architecture of the human long non-coding RNA, steroid receptor RNA activator, Nucleic Acids Res. 40 (2012) 5034–5051.

[55] S. Maenner, M. Blaud, L. Fouillen, A. Savoye, V. Marchand, A. Dubois, S. Sanglier-Cianferani, A. Van Dorsselaer, P. Clerc, P. Avner, A. Visvikis, C. Branlant, 2-D structure of the A region of Xist RNA and its implication for PRC2 association, PLoS Biol. 8 (2010) e1000276.

[56] M. C. Tsai, O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal, H. Y. Chang, Long noncoding RNA as modular scaffold of histone modification complexes, Science 329 (2010) 689–693.

[57] K. Sato, M. Hamada, K. Asai, T. Mituyama, CENTROIDFOLD: a web server for RNA secondary structure prediction, Nucleic Acids Res. 37 (2009) W277–280.

[58] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, K. Asai, Prediction of RNA

secondary structure using generalized centroid estimators, Bioinformatics 25 (2009) 465–473.

[59] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (2003) 3406–3415.

[60] R. Lorenz, S. H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker, ViennaRNA Package 2.0, Algorithms Mol Biol 6 (2011) 26.

[61] D. H. Mathews, RNA Secondary Structure Analysis Using RNAstructure, Curr Protoc Bioinformatics 46 (2014) 1–12.

[62] D. H. Mathews, Using the RNAstructure Software Package to Predict Conserved RNA Structures, Curr Protoc Bioinformatics 46 (2014) 1–12.

[63] H. Kiryu, T. Kin, K. Asai, Rfold: an exact algorithm for computing local base pairing probabilities, Bioinformatics 24 (2008) 367–373.

[64] M. Hamada, K. Sato, H. Kiryu, T. Mituyama, K. Asai, Predictions of RNA secondary structure by combining homologous sequence information, Bioinformatics 25 (2009) i330–338.

[65] M. Hamada, K. Yamada, K. Sato, M. C. Frith, K. Asai, CentroidHomfold-LAST: accurate prediction of RNA secondary structure using automatically collected homologous sequences, Nucleic Acids Res. 39 (2011) W100–106.

[66] K. Sato, Y. Kato, M. Hamada, T. Akutsu, K. Asai, IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming, Bioinformatics 27 (2011) 85–93.

[67] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, D. Haussler, FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing, Nat. Methods 7 (2010) 995–1001.

[68] D. Loughrey, K. E. Watters, A. H. Settle, J. B. Lucks, SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing, Nucleic Acids Res. 42 (2014).

[69] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, S. M. Assmann, In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features, Nature 505 (2014) 696–700.

[70] S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, J. S. Weissman, Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo, Nature 505 (2014) 701–705.

[71] J. Talkish, G. May, Y. Lin, J. L. Woolford, C. J. McManus, Mod-seq: high-throughput sequencing for chemical probing of RNA structure, RNA 20 (2014) 713–720.

[72] M. G. Seetin, W. Kladwang, J. P. Bida, R. Das, Massively parallel RNA chemical mapping with a reduced bias MAP-seq protocol, Methods Mol. Biol. 1086 (2014) 95–117.

[73] Y. Wan, K. Qu, Q. C. Zhang, R. A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R. C. Spitale, M. P. Snyder, E. Segal, H. Y. Chang, Landscape and variation of RNA secondary structure across the human transcriptome, Nature 505 (2014) 706–709.

[74] S. A. Mortimer, M. A. Kidwell, J. A. Doudna, Insights into RNA structure and function from genome-wide studies, Nat. Rev. Genet. 15 (2014) 469–479.

[75] Z. Ouyang, M. P. Snyder, H. Y. Chang, SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data, Genome Res. 23 (2013) 377–387.

[76] C. Y. Chan, C. E. Lawrence, Y. Ding, Structure clustering features on the Sfold Web server, Bioinformatics 21 (2005) 3926–3928.

[77] K. E. Deigan, T. W. Li, D. H. Mathews, K. M. Weeks, Accurate SHAPE-directed RNA structure determination, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 97–102.

[78] K. Zarringhalam, M. M. Meyer, I. Dotu, J. H. Chuang, P. Clote, Integrating chemical footprinting data into RNA secondary structure prediction, PLoS ONE 7 (2012) e45160.

[79] S. Washietl, I. L. Hofacker, P. F. Stadler, M. Kellis, RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction, Nucleic Acids Res. 40 (2012) 4261–4272.

[80] M. Hamada, Direct updating of an RNA base-pairing probability matrix with marginal probability constraints, J. Comput. Biol. 19 (2012) 1265–1276.

[81] M. Hamada, K. Sato, K. Asai, Improving the accuracy of predicting secondary structure for aligned RNA sequences, Nucleic Acids Res. 39 (2011) 393–402.

[82] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, P. F. Stadler, RNAalifold: improved consensus structure prediction for RNA alignments, BMC Bioinformatics 9 (2008) 474.

[83] M. Hamada, RNA secondary structure prediction from multi-aligned sequences, Methods Mol. Biol. 1269 (2015) 17–38.

[84] Z. Sukosd, B. Knudsen, J. Kjems, C. N. Pedersen, PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data, Bioinformatics 28 (2012) 2691–2692.

[85] C. B. Do, M. S. Mahabhashyam, M. Brudno, S. Batzoglou, ProbCons: Probabilistic consistency-based multiple sequence alignment, Genome Res. 15 (2005) 330–340.

[86] K. Katoh, H. Toh, Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework, BMC Bioinformatics 9 (2008) 212.

[87] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, R. Backofen, LocARNA-P: accurate boundary prediction and improved detection of structural RNAs, RNA 18 (2012) 900–914.

[88] M. Bauer, G. W. Klau, K. Reinert, Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization, BMC Bioinformatics 8 (2007) 271.

[89] Y. Tabei, H. Kiryu, T. Kin, K. Asai, A fast structural multiple alignment method for long RNA sequences, BMC Bioinformatics 9 (2008) 33.

[90] M. Hamada, K. Sato, H. Kiryu, T. Mituyama, K. Asai, CentroidAlign: fast and accurate aligner for structured RNAs by maximizing expected sum-of-pairs score, Bioinformatics 25 (2009) 3236–3243.

[91] H. Yonemoto, K. Asai, M. Hamada, CentroidAlign-Web: A Fast and Accurate Multiple Aligner for Long Non-Coding RNAs, Int J Mol Sci 14 (2013) 6144–6156.

[92] S. E. Seemann, P. Menzel, R. Backofen, J. Gorodkin, The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences, Nucleic Acids Res. 39 (2011) W107–111.

[93] K. Rother, M. Rother, P. Skiba, J. M. Bujnicki, Automated modeling of RNA 3D structure, Methods Mol. Biol. 1097 (2014) 395–415.

[94] M. Halvorsen, J. S. Martin, S. Broadaway, A. Laederach, Disease-associated mutations that alter the RNA structural ensemble, PLoS Genet. 6 (2010) e1001074.

[95] H. Kiryu, K. Asai, Rchange: algorithms for computing energy changes of RNA secondary structures in response to base mutations, Bioinformatics 28 (2012) 1093–1101.

[96] D. Barash, A. Churkin, Mutational analysis in RNAs: comparing programs for RNA deleterious mutation prediction, Brief. Bioinformatics 12 (2011) 104–114.

[97] T. Fukunaga, H. Ozaki, G. Terai, K. Asai, W. Iwasaki, H. Kiryu, CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data, Genome Biol. 15 (2014) R16.

[98] H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, Q. Morris, RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins, PLoS Comput. Biol. 6 (2010) e1000832.

[99] M. Hiller, R. Pudimat, A. Busch, R. Backofen, Using RNA secondary structures to guide sequence motif finding towards single-stranded regions, Nucleic Acids Res. 34 (2006) e117.

[100] J. Iwakiri, T. Kameda, K. Asai, M. Hamada, Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions, Bioinformatics 29 (2013) 2524–2528.

[101] S. R. Eddy, Computational analysis of conserved RNA secondary structure in transcriptomes and genomes, Annu Rev Biophys 43 (2014) 433–456.

[102] J. Gorodkin, W. L. Ruzzo (Eds.), RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods (Methods in Molecular Biology), 2014 ed., Humana Press, 2014. URL: http://amazon.com/o/ASIN/162703708X/.

[103] E. Picardi (Ed.), RNA Bioinformatics (Methods in Molecular Biology), 2015 ed., Humana Press, 2015. URL: http://amazon.com/o/ASIN/1493922904/.

[104] P. J. Volders, K. Verheggen, G. Menschaert, K. Vandepoele, L. Martens, J. Vandesompele, P. Mestdagh, An update on LNCipedia: a database for annotated human lncRNA sequences, Nucleic Acids Res. 43 (2015) D174–180.

[105] D. Bhartiya, K. Pal, S. Ghosh, S. Kapoor, S. Jalali, B. Panwar, S. Jain, S. Sati, S. Sengupta, C. Sachidanandan, G. P. Raghava, S. Sivasubbu,

V. Scaria, lncRNome: a comprehensive knowledgebase of human long noncoding RNAs, Database (Oxford) 2013 (2013) bat034.

[106] S. R. Eddy, Accelerated Profile HMM Searches, PLoS Comput. Biol. 7 (2011) e1002195.

[107] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, A. J. Giraldez, Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation, EMBO J. 33 (2014) 981–993.

[108] X. Wang, I. M. El Naqa, Prediction of both conserved and nonconserved microRNA targets in animals, Bioinformatics 24 (2008) 325–332.

[109] M. Cesana, D. Cacchiarelli, I. Legnini, T. Santini, O. Sthandier, M. Chinappi, A. Tramontano, I. Bozzoni, A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA, Cell 147 (2011) 358–369.

[110] X. C. Quek, D. W. Thomson, J. L. Maag, N. Bartonicek, B. Signal, M. B. Clark, B. S. Gloss, M. E. Dinger, lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs, Nucleic Acids Res. 43 (2015) D168–173.

[111] C. Park, N. Yu, I. Choi, W. Kim, S. Lee, lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs, Bioinformatics 30 (2014) 2480–2485.

[112] D. Incarnato, F. Neri, F. Anselmi, S. Oliviero, Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome, Genome Biol. 15 (2014) 491.

[113] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, E. Segal, Genome-wide measurement of RNA secondary structure in yeast, Nature 467 (2010) 103–107.

[114] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, A. P. Arkin, Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq), Proc. Natl. Acad. Sci. U.S.A. 108 (2011) 11063–11068.

[115] N. A. Siegfried, S. Busan, G. M. Rice, J. A. Nelson, K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP), Nat. Methods 11 (2014) 959–965.

Table 1: Available resources of lncRNA expressions in Expression Atlas and ArrayExpress

Species	Sample sources	Expression Atlas ID	# of tissues	# of cell lines	GEO/ENA/EGA ID	Reference
Human	Normal tissues	E-MTAB-2836	32	–	ERP006650(ENA)	[17]
		E-MTAB-513	16	–	GSE30611(GEO)	[18]
		E-GEOD-30352	8	–	GSE30352(GEO)	[19]
	Cancer cell lines	E-MTAB-2706	–	675	EGAS00001000610(EGA)	[20]
		E-MTAB-2980	–	39	PRJNA169425(ENA)	–
	Subcellular compartments	E-GEOD-26284	–	15	GSE26284(GEO)	[21]
Mouse	Normal tissues	E-MTAB-2801	9	–	GSE41637(GEO)	[22]
		E-MTAB-599	6	–	GSE30617(GEO)	[23]
		E-GEOD-30352	6	–	GSE30352(GEO)	[19]
		E-GEOD-41338	5	–	GSE41338(GEO)	[24]
Rat	Normal tissues	E-GEOD-53960	11	–	GSE53960(GEO)	[25]
		E-MTAB-2800	9	–	GSE41637(GEO)	[22]

Table 2: Summary of available public data for structural probing techniques in combination with high-throughput sequencing

Chemical	Method	Species	Target	GEO	SRA	Platform	read type	length	Ref
DMS	structure-seq	A.thaliana	total RNA (polyA)	–	SRP027216	HiSeq2000	single	40	[69]
DMS	DMS-seq	S.cerevisiae H.sapiens	total RNA (polyA)	GSE45803	SRP020556	HiSeq2000	single	30-50	[70]
DMS	Mod-seq	S.cerevisiae	total RNA	–	SRP029192	HiSeq2000	single	50	[71]
DMS/CMCT	CIRS-seq	M.musculus	total RNA	GSE54106	SRP035427	HiScanSQ	single	50	[112]
RNase P1	Frag-seq	M.musculus	total RNA	GSE24622	SRP004883	SOLiD 3.0	single	50	[67]
RNase V1,S1	PARS-seq	S.cerevisiae	total RNA (polyA)	GSE22393	SRP003175	SOLiD	single	50	[113]
RNase V1,S1	PARS-seq	H.sapiens	total RNA (polyA)	GSE50676	SRP029656	HiSeq	pair	101(*)	[73]
1M7	SHAPE-seq	B.subtilis	total RNA	GSE31573	SRP007955	GAIIx	pair	50	[114]
1M6,1M7,NMIA	SHAPE-seq	HIV-1	genome (RNA)	–	SRP042347	HiSeq MiSeq	pair/single		[115]

(*) 50bp at the $5'$ end was utilized.
The columns "type" and "length" indicate the read type and length of the data, respectively. SRA, short read archive (http://www.ncbi.nlm.nih.gov/sra); GEO, Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/); DMS, dimethyl sulfate; CMCT, 1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate; NMIA, N-methylisotoic anhydride; 1M7, 1-methyl-7-nitroisatoic anhydride;

Table 3: Summary of lncRNA databases

Database	LNCipedia	lncRNome	lncRNAdb	lncRNAtor
# of lncRNA	111,685	113,513	287	34,605
Species	human	human	human, mouse, rat, zebrafish, fly, yeast, chicken, etc.	human, mouse, zebrafish, fly, worm, yeast
Sequence	Y	Y	Y	N
Secondary structure	Y	Y	N	N
Genomic position	Y	Y	Y	Y
Coding potential	CPC, HMMER, PRIDE, PhyloCSF, Ribosome-profiling	PhyloCSF	-	CPC
Conservation	Ensembl Compara	Phastcons	literature	Phastcons
Orthologue	N	N	Y	Y
Expression profile	N	N	Y (Human Body Atlas)	Y (GEO, ENCODE, modENCODE, TCGA)
# of human tissues[1]	-	-	16	10 (cancer)
RNA–protein[2]	N	Y(exp., pred.)	Y (literature)	Y (exp.)
RNA–RNA[2]	Y (pred.)	N	Y (literature)	N
SNP	N	Y	N	N
Epigenome[3]	N	Y	N	N
GO term	N	N	N	Y
Link to references	Y	Y	Y	N
External links	UCSC genome, PubMed, Ensembl	HGNC, OMIM, NCBI, PubMed, Ensembl	PubMed, NCBI	UCSC genome, Ensembl
Reference	[104]	[105]	[110]	[111]

[1] Number of human tissues for expression profiles of lncRNAs.
[2] Intermolecular interactions supported by experiment (exp.) or prediction (pred.).
[3] DNA methylations and histone modifications of lncRNA genes.

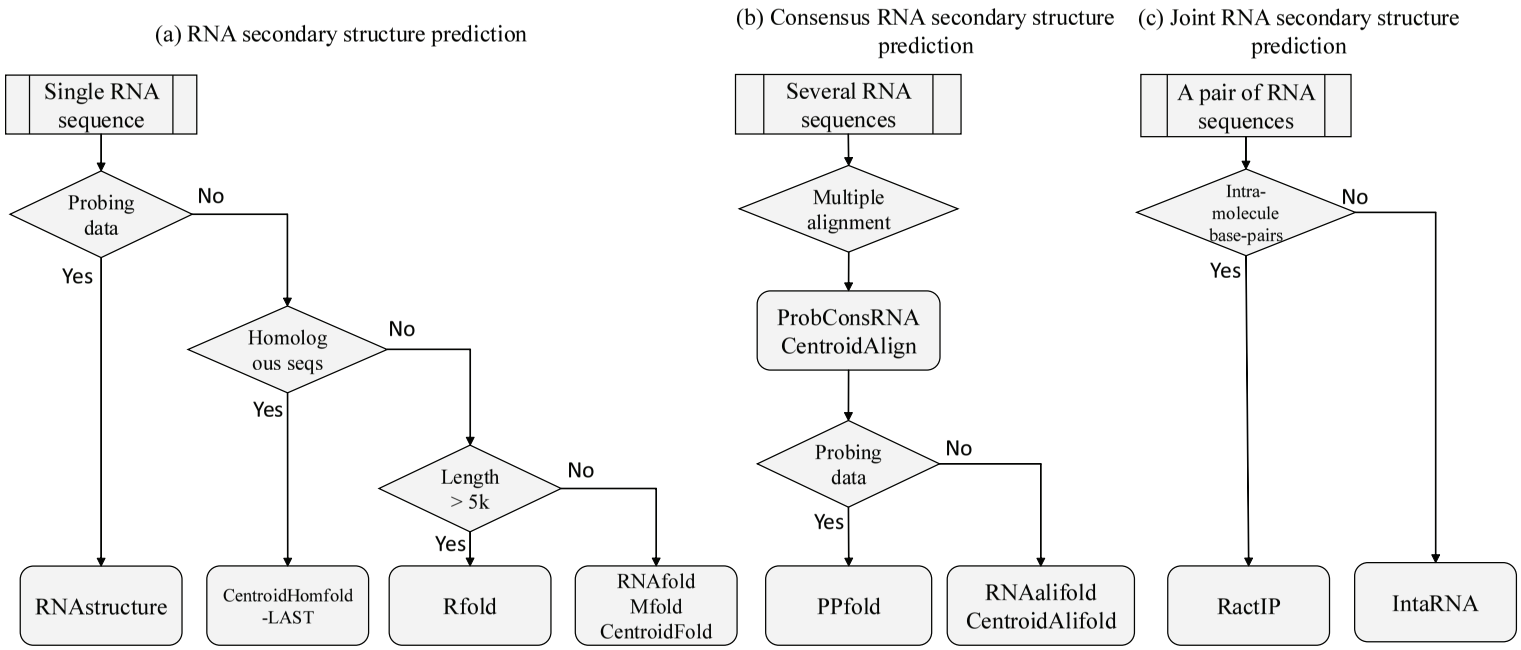

Figure 1: Flowcharts of structure predictions for lncRNAs: (a) RNA secondary structure predictions for a single RNA sequence (see Section 5.1); (b) Consensus/common RNA secondary structure predictions for several RNA sequences (see Section 5.2); (c) Joint RNA secondary structure predictions for a pair of RNA sequences (see Section 5.3).

Highlights

We review bioinformatics tools and databases that are useful for lncRNA research.

We show purely computational tools as well as those utilizing experimental data.

We review tools for analyzing 2D and 3D structures and the interactions.

We review tools and databases of expression profiles for analyzing localization.