Chapter 4

Structural and Functional Annotation of Long Noncoding RNAs

Martin A. Smith and John S. Mattick

Abstract

Protein-coding RNAs represent only a small fraction of the transcriptional output in higher eukaryotes. The remaining RNA species encompass a broad range of molecular functions and regulatory roles, a consequence of the structural polyvalence of RNA polymers. Albeit several classes of small noncoding RNAs are relatively well characterized, the accessibility of affordable high-throughput sequencing is generating a wealth of novel, unannotated transcripts, especially long noncoding RNAs (lncRNAs) that are derived from genomic regions that are antisense, intronic, intergenic, and overlapping protein-coding loci. Parsing and characterizing the functions of noncoding RNAs—lncRNAs in particular—is one of the great challenges of modern genome biology. Here we discuss concepts and computational methods for the identification of structural domains in lncRNAs from genomic and transcriptomic data. In the first part, we briefly review how to identify RNA structural motifs in individual lncRNAs. In the second part, we describe how to leverage the evolutionary dynamics of structured RNAs in a computationally efficient screen to detect putative functional lncRNA motifs using comparative genomics.

Key words lncRNA, Comparative genomics, RNA secondary structure, Homology search, Functional genome annotation

1 Introduction

Functional genome annotation involves the identification of both known and hypothetical genes in uncharacterized genomic DNA sequence. This largely includes protein-coding genes and noncoding RNAs, as well as other genomic features such as telomeric/subtelomeric regions and centromeres. The identification of protein-coding genes can unravel the molecular repertoire of the majority of the genomes of microorganisms, especially prokaryotes, whose genomes are largely composed of protein-coding sequences. However, protein-coding sequences encompass only a small fraction of the genome in higher eukaryotes, which decreases with increasing developmental and cognitive complexity [1, 2] and comprise less than 1.5 % of the human genome.

Jonathan M. Keith (ed.), *Bioinformatics: Volume II: Structure, Function, and Applications*, Methods in Molecular Biology, vol. 1526, DOI 10.1007/978-1-4939-6613-4_4,

Most of the human genome is dynamically transcribed into RNA [3, 4], which implies that untranslated RNAs compose the most abundant class of genomic output. In particular, noncoding transcripts greater than 200 nt in length—long noncoding RNAs (lncRNAs)—are emerging as master regulators of development and differentiation in higher eukaryotes [5–10]. There are currently 15,767 lncRNA genes (excluding alternative isoforms and pseudogenes) listed in version 25 of the GENCODE human genome annotation database, compared to 19,950 protein-coding genes. Contrary to protein-coding genes, whose set is relatively well characterized and has remained relatively stable in number and repertoire throughout metazoan evolution [1, 2, 11], although there are novel genes mainly encoding small proteins being discovered [12], the number of identified lncRNAs is steadily increasing as more and more biological conditions are investigated with high-throughput RNA sequencing technologies.

Many lncRNAs appear to regulate gene expression through their association with epigenetic proteins, such as histone modification enzymes and DNA methyltransferases, with which they synergistically organize the nuclear environment [13–15]. Other lncRNA functions include acting as molecular decoys and macromolecular scaffolds, as well as the regulation of splicing and translation, mRNA stability, and the formation of subcellular organelles [5, 16]. A small but growing number of lncRNAs have been functionally validated through knockout and ectopic expression in vivo and in cell culture, and other biochemical studies [17–20], but the precise molecular mechanisms and structures guiding their function remain largely unresolved.

At present, lncRNAs are largely categorized by their position relative to neighboring protein-coding genes, i.e., intergenic, antisense, intronic, or bidirectional. However, the particular functions of lncRNAs do not necessarily correlate with their genomic context. For example, the lncRNA *HOTAIR* functions by recruiting a chromatin modification complex (PRC2) to repress gene expression in *trans* [21], whereas the lncRNA *HOTTIP* recruits another epigenetic complex (WDR5-MLL1) in *cis* to activate gene expression via chromosomal looping [22]. Both are situated in the intergenic regions surrounding HOX genes. The functional annotation of lncRNAs at a genome- or transcriptome-wide scale therefore requires the consideration of additional molecular features that may be unique to each transcript.

A unifying feature of ncRNAs is their propensity to form discrete secondary and tertiary structures through canonical and noncanonical nucleotide base pairings that often dictate their function. Many lncRNAs appear to be very plastic, evolve quickly, and/or have arisen relatively recently in evolution, as evidenced by high turnover rates and reduced primary sequence conservation [23, 24], although there are exceptions that have extraordinarily high

levels of sequence conservation [25–27]. Their evolutionary dynamics are different from protein-coding genes, displaying relaxed structure-function constraints that are synonymous with being under positive selection for adaptive radiation. They are in general (although there are likely to be exceptions) unlikely to have catalytic activities, such as ribosomal RNAs, yet may nonetheless form evolutionarily stable, functional secondary and tertiary structures with different functions, as well as shorter primary sequences that may interact with other RNAs and DNA. For instance, the widespread presence of repetitive sequences derived from mobile elements in the human genome is believed to contribute to modular lncRNA biogenesis by forming a reservoir of functional motifs—or structured templates for RNA-binding proteins—that can be co-opted into RNA regulatory networks via positive selection [28–30].

Computational identification of functional RNA structural motifs encoded in genomic sequences is a challenging task, mainly because almost any RNA sequence can form internal base pairs via classical Watson–Crick, Hoogstein, or ribose 2′OH hydrogen bond formation, and fold into discrete structures [31, 32], but also because RNA structures themselves are dynamic, flexible, and are contingent on the cellular environment (i.e., temperature, ion concentrations, ligand binding, transcriptional kinetics). Functional RNA structures can nonetheless be identified through comparative genomics by observing nucleotide substitutions that are consistent and compatible with a common structural topology. Indeed, a much larger fraction of the human genome seems to function through the formation of RNA structure motifs than through sequence-constrained elements, as evidenced by considering nucleotide covariation events in evolutionary information [33].

In this chapter, we describe how to annotate ncRNAs in genomic or transcriptomic data, where known or putative functions are assigned to uncharacterized sequences to gain insight into their biology. First, we summarize how to identify functional RNA elements in single sequences via homology search as well as prediction of local structures in long transcripts. Finally, we describe how to identify putative functional motifs in lncRNAs that are supported by evolutionarily conserved RNA secondary structures. We provide user friendly, step-by-step instructions on how to perform a multiple genome-wide screen for functional RNA motifs similar to that published in [33].

2 Materials

A UNIX-based computing environment should be employed for most of the described methods, preferably with access to a high-performance computing infrastructure. Alternatively, a computer or server with multiple processors and over 4 GB of RAM may be employed.

2.1 Genomic Data

Genomic or transcriptomic sequence data should be downloaded and converted (if required) to *fasta* file format, unless it is already available. Genomic data for reference organisms can be obtained from the following sources:

1. UCSC genome browser—select the organism and the desired genome version, then full data set, then the file with suffix "*.fa.gz*" at http://hgdownload.cse.ucsc.edu/downloads.html.
2. NCBI—select the species of interest and then sequence data can be downloaded for each chromosome individually at (ftp://ftp.ncbi.nih.gov/genomes/). A FTP batch download tool or interface should be considered to automate the process.
3. ENSEMBL genome browser—select the appropriate release version, then 'fasta' at ftp://ftp.ensembl.org/pub/.

2.2 Transcriptomic Data

LncRNAs are often spliced (including alternatively spliced), generating sequences and structures that would otherwise be missed during computational screens of unprocessed genomic sequences. Depending on the task at hand and the availability of suitable data, the sequences corresponding to processed transcripts should also be considered to improve the robustness of functional lncRNA annotation. For RNA sequencing data, algorithms for de novo assembly should be considered provided the depth of coverage is sufficient. These programs usually produce output files containing genomic coordinates in *.bed* (browser extendible data file, preferably in 12-field format with exon boundary information), *.gtf* (gene transfer format), *.gff* (general feature format), or similar formats. The popular *Cufflinks* program from the *Tuxedo* suite of RNAseq tools [34] produces a *.gtf* file and includes the appropriate software—a program called *gffread* located in the Cufflinks binary folder—to extract and process sequence information from a reference genome into a *.fasta* file. Alternatively, the *Trinity* program for de novo transcriptome assembly without aligning to a reference genome [35] directly outputs a *.fasta* file of assembled transcripts from the *.fastq* files containing deep sequencing data.

2.3 Multiple Genome Alignments

Comparative genomics approaches for functional annotation of noncoding RNAs require pairwise or multiple genome alignments for the species of interest. Prealigned genomic sequence alignments for most well-studied vertebrates can be downloaded in *.maf* (multiple alignment format) from the ENSEMBL comparative genomics database [36] or from the UCSC genome browser [37]—which also hosts alignments for nonvertebrate species—as follows:

1. ENSEMBL Compara—Information about downloading multiple genome alignments is available at http://ensembl.org/info/data/ftp/index.html. Multiple alignments in *.maf* from the latest release at the time this was written can be downloaded

via FTP protocol at ftp://ftp.ensembl.org/pub/release-85/maf/ensembl-compara/multiple_alignments/.

2. UCSC Genome Browser—Navigate to the table browser tab at http://genome.ucsc.edu (select 'tools,' then 'table browser' from the drop-down menu bar on the top of the page). Select the reference species of interest, then 'Comparative Genomics' from the group menu, 'Conservation' from the track menu, and 'Multiz Align,' form the table menu. Optionally, regions can be limited to an existing UCSC or custom track (which needs to be uploaded independently prior to this step). This can significantly reduce the size of the download when only interested in a set of transcripts, for example. Next, ensure that 'MAF—multiple alignment format' appears in the output format menu, otherwise the appropriate track or table must be selected. Finally, name the output file and get output (ideally, compressed) or send the output to the Galaxy [38] platform for post-processing (*see later*).

Multiple alignments form the UCSC Genome Browser employ a different synteny and alignment algorithm than those from ENSEMBL. The latter usually present contiguous alignments for large syntenic blocks via the *Enredo* (or *Mercator*) and *Pecan* algorithms [39, 40], whereas the former is optimized for total genomic coverage and presents smaller, fragmented alignment blocks as produced with the *TBA* and *MULTIZ* algorithms [41]. Because of their highly fragmented nature and variable presence of each species in each block, TBA/MULTIZ alignments may require additional processing, such as being 'stitched' together. A good summary of approaches for processing *.maf* files is described by Blankenberg et al. [42]. The ENSEMBL alignments require less processing, as the syntenic blocks are much longer. These alignments can also contain segmental duplications, which should be removed at the user's discretion (ensuring that the coordinates of the segmental duplications for the reference species are saved for future reference).

3 Methods

The first step in any analysis of a putative noncoding RNA is to estimate its protein-coding potential. This typically involves excluding known protein-coding genes from a reference genome annotation, from mass spectrometry data (when available), as well as computational estimation of coding potential via the analysis of open reading frames and evolutionary information, such as synonymous codon usage. The *Pinstripe* software suite is one example of a recently developed bioinformatics resource that enables the discrimination of coding versus noncoding transcripts, which is accompanied by a well-described usage manual [12]. Such methods

and additional considerations—i.e., bifunctional RNA transcripts that are both mRNAs and ncRNAs—are reviewed in [43, 44].

There are two general approaches for the functional annotation of noncoding RNAs: (1) homology search against known RNAs; and (2) de novo identification of putative functional domains. The former is more suitable for the annotation of small RNAs (e.g., tRNAs, snoRNAs, 5S rRNAs, snRNAs, miRNAs, etc.); however, an increasing number of lncRNAs have been sufficiently characterized and are amenable to this approach (*see* [45] and the most recent release of RFAM). De novo computational annotation of noncoding RNAs can be applied to both size categories of transcripts and involves the elucidation of both sequence and structural characteristics that are indicative of function. Comparison of sequence similarity to orthologous genes, for instance, with BLAST [46], is a commonly employed method for the identification of protein-coding genes and ribosomal RNAs given their strong dependence on sequence composition as well as crucial cellular functions. However, when comparing genes with similar functions across larger evolutionary distances, sequence homology is outclassed by structural homology, where classical sequence alignment methods are inefficient. Hidden Markov models [47, 48] and codon substitution matrices (e.g., PAM [49] or BLOSUM [50]) are employed to overcome the sequence alignment barrier when faced with greater sequence divergence than for protein-coding genes.

For noncoding RNAs, alternative computational strategies must be employed to overcome the increased diversity of sequences that are compatible with a given secondary or tertiary structure. The evolutionary dynamics of noncoding RNAs are governed by three factors: (1) They do not require the preservation of sequence composition to convey a genetic code, i.e., codons, with the notable exception of the anticodon loop in tRNAs. (2) RNA structures are more tolerant to nucleotide substitutions than proteins for mutated codons. Indeed, 6 out of 16 possible canonical ribonucleotide combinations will form canonical base pairings, which include Watson–Crick and G-U/U-G 'wobble' base pairs. Because RNA structures can accommodate a higher frequency of base substitutions than mRNAs—as long as they are consistent or compatible with their paired nucleotide—bioinformatics tools investigating noncoding RNAs must focus on secondary and tertiary structural characteristics as well as primary sequence, where short patches of high conservation may indicate important biochemical interactions. (3) Since their biological function is often of regulatory nature, they are more likely to be under positive selection for adaptive radiation. This is most notable for lncRNAs.

3.1 Detecting Homology to Known Functional RNAs

The RFAM database encompasses several well-characterized noncoding RNA families that are presented in multiple alignments based on both their sequence and higher order structure topologies [51]. Until recently, the RFAM repository was mostly limited

to entire RNA sequences, mainly small noncoding RNAs. Recent updates to RFAM have expanded the repository to include some lncRNAs as well as *bona fide* RNA structural motifs [52]. The latter are defined as "a non-trivial, recurring RNA sequence and/or secondary structure that can be predominantly described by local sequence and secondary structure elements" and can be part of a larger structure or noncoding RNA [53]. RFAM includes Covariance Models (CMs) for each entry, or family, in the database. CMs are a probabilistic representation of RNA structure profiles that can be used to scan a genome (or transcriptome) for sequences compatible with a given consensus structure. They can be used by the *Infernal* program to scan large metazoan genomes in minutes and report homologous hits with high accuracy [54]. The *Infernal* software package can also generate a CM from a given multiple sequence and structure alignment and thus permits using custom CMs to perform a search. Detailed instructions on how to use *Infernal* can be found at http://infernal.janelia.org/ as well as in [55].

There are also alternative bioinformatics resources for RNA structural homology search. The *RNAmotif* program enables users to construct descriptors of a target RNA structure, then scans a sequence database, and reports all compatible sequences [56]. Although the software is somewhat out of date, *RNAmotif*'s capacity to construct detailed and customized RNA structure descriptors manually and with relative ease justify its pertinence. It also enables the inclusion of tertiary structural elements such as pseudoknots, triplexes, and quadruplexes. Unfortunately, it does not consider thermodynamic stability or base-pairing probabilities and, consequently, can produce a large amount of biologically irrelevant hits unless the results are filtered appropriately (for a practical example of how this may be performed, please refer to the last paragraph of Subheading 3). Alternatively, the recently developed *LocaRNAscan* algorithm [57] can consider the local structural environment in the target sequence when performing a scan using a base pair probability matrix (*see later*) as a query, which can be generated from a single sequence or an alignment of several sequences.

3.2 Predicting the Structural Landscape of Individual lncRNA Sequences

The computational prediction of RNA secondary structures from sequence alone was one of the first challenges in bioinformatics. Consequently, modern software packages such as *RNAfold* [58], *UNAfold* [59], and *RNAstructure* [60, 61] are quite efficient at predicting the most thermodynamically stable RNA secondary structure—Minimum Free Energy (MFE)—for a given input sequence. Unfortunately, MFE structural predictions do not always represent the biological reality and, on their own, are not usually considered as a robust qualification of function. This is particularly true for lncRNAs, which can be tens of thousands of nucleotides

long. Locally stable RNA secondary structures, which might compose functional units (or modules) of a lncRNA, can be overlooked in favor of long-range base pairings that contribute more toward lowering the overall free energy score. Furthermore, the dynamic structural nature of RNA macromolecules also confounds RNA structure prediction, as noncoding RNAs can form more than a single functional structural topology (riboswitches are a good example). It is therefore beneficial to consider an ensemble of suboptimal structures when characterizing the function of noncoding RNAs, as exemplified in Fig. 1.

A more biologically relevant alternative to the MFE structure is the centroid, which consists of the structure with minimal distance to all other structures in a set of suboptimal structures. The centroid is usually generated through the partition function, which estimates the statistical distribution of all possible RNA structures within a given thermodynamic range (Boltzmann ensemble). Although centroid estimators have been shown to outperform MFE predictions on known RNA structures [62], they do not necessarily inform about the stability or diversity of the structural landscape for a given query sequence. The latter can be evaluated in two ways: (1) through direct visual inspection of a base-pairing probability matrix, such as that produced by the "*RNAfold –p*" program in the Vienna RNA package (Fig. 1a)—a greater quantity of smaller dots is indicative of a larger diversity of compatible base pairings for a particular nucleotide, which is consistent with a reduced likelihood of forming a stable structure; and (2) through the command-line output of *RNAfold*, or the *RNAfold* webserver [63], which produce a numerical estimate of the ensemble diversity, as well as the frequency of the MFE within the ensemble (i.e., how credible the MFE structure prediction is). A larger ensemble diversity value suggests that the queried RNA sequence may form a broader repertoire of structures or dynamically fluctuate between intermediary structures.

As mentioned earlier, secondary structure prediction of individual lncRNA sequences is not a trivial task. Fortunately, the computational prediction of locally stable structural elements has been shown to be more accurate than global RNA structural predictions for long RNA polymers [64]. This finding is consistent with the general hypothesis that lncRNAs function via local structural (or unstructured) domains, such as protein-binding motifs or RNA–DNA interactions (*see* Subheading 1). *RNAplfold* from the Vienna RNA package [58] and its enhancement in *LocalFold* [64] both offer a useful solution for the manual inspection of local structural topologies in long noncoding RNAs. The tools produce a base-pairing probability matrix that spans the entire RNA sequence but limits the range of base-pairing interactions to a user-definable threshold (Fig. 1d). This facilitates the identification of locally stable (or unstable) structures, which can reveal putative

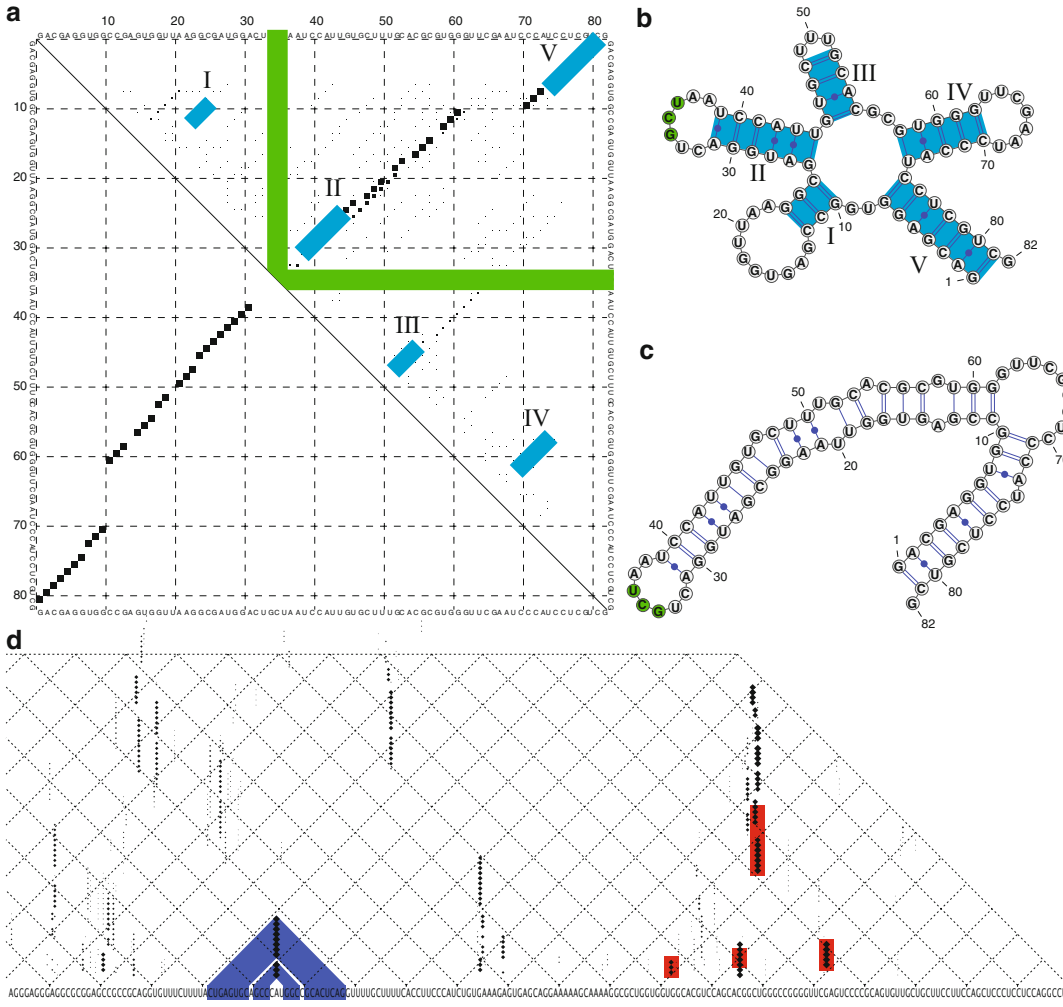

Fig. 1 Representation of RNA secondary structure predictions for single sequences. (**a**) An RNA base-pairing probability matrix representing both the minimum free energy structure prediction (below the diagonal) and suboptimal base-pairing probabilities (above the diagonal) of a serine tRNA that forms five helices. The RNA sequence of interest is displayed on the *X* and *Y* axes, where each dot represents possible base pairings between bases (*x*,*y*). The size of the dots is indicative of the frequency (or probability) of the base pairings in a Boltzmann ensemble of suboptimal structures, as calculated by McCaskill's partition function algorithm in the Vienna RNA package [58]. The base pairs forming the validated biological structure (**b**) are highlighted in *blue* and numbered accordingly, whereas the unpaired bases forming the anticodon are highlighted in *green*. (**c**) The MFE prediction forms a structure that is quite divergent to the actual tRNA, although the biological structure is perceptible in the suboptimal base pairings. (**d**) A base-pairing probability matrix generated by the *RNAplfold* algorithm on a ~400 nt section of the 3′ end of the NEAT1 lncRNA. Locally stable base pairings are displayed as described for (**a**), however the sequence is represented on the diagonal (i.e., the upper quadrant of (**b**) is rotated 45°). In the lower left, the bases associated to the base pairs (*dots*) are highlighted in *blue*. In the lower right, the tRNA-like structure at the 3′ end of NEAT1 (as illustrated in Fig. 2c) is highlighted in *red*

functional regions as well as guide the design of small interfering RNAs for knockdown experiments. Alternatively, there are software tools, such as *Rnall* [65], *RNAsurface* [66], *RNAlfoldz* (part of the Vienna RNA package [58]), that can facilitate the identification of RNA subsequences presenting strong local structural stability, although a user-defined maximal base-pairing span is required.

3.3 Inferring Function from an Individual RNA Sequence

If noncoding RNAs function through the formation of stable secondary structures, can structure predictions alone be used for de novo functional annotation of ncRNAs? This question was first examined over 30 years ago by comparing the RNA structure (or 'folding') score of a native RNA sequence to that of shuffled sequences, under the premise that functional RNAs should form more stable structures than random sequences [67–69]. This strategy produced promising results, but it was consequently shown that the relatively higher stability of native noncoding RNA sequences reflected local biases in sequence composition rather than structural features alone [70]. In particular, the energetic contributions of base-stacking interactions were ignored (the order of consecutively arranged base pairs can significantly alter the free energy score). Some reports have since successfully applied this approach to certain classes of noncoding RNAs by using adequate background models that control for dinucleotide content [71, 72]. Known and novel RNA elements have also been predicted in the yeast genome using a similar strategy, several of which were subsequently experimentally validated [73].

3.4 Detecting Functional 2D Motifs via Comparative Genomics

The biological significance of lncRNAs has often been questioned since they (generally) display lower conservation of primary sequence than proteins in evolutionary comparisons [24, 74]. Conservation of RNA secondary or tertiary structure has rarely been considered in such analyses, partially due to the more complex bioinformatic analyses required to investigate such phenomena. However, probing evolutionary data for evidence of RNA structural conservation is not substantially more difficult in practice than evaluating primary sequence conservation. In this section, we describe how to leverage the hallmark signature of RNA structural conservation, i.e., base pair covariation, to identify putative functional RNA motifs in multiple sequence alignments, using existing software.

We recently showed that measuring RNA structure conservation from genomic sequence alignments of 32 mammals could identify evidence of purifying selection on RNA structure motifs that span over 13 % of the human genome, while presenting little overlap with known sequence-constrained regions [33]. Evolutionarily Conserved Structure (ECS) predictions with the human genome as reference can be visualized in the UCSC genome browser (Fig. 2) as follows:

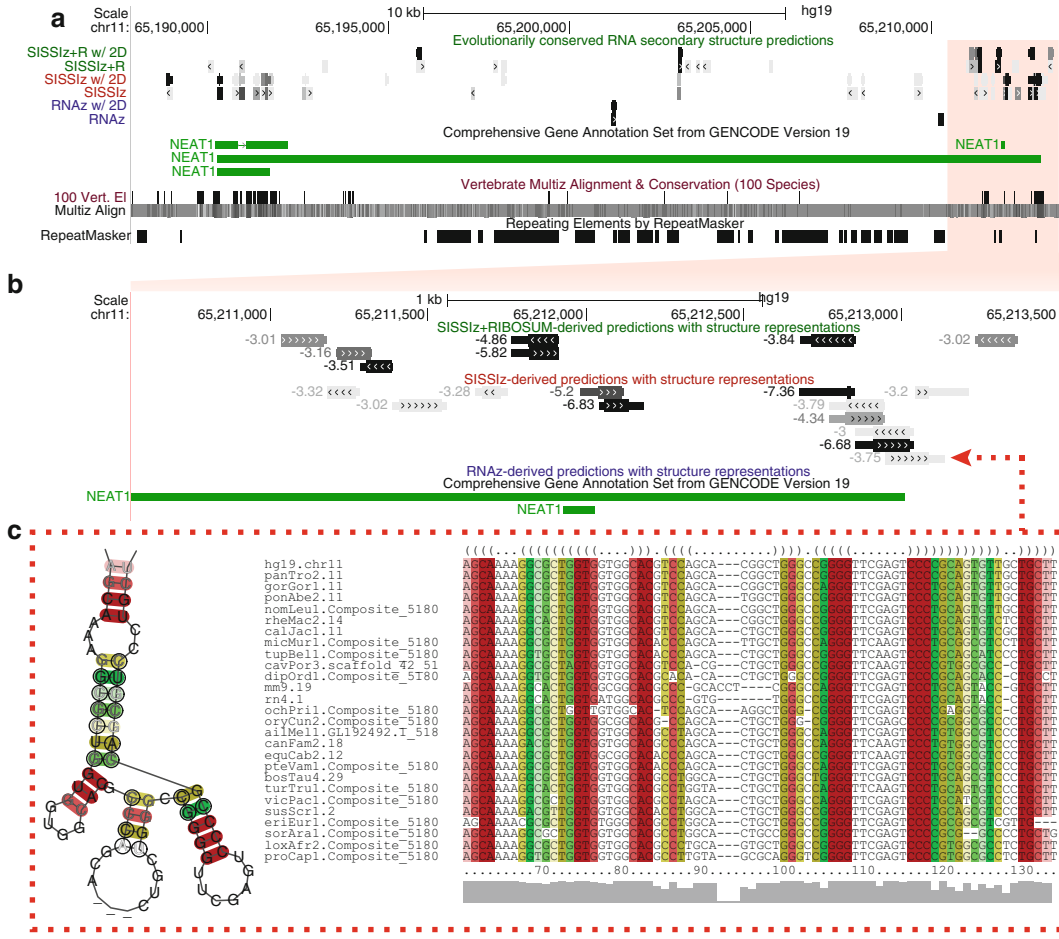

Fig. 2 Visualization of ECS predictions in the UCSC Genome Browser. (**a**) The NEAT1 lncRNA locus presenting several ECS predictions from [33]. Six subtracks are displayed: *SISSIz*, *SISSIz* with *RIBOSUM* scoring, and *RNAz*-derived results for all significant predictions and those with structure topologies and alignments available to view on a web server (*see* Subheading 3). (**b**) Expanded, zoomed in view of the tracks with structure representations. The RNA secondary structure consensus, flanked by the outermost base pair, is represented by a thicker rectangle. The color of the bars corresponds to a relative measure of their scores (darker = stronger score). (**c**) Detailed illustration of a segment of the predicted structure and alignment obtained by clicking on an ECS prediction from (**b**), which also provides general predictions statistics, a dot-bracket representation of the consensus structure and the consensus sequence generated on the spot via the Vienna RNA package [58]

1. Browse to http://genome.ucsc.edu (or any UCSC Genome Browser mirror), navigate to the 'Genomes' tab, then select the hg19 human genome assembly.
2. Click on the 'track hubs' button, then select the 'My hubs' tab.
3. Paste in the URL for the ECS track hub (http://www.martinalexandersmith.com/hubs/ecs/hub.txt), then 'Add Hub.' The URL can also be obtained via the supplementary information from [33].

4. Browse to any region of interest, zooming out if the ECS track hub titles appear and nothing is displayed under them in the browser (usually, >1 KB of genomic span should be sufficient). ECS predictions are split according to the algorithms that were used to make the predictions (*RNAz*, *SISSIz*, and *SISSIz* + *RIBOSUM*). Although all the ECS predictions are statistically significant (with a ≤ 1 % false-positive rate), they are color coded based on their relative scores (darker = less likely to arise by chance). After fully expanding the tracks, either by clicking on the title of the track or in the individual track configuration below the browser, the scores associated to the predictions are displayed as the name of each ECS prediction. *SISSIz*-derived predictions will display Z-scores, which represent the degree of observed structural conservation (in number of standard deviations) from the mean of a background distribution produced from *SISSIz*'s null model. There are two subtracks for each employed algorithm: one supporting structure representations, one without. Those with structure representation also have larger segments annotated within individual ECSs; these correspond to the positions within the sampled genomic alignments that contain the outermost base pairs forming the conserved structure prediction (Fig. 1).
5. Expand the ECS track display settings to 'pack' or 'full' view by clicking on the title bar or by selecting the appropriate view in the drop-down menu below the browser interface window.
6. Directly click on a bar corresponding to an ECS prediction of interest. Depending on the nature of the subtrack, this will either: (1) link to a page with a rundown of the statistics for the ECS of interest as well as a description of the methodology; or (2) link to an external page with detailed statistics for the selected ECS, a colored and annotated figure of the consensus secondary structure corresponding, the multiple sequence alignment (colored and annotated) that was used to make the prediction, as well as the consensus structure and sequence in dot-bracket format (Fig. 1c). The ECS tracks with structure representations that link to an external page (as described earlier) will display bars with thin and thick segments; the thinner extremities correspond to regions in the sampled alignment that are not contained within the predicted secondary structure, whereas the thicker internal portion of the bars represents regions contained within the ECS prediction (*see* **Note 1**).
7. Any combination of subtracks (i.e., all ECS predictions, predictions with structure representations, or the results for individual algorithms) can be hidden (or redisplayed) by clicking on the link in the title of the ECS predictions track, located in the drop-down controls section of the UCSC browser below the main window.

There are several caveats pertaining to the data currently contained within the ECS track hub for the UCSC browser. These data are derived from genome-wide screens that are resource intensive and, consequently, were applied to heuristic and not necessarily accurate genome-scale multiple sequence alignments (alignment errors can often be observed via close inspection of alignments from **step 6**). The quality and amount of significant ECS predictions will undoubtedly improve by realigning the queried sequences with more robust algorithms, such as *Clustal Omega* [75], *MAFFT* [76] or, ideally, RNA structure alignment algorithms (reviewed in [77]).

Another caveat is that the above-mentioned ECS predictions are generated from sliding windows of ≤200 nucleotides (nt), which includes multiple genome alignment columns that can primarily be composed of indels. This means RNA base pairs that are more than 200 nt apart are ignored. Furthermore, the sampled alignment windows are offset by 100 nt, therefore conserved RNA structures smaller than 200 nt may also be missed given an incomplete sampling of the structure's boundaries.

An additional issue with the functional annotation of lncRNAs is that many are spliced, often comprising relatively small exons. Although the biological motives for lncRNA splicing remain enigmatic, one possibility is that constitutively spliced exons are joined to maintain the formation of higher order structures, whereas alternatively spliced exons contain self-contained modular units. Probing multiple alignments for evidence of RNA structural conservation in spliced transcripts would thus require pasting the alignment blocks together first (reviewed in [42]), as well as additional considerations like splice site conservation and syntenic continuity in other species.

Performing a de novo scan for ECSs in multiple sequence alignments, either from another reference species or from a set of spliced alignments, can be quite computationally intensive. The approach used for the genomic screen published in [33] can nonetheless be performed by anyone with basic command-line experience. For large alignments (whole genomes or chromosomes)

1. Download and install the following software packages (requires compilation and linking the binaries to the environmental $PATH variable):
   (a) *SISSIz* 2.0 and *RNAz* 2.0 [78] available at http://martinalexandersmith.com/ecs or via links provided in their original manuscripts (N.B. *SISSIz* 2.0 was released in [33]).
   (b) The Vienna RNA package at http://www.tbi.univie.ac.at/RNA [58], preferably version 1.8.5 (newer versions may not be compatible with the software in **step 2**).

2. Download the JAVA archive containing the binary code required to scan .maf files from the following URL (in the software section): http://martinalexandersmith.com/ecs.
3. Ensure that the multiple (genome) sequence alignments have the reference species in the first row with genomic coordinates in the appropriate field of the *.maf* file. This will be used to output the genomic coordinates of the predictions during the scan. Also, ensure that the alignments present sufficiently long blocks (*see* Subheading 2 and **Note 2**).
4. Launching the following command (in the appropriate directory) from a UNIX terminal will provide more verbose information on the basic usage and available parameters: '*java –jar MafScanCcr.jar*.' Some options include window size, step or 'sliding' distance, realignment of the input with the multiple sequence alignment program *MAFFT*, number of processors to use, etc.
5. Execute the program with the selected parameters. The program will load one alignment block of the *.maf* input file at a time, with an optional realignment step to increase accuracy at the expense of computation time. Next, *N* windows are sampled concurrently, where *N* is the number of specified processors (the alignments can also be run in parallel on a computer cluster).
6. The program will save all sampled subalignments that score above the respective thresholds for each employed algorithm. Genomic coordinates associated to significant ECS predictions for the alignment's reference species are also emitted to the standard output in browser extendable (*.bed*) format. Simply redirect the standard output to a file, e.g., '> *output.bed*' from the UNIX terminal. Alternatively, genomic coordinates can be recovered from the file names of the saved alignments, which encode a 6-field underscore delimited bed-compatible entry. Furthermore, the name field of the *.bed* entries also encodes colon-delineated statistical information about the alignment used to make the ECS prediction. This includes (in order):
   (a) Number of retained sequences.
   (b) Raw mean pairwise identity (including indels).
   (c) Mean pairwise identity (normalized to the shortest gapless sequence length).
   (d) Relative gap (indel) content.
   (e) Standard deviation of the (normalized) mean pairwise identity.
   (f) Normalized Shannon entropy.
   (g) Relative GC content.

(h) Scoring algorithm employed: s = *SISSIz* 2.0; r = *SISSIz* 2.0 with *RIBOSUM* scoring; z = *RNAz*-2.0.

The fifth field of the *.bed* entries represents the score associated with the predictions. The scores have been modified to accommodate representation in the UCSC genome browser, which only supports integer values. Z-scores from *SISSIz* predictions are multiplied by −100 (−2.54 = 254), whereas *RNAz*-derived scores are simply multiplied by 100 (0.85 = 85).

7. The topology of a given ECS prediction can be visualized by running the *RNAalifold* program from the Vienna RNA package on the multiple alignment associated to the predicted ECS. The default *RNAalifold* options are suitable for ECS predictions from *SISSIz* and *RNAz*, but the *RIBOSUM* scoring option '*-r*' should be used otherwise.

8. Because the ECS predictions are based on a consensus, it is possible that the reference species forms a structure that is not compatible with the consensus. To evaluate the likelihood of this structural congruence, an auxiliary program is available to process the alignments output from **step 6** (see the supplementary information of [33]). The *ParseAlifold.jar* program performs two main tasks: (1) trimming the genomic coordinates of the reference species to the outermost base pairs of the consensus structure; (2) measuring the relative difference between the native secondary structure for the sampled reference sequence and that produced from constraining the structure to the consensus, as produced from the '*RNAfold −C*' command from the Vienna RNA package [58]. This is done for both the minimum free energy and the base-pairing probabilities generated from the partition function implemented in *RNAfold*, where the probabilities of base pairs from the consensus are extracted from the base-pairing probability matrix. The *.bed* 6 plus formatted output prints to the terminal's standard output and contains the following additional fields:

   (a) Average base-pairing probability of the minimum free energy structure for the reference species. If the base is unpaired, this value is calculated as 1—the sum of all probabilities for the given base.

   (b) Average base-pairing probability of consensus-constrained reference structure.

   (c) Base-pairing probability ratio (constrained/native).

   (d) Free energy (kcal/mol) of the consensus-constrained reference sequence.

(e) Minimum free energy (kcal/mol) of the native reference sequence.

(f) Free energy ratio (constrained/native).

(g) Length of prediction (nt).

(h) Dot-bracket secondary structure mask of *RNAalifold* consensus. Ex: (((((...))))).

3.5 The Next Frontier: Functional Parsing of lncRNAs

In higher eukaryotes, recurring RNA structural motifs that display evidence of evolutionary conservation provide a tangible basis for the functional annotation of noncoding sequences, as they may indicate protein-interaction domains that potentially nucleate regulatory networks. For example, Parker et al. [79] performed a similar analysis using evolutionarily conserved RNA secondary structures predicted with *EvoFold* [80] to generate profile Stochastic Context-Free Grammars (SCFGs), which were then used to scan the human genome for paralogs to the RNA structural predictions. The results were grouped into RNA families based on their structural similarities and revealed 220 families of RNA structures, including 172 novel RNA structure families.

However, as effective as bioinformatic methods may be, they seldom indicate what biological functions or processes are involved (unless, of course, there is a high level of homology to well-characterized RNAs). Assigning biological functions to novel RNA structural motifs can be achieved via modern experimental techniques predicated on high-throughput sequencing, such as RNA immunoprecipitation (RIP-Seq), crosslinking immunoprecipitation (CLIP-Seq), and chromatin isolation by RNA purification (ChiRP-Seq). These methods can identify the RNAs interacting with specific proteins, providing sets of RNA sequences that share the same protein-binding characteristics. The increasing availability of next-generation sequencing technologies will likely increase contributions to public specialized databases such as *starBase* [81], which contains numerous RNAseq data sets relating to RNA–protein interactions. Mining these data with advanced bioinformatics tools will bridge the gap between functional annotation of lncRNAs and RNA structure prediction.

Computational identification of RNA structures common to a set of sequences can currently be performed via clustering algorithms based on pairwise comparison scores, obtained through either RNA structure alignment algorithms (e.g., *CARNA* [82], *LocaRNA* [83], *FOLDALING* [84, 85]) or other secondary structure comparison strategies (e.g., *GraphClust* [86], RNACluster [87], and *NoFold* [88]). These approaches have been applied to small RNA sequences and have successfully identified both known and yet to be characterized noncoding RNA families based on their shared secondary structures [79, 83, 85, 87, 88]. Unfortunately, lncRNA sequences are not directly amenable to such structure-motif enrichment approaches because they may harbor extraneous

sequence elements, thus requiring additional processing such as the extraction of subsequences presenting stable RNA structure domains. Refining the aforementioned methods and applying them to sequencing data that target RNA–protein interactions will help identify new functional RNA structure motifs, which can, in turn, serve to index genomic sequences. This strategy will lay the foundations required to unravel the structure–function relationships of lncRNAs, categorize their repertoires, and annotate the expanses of noncoding sequences in vertebrate genomes.

4 Notes

1. *Sense or antisense?* Given the complementary nature of canonical RNA base pairs (G–C/C–G), it is not uncommon to find that both strands of DNA produce high scoring, consensus secondary structure predictions. When these bidirectional structure predictions arise in regions with little or no associated transcription, determining the most likely orientation of the putative transcript can be quite difficult. Sequences with high GC content are more susceptible to this phenomenon because there are fewer G–U base pairs, which can effectively be used to discriminate the host transcript's orientation (the antisense A–C base pair does not contribute to canonical Watson–Crick base pairing). Occasionally, visual inspection of the alignments and consensus RNA secondary structures can be sufficient to identify the most likely orientation, i.e., the strand that produces more base pairs (G–U in particular). Otherwise, the most likely orientation can sometimes be determined by using the *RNAstrand* program [89], a machine learning algorithm which was specifically developed for this purpose (not covered here). *RNAstrand* generates a score which estimates the orientation of a consensus RNA secondary structure from a given multiple sequence alignment used as input.
2. *Genomic alignments and block sizes.* As a strict minimum, the blocks should be at least the length of the window size for sampling structure conservation (by default, 200 nt). The longer the blocks are, the more consecutive overlapping windows will be sampled, which will provide greater genomic coverage of the computational screen. Usually, alignments with more species will present shorter blocks given the greater diversity of synteny. In this case, 'stitching' the alignment blocks together can also abrogate synteny in nonreference sequences (i.e., all but the first row in the alignment), which may introduce uncertainty in the consensus structure evaluation as noncontiguous sequences are treated as contiguous. For example, a 500 nt segment from human chromosome 12 might align to a

250 nt segment from mouse chromosome 3 and 250 nt from mouse chromosome 6, therefore any windows sampled between the segment joining both mouse chromosomes will not reflect the biological reality (unless these regions are prone to fusion or trans-splicing events, an unlikely predicament). From a practical viewpoint, the multiple genome alignments produced by the *Enredo-Pecan-Ortheus* pipeline [39, 90] (available via the ENSEMBL comparative genomics portal: http://ensembl.org/info/genome/compara/index.html) present much longer syntenic blocks than those from *TBA/Multiz* [41] (accessible via the UCSC Genome Browser comparative genomics tracks), thus avoiding the need to 'stitch' several small alignments together.

References

1. Liu G, Mattick JS, Taft RJ (2013) A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle 12 (13):2061–2072
2. Taft RJ, Pheasant M, Mattick JS (2007) The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29 (3):288–299
3. Djebali S, Davis CA, Merkel A et al (2012) Landscape of transcription in human cells. Nature 489(7414):101–108
4. Mercer TR, Gerhardt DJ, Dinger ME et al (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30(1):99–104
5. Morris KV, Mattick JS (2014) The rise of regulatory RNA. Nat Rev Genet 15(6):423–437
6. Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15(1):7–21
7. Mattick JS (1994) Introns: evolution and function. Curr Opin Genet Dev 4(6):823–831
8. Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2(11):986–991
9. Mattick JS (2011) The central role of RNA in human development and cognition. FEBS Lett 585(11):1600–1616
10. Mattick JS (2010) RNA as the substrate for epigenome-environment interactions: RNA guidance of epigenetic processes and the expansion of RNA editing in animals underpins development, phenotypic plasticity, learning, and cognition. Bioessays 32(7):548–552
11. Ezkurdia I, Juan D, Rodriguez JM et al (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23(22):5866–5878
12. Gascoigne DK, Cheetham SW, Cattenoz PB et al (2012) Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. Bioinformatics 28(23):3042–3050
13. Mercer TR, Mattick JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 20 (3):300–307
14. Koziol MJ, Rinn JL (2010) RNA traffic control of chromatin complexes. Curr Opin Genet Dev 20(2):142–148
15. Mattick JS, Amaral PP, Dinger ME et al (2009) RNA regulation of epigenetic processes. Bioessays 31(1):51–59
16. Wang KC, Chang HY (2011) Molecular mechanisms of long noncoding RNAs. Mol Cell 43(6):904–914
17. Li L, Chang HY (2014) Physiological roles of long noncoding RNAs: insight from knockout mice. Trends Cell Biol 24(10):594–602
18. Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5(4):e1000459
19. Quek XC, Thomson DW, Maag JL et al (2014) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 43:D168–D173. doi:10.1093/nar/gku988
20. Sauvageau M, Goff LA, Lodato S et al (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. Elife 2:e01749
21. Rinn JL, Kertesz M, Wang JK et al (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129 (7):1311–1323

22. Wang KC, Yang YW, Liu B et al (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature 472(7341):120–124
23. Ulitsky I, Shkumatava A, Jan CH et al (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147(7):1537–1550
24. Johnsson P, Lipovich L, Grander D et al (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta 1840 (3):1063–1071
25. Bejerano G, Haussler D, Blanchette M (2004) Into the heart of darkness: large-scale clustering of human non-coding DNA. Bioinformatics 20(Suppl 1):i40–i48
26. Calin GA, Liu CG, Ferracin M et al (2007) Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. Cancer Cell 12(3):215–229
27. Stephen S, Pheasant M, Makunin IV et al (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol Biol Evol 25 (2):402–408
28. Kapusta A, Kronenberg Z, Lynch VJ et al (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet 9(4):e1003470
29. Matylla-Kulinska K, Tafer H, Weiss A et al (2014) Functional repeat-derived RNAs often originate from retrotransposon-propagated ncRNAs. Wiley Interdiscip Rev RNA 5 (5):591–600
30. Smith M, Bringaud F, Papadopoulou B (2009) Organization and evolution of two SIDER retroposon subfamilies and their impact on the Leishmania genome. BMC Genomics 10:240
31. Stombaugh J, Zirbel CL, Westhof E et al (2009) Frequency and isostericity of RNA base pairs. Nucleic Acids Res 37 (7):2294–2312
32. Cruz JA, Westhof E (2009) The dynamic landscapes of RNA architecture. Cell 136 (4):604–609
33. Smith MA, Gesell T, Stadler PF et al (2013) Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 41 (17):8220–8236
34. Trapnell C, Hendrickson DG, Sauvageau M et al (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31(1):46–53
35. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8(8):1494–1512
36. Flicek P, Amode MR, Barrell D et al (2014) Ensembl 2014. Nucleic Acids Res 42(Database issue):D749–D755
37. Karolchik D, Barber GP, Casper J et al (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42(Database issue): D764–D770
38. Goecks J, Nekrutenko A, Taylor J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86
39. Paten B, Herrero J, Beal K et al (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res 18(11):1814–1828
40. Dewey CN (2007) Aligning multiple whole genomes with Mercator and MAVID. Methods Mol Biol 395:221–236
41. Blanchette M, Kent WJ, Riemer C et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14 (4):708–715
42. Blankenberg D, Taylor J, Nekrutenko A et al (2011) Making whole genome multiple alignments usable for biologists. Bioinformatics 27 (17):2426–2428
43. Ilott NE, Ponting CP (2013) Predicting long non-coding RNAs using RNA sequencing. Methods 63(1):50–59
44. Dinger ME, Pang KC, Mercer TR et al (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol 4(11):e1000176
45. Burge SW, Daub J, Eberhardt R et al (2013) Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 41(Database issue):D226–D232
46. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
47. Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6(3):361–365
48. Krogh A, Brown M, Mian IS et al (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235(5):1501–1531
49. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. National Biomedical Research Foundation, Washington, DC
50. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks.

Proc Natl Acad Sci U S A 89 (22):10915–10919
51. Griffiths-Jones S, Bateman A, Marshall M et al (2003) Rfam: an RNA family database. Nucleic Acids Res 31(1):439–441
52. Nawrocki EP, Burge SW, Bateman A et al (2014) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43: D130–D137. doi:10.1093/nar/gku1063
53. Gardner PP, Eldai H (2014) Annotating RNA motifs in sequences and alignments. Nucleic Acids Res 43:691–698. doi:10.1093/nar/gku1327
54. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935
55. Griffiths-Jones S (2005) Annotating noncoding RNAs with Rfam. Curr Protoc Bioinformatics Chapter 12, Unit 12.15
56. Macke TJ, Ecker DJ, Gutell RR et al (2001) RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 29(22):4724–4735
57. Will S, Siebauer MF, Heyne S et al (2013) LocARNAscan: incorporating thermodynamic stability in sequence and structure-based RNA homology search. Algorithms Mol Biol 8:14
58. Lorenz R, Bernhart SH, Honer Zu Siederdissen C et al (2011) ViennaRNA Package 2.0. Algorithms Mol Biol 6:26
59. Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453:3–31
60. Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA 10(8):1178–1190
61. Mathews DH, Disney MD, Childs JL et al (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 101 (19):7287–7292
62. Hamada M, Kiryu H, Sato K et al (2009) Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics 25(4):465–473
63. Gruber AR, Lorenz R, Bernhart SH et al (2008) The Vienna RNA websuite. Nucleic Acids Res 36(Web Server issue):W70–W74
64. Lange SJ, Maticzka D, Mohl M et al (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Res 40(12):5215–5226
65. Wan XF, Lin G, Xu D (2006) Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. J Bioinform Comput Biol 4(5):1015–1031
66. Soldatov RA, Vinogradova SV, Mironov AA (2014) RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. Bioinformatics 30 (4):457–463
67. Seffens W, Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res 27 (7):1578–1584
68. Chen JH, Le SY, Shapiro B et al (1990) A computational procedure for assessing the significance of RNA secondary structure. Comput Appl Biosci 6(1):7–18
69. Le SY, Maizel JV Jr (1989) A method for assessing the statistical significance of RNA folding. J Theor Biol 138(4):495–510
70. Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16(7):583–605
71. Bonnet E, Wuyts J, Rouze P et al (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 20(17):2911–2917
72. Clote P, Ferre F, Kranakis E et al (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA 11(5):578–591
73. Kavanaugh LA, Dietrich FS (2009) Noncoding RNA prediction and verification in Saccharomyces cerevisiae. PLoS Genet 5(1): e1000321
74. Kutter C, Watt S, Stefflova K et al (2012) Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet 8 (7):e1002841
75. Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 1079:105–116
76. Katoh K, Standley DM (2014) MAFFT: iterative refinement and additional methods. Methods Mol Biol 1079:131–146
77. Gorodkin J, Hofacker IL (2011) From structure prediction to genomic screens for novel non-coding RNAs. PLoS Comput Biol 7(8): e1002100
78. Gruber AR, Findeiss S, Washietl S et al (2010) RNAz 2.0: improved noncoding RNA detection. Pac Symp Biocomput, 69–79
79. Parker BJ, Moltke I, Roth A et al (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Res 21(11):1929–1943

80. Pedersen JS, Bejerano G, Siepel A et al (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2(4):e33
81. Li JH, Liu S, Zhou 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(Database issue):D92–D97
82. Sorescu DA, Mohl M, Mann M et al (2012) CARNA—alignment of RNA structure ensembles. Nucleic Acids Res 40(Web Server issue): W49–W53
83. Will S, Reiche K, Hofacker IL et al (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 3(4):e65
84. Havgaard J, Kaur S, Gorodkin J (2012) Comparative ncRNA gene and structure prediction using Foldalign and FoldalignM. Curr Protoc Bioinformatics Chapter 12, Unit12.11
85. Torarinsson E, Havgaard JH, Gorodkin J (2007) Multiple structural alignment and clustering of RNA sequences. Bioinformatics 23(8):926–932
86. Heyne S, Costa F, Rose D et al (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics 28(12):i224–i232
87. Liu Q, Olman V, Liu H et al (2008) RNACluster: an integrated tool for RNA secondary structure comparison and clustering. J Comput Chem 29(9):1517–1526
88. Middleton SA, Kim J (2014) NoFold: RNA structure clustering without folding or alignment. RNA 20(11):1671–1683
89. Reiche K, Stadler PF (2007) RNAstrand: reading direction of structured RNAs in multiple sequence alignments. Algorithms Mol Biol 2:6
90. Paten B, Herrero J, Fitzgerald S et al (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res 18 (11):1829–1843