Briefings in Bioinformatics Advance Access published April 21, 2009

An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes

Giulia Soldà, Igor V. Makunin, Osman U. Sezerman, Alberto Corradin, Giorgio Corti and Alessandro Guffanti

Submitted: 5th December 2008; Received (in revised form): 12th March 2009

Abstract

Non-protein coding RNAs (ncRNAs) have emerged as a vast and heterogeneous portion of eukaryotic transcriptomes. Several ncRNA families, either short (<200 nucleotides, nt) or long (>200 nt), have been described and implicated in a variety of biological processes, from translation to gene expression regulation and nuclear trafficking. Most probably, other families are still to be discovered. Computational methods for ncRNA research require different approaches from the ones normally used in the prediction of protein-coding genes. Indeed, primary sequence alone is often insufficient to infer ncRNA functionality, whereas secondary structure and local conservation of portions of the transcript could provide useful information for both the prediction and the functional annotation of ncRNAs. Here we present an overview of computational methods and bioinformatics resources currently available for studying ncRNA genes, introducing the common themes as well as the different approaches required for long and short ncRNA identification and annotation.

Keywords: *small and long noncoding RNA; gene prediction; genome annotation; bioinformatics analysis; regulatory RNA; bioinformatics programming*

INTRODUCTION TO THE NONCODING RNA WORLD

In the last decade, non-protein coding RNAs (ncRNAs) have emerged as a diverse and vast portion of mammalian transcriptome, accounting for the majority of all annotated transcripts [1, 2]. Before that, the number of known ncRNAs was restricted to 'housekeeping' RNAs, such as ribosomal RNAs (rRNA), transfer RNAs (tRNA) and spliceosomal RNAs (snRNA), together with few regulatory RNAs, such as *H19* and *Xist* (X-Inactive Specific Transcript) in mammals and the microRNA *lin-4* in *C.elegans* [3–5]. Currently, thousands of short ncRNAs have been identified, including

Corresponding author. Giulia Soldà, Department of Biology and Genetics for Medical Sciences, University of Milano, Via Viotti 3/5, 20133 Milan, Italy. Tel: +390250315852; fax: +390250315864; E-mail: giulia.solda@unimi.it

Giulia Soldà has done her Bsc and Msc in Medical Biotechnology and PhD in Molecular and Cellular Biology, and is currently Assistant Professor in Applied Biology at the University of Milan. She currently works on noncoding RNAs involvement in complex human diseases.

Igor Makunin was a research officer at the Institute for Molecular Bioscience, University of Queensland, until 2009. Currently he affiliates with the Institute of Chemical Biology and Fundamental Medicine, Novosibirsk, Russia. His main interests are comparative genomics and noncoding RNA.

Osman Ugur Sezerman has done his Bsc in Electrical Engineering, MSc and PhD in Biomedical Engineering. He is heading the Computational Biology Laboratory at Sabanci University since 1999. He develops algorithms in computational biology, and also in the RNA field.

Alberto Corradin has his Bsc in Electronic Engineering. He is attending the PhD program in Bioengineering at the University of Padua. He currently works on noncoding RNA analysis and on modeling of biological systems at the Istituto Oncologico Veneto (IOV), Padua, Italy.

Giorgio Corti has completed his Bsc in Physics and MSc in Bioinformatics. After a period of research in astronomy, he works since 2006 in the bioinformatics analysis of deep sequencing data and on bioinformatics programming at the ITB-CNR, Milan, Italy.

Alessandro Guffanti has done Bsc in Biological sciences, Msc in Computer Science, and again Msc in Medical Statistics. He currently works at Genomnia s.r.l., Milan, Italy, on the development of bioinformatics analysis pipelines and in the analysis of deep sequencing data from transcriptome and genome research projects, including microRNA identification and annotation.

© The Author 2009. Published by Oxford University Press. For Permissions, please email: journals.permissions@oxfordjournals.org

microRNAs (miRNA), endogenous short interfering RNAs (siRNA), PIWI-interacting RNAs (piRNA) and small nucleolar RNAs (snoRNA) [6, 7]. Moreover, tens of thousands of long ncRNAs (>200 nucleotides, nt) have been discovered using full-length complementary DNA (cDNA) cloning and genomic tiling arrays to comprehensively profile plants and animals transcriptomes [8–12].

ncRNAs have been implicated in a variety of regulatory processes, ranging from X chromosome inactivation, genomic imprinting and chromatin modification to transcriptional activation, transcriptional interference, post-transcriptional gene silencing and nuclear trafficking [13, 14]. Small RNA functions are better defined, and the mechanisms by which they exert their effects are, at least partially, understood [7]. For instance, miRNAs are endogenous 21–23 nt RNA molecules that act as post-transcriptional regulators of gene expression by targeting cognate messenger RNAs (mRNAs) for translational repression/degradation, via the association with Argonaute proteins [14]. On the contrary, it is still largely unclear how long ncRNAs work. However, it has become apparent that long ncRNAs can act both in *cis* and in *trans* [15–19], and that some function as precursors for short ncRNAs [20, 21].

Despite the recent progress in understanding this heterogeneous and previously hidden layer of regulatory transcripts, the majority of ncRNAs is still uncharacterized, and doubts have been raised as to how many of them are functional at all [22]. Certainly, unlike protein-coding genes where sequence motifs are usually indicative of function, ncRNA primary sequence information may be insufficient for predicting their function *a priori* [23]. Therefore, correct annotation of ncRNAs and discrimination between protein-coding and noncoding transcripts requires novel strategies and poses novel challenges.

We present here an overview of current computational methods and bioinformatic resources for the identification, annotation and characterization of ncRNAs, focusing on regulatory ncRNAs. Annotation of housekeeping or 'classical' ncRNAs has been recently reviewed elsewhere [24]. After a description of structural approaches, which can be generally applied to ncRNA prediction and annotation, separate sections are devoted to specific strategies for long and short ncRNA annotation (as they present different computational challenges), and a final section collects the most relevant databases and software for ncRNA research. This review aims at providing a broad picture of the main ncRNA analytical approaches, commenting some selected methods and identifying the most problematic issues.

STRUCTURAL APPROACHES FOR ncRNA ANNOTATION

Reliable genome-wide ncRNA annotation is currently restricted to homologs of known structured RNA families, which mainly includes housekeeping RNAs (e.g. transfer, ribosomal and spliceosomal RNAs) and small RNAs (mainly miRNAs).

The *Rfam* database (Table 1) is built from structure-annotated multiple sequence alignments, covariance models (CMs) and family annotation for noncoding RNAs, *cis*-regulatory and self-splicing intron families [25]. Over a million of sequences have been aligned to form over 1300 families (current release 9.1, January 2009). This is a powerful resource because functional ncRNAs often have a secondary structure which is more conserved than the simple nucleotide sequence. CMs used in *Rfam* can efficiently model both the sequence and the structure, leading to the predicted functional classification of a ncRNA. The INFERNAL software (which incorporates also a Blast search engine) is at the core of the *Rfam* CMs build process and search and can be downloaded from the *Rfam* site, together with the target database. Searches with known miRNA precursors from mirBase or tRNA sequences, for example, invariably gave a clear-cut result, with alignments and the secondary structure prediction included. It is also possible to browse from the website all the families or the genomes (>1100) by species name or specific kingdoms.

Each *Rfam* family is hand-curated, both in the alignment and in threshold used for the CM, and is annotated with multiple useful information, such as a link to the related Wikipedia page describing the family and giving literature references. Moreover, the predicted phylogenetic tree for any alignment, generated using either a maximum likelihood approach or neighbour-joining, is displayed along with the secondary structure. Finally, a dedicated section shows detailed information about the *Rfam* family, such as data curation, model feature and the CM bit-scores distributions.

Rfam and other structural approaches have been extensively used for both long and short ncRNA

Table 1: Databases for non coding RNA research

Database	Source	Small RNAs	Long ncRNAs	Description	Availability
RNAdb	http://jsm-research.imb.uq.edu.au/rnadb/	Y	Y	Mammalian ncRNA database	W, D
NONCODE	http://www.noncode.org/index.htm	Y	Y	Database of ncRNAs from 861 organisms (Eukaryotes, Bacteria, Archea, Viruses)	W, D
fRNAdb	http://www.ncrna.org/fRNAdb/	Y	Y	A comprehensive non-coding RNA sequence database, also including data from RNAdb, NONCODE, Rfam and mirBase	W, D
Rfam	http://rfam.sanger.ac.uk/	Y	Few	collection of structural RNA families	W, D
NRED	http://jsm-research.imb.uq.edu.au/NRED/	N	Y	Repository of ncRNA expression information	W, D
NATsDB	http://natsdb.cbi.pku.edu.cn/	N	Y	Specific for *cis* antisense transcripts	W, D
Trans-SAMap	http://trans.cbi.pku.edu.cn/	N	Y	Specific for *trans* antisense transcripts	W, D
antiCODE	http://bioinfo.ibp.ac.cn/ANTICODE/index.htm	N	Y	Specific for *cis* and *trans* antisense transcripts	W, D
miRBase	http://microrna.sanger.ac.uk/sequences/	Y	N	Main repository for microRNA data	W, D
snoRNABase	http://www-snorna.biotoul.fr/index.php	Y	N	Human snoRNAs database	W
Plant snoRNA database	http://bioinf.scri.sari.ac.uk/cgi-bin/plant.snorna/home	Y	N	Plant snoRNAs database	W, D
smiRNAdb	http://www.mirz.unibas.ch/smiRNAdb/cgi/smiRNAdb	Y	N	Database of miRNA expression information (small RNAs cloned by the Tuschl Lab)	W, D
microrna.org	http://www.microrna.org/microrna/home.do	Y	N	Database of miRNA targets and expression	W, D
Argonaute	http://www.ma.uni-heidelberg.de/apps/zmf/argonaute/	Y	N	Database of mammalian miRNAs expression and their known or predicted targets	W
Tarbase	http://diana.cslab.ece.ntua.gr/tarbase/	Y	N	Database of experimentally supported miRNA target interactions	W, D
MirGator	http://genome.ewha.ac.kr/miRGator/miRGator.html	Y	N	Database and navigator tool for functional interpretation of miRNAs	W

W = web-based resource, D = downloadable data.

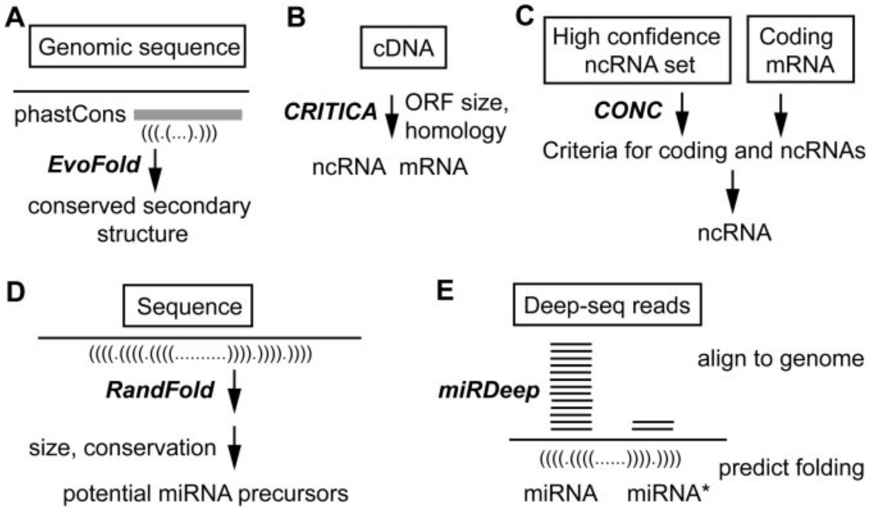

Figure 1: Schematic representations of bioinformatic strategies for detection and functional annotation of ncRNAs. Dots and parentheses conventionally represent secondary structure prediction. (**A**) Secondary structure approach: consider all conserved phastCons elements and run *EvoFold* to detect interesting hairpin structures. (**B**) *CRITICA* analysis: consider all transcripts aligned to the genome scaffold and discriminate between coding and ncRNAs on the basis of predicted ORF size and conservation. (**C**) Machine learning approach (*CONC*): train the software with high confidence ncRNA set (i.e. datasets from *RNAdb* and *NONCODE* databases) and coding mRNAs and then feed the classifier with real sequence samples. (**D**) Predict miRNA precursors from secondary structure (hairpin), size and conservation. (**E**) *MirDeep* approach for miRNA identification from deep-sequencing data: differential recovery of sequences from the miRNA and its cognate star form, coupled with secondary structure prediction of the hairpin precursor.

discovery and annotation, but proved less successful with longer transcripts (for a more detailed description see also [26, 27]). All these methods rely on the potential of a ncRNA sequence to fold into a stable secondary structure, and on the hypothesis that ncRNA function is mediated by its secondary structure. However, the formation of a stable secondary structure alone, evaluated by thermodynamics parameters, is generally not sufficient to reliably detect ncRNAs, as many of them do not appear to adopt significantly more stable structure than random RNA sequences [28]. As ncRNAs often conserve a base-paired secondary structure with low primary sequence similarity, the combination of secondary structure prediction with conservation of that structure in related species has proved successful for the identification of functional ncRNAs [29–31]. Several programs (i.e. *QRNA*, *RNAz* and *Evofold*) used the prediction of conserved secondary structures for ncRNA identification (Figure 1A). However, although these methods can work as long noncoding RNA gene predictors, they actually identify conserved elements of RNA secondary structure that can and do occur in mRNAs as well as ncRNAs. For instance, conserved local secondary structures are particularly abundant within 5' and 3' untranslated regions (UTR) of mRNAs, where regulatory proteins bind [32]. Therefore, programs based on secondary structure prediction might lead to significant false positive and false negative discoveries. A further issue for the efficient annotation of secondary structure prediction of long ncRNAs is the accurate knowledge of transcript boundaries. Indeed, if the transcript sequence is incomplete (not full length), it is very difficult to assign the correct functional annotation to that RNA transcript. The transcript length also affects the performance of the folding program. Finally, important but essentially unstructured long ncRNAs like *Xist* and *IPW* (Imprinted in Prader–Willi) are not detected by these methods.

BIOINFORMATIC APPROACHES FOR LONG ncRNAS IDENTIFICATION AND ANNOTATION

The identification and functional annotation of putative regulatory long 'mRNA-like' ncRNAs is a primary focus of computational RNA research, but currently available methods to predict ncRNAs on a genome scale are still immature. The annotation of large eukaryotic genomes is computationally expensive, thus limiting the feasibility of approaches that are otherwise efficient in searching smaller genomes. In addition, genomes of higher eukaryotes are frequently associated with ncRNA-derived pseudogenes and repeats, which make the discrimination of functional copies from the nonfunctional ones challenging. To date, in addition to the detection of conserved secondary structure, long ncRNA discovery and annotation has been based mainly on protein-coding potential determination, primary sequence conservation among different species and approaches which combine the previous information.

Assessment of the protein-coding potential

As long ncRNAs generally lack discernable features to facilitate categorization and functional prediction, the most widely used strategy to annotate a ncRNA is to exclude that it possesses protein-coding features, thus discriminating it from mRNA [33, 34].

Typically, the starting data are novel cDNAs or Expressed Sequence Tags (EST) obtained by high-throughput experiments (e.g. full-length cDNA cloning, tiling arrays, deep sequencing data) or selected known transcripts (again mainly cDNAs or EST) retrieved from public databases, such as GenBank, FANTOM, etc. In many cases, these transcripts are derived from libraries aimed at isolating mRNAs, which preferentially include the polyadenylated RNA fraction. As a result, non-polyadenylated transcripts tend to be underrepresented among available cDNA and/or EST collections.

A number of bioinformatic methods, such as *DIANA-EST* [35], *ESTScan* [36], *BlastX* [37], *CSTminer* [38] and *CRITICA* [39], have been applied to estimate the protein-coding potential of an RNA sequence. The protein-coding potential is mainly determined on the basis of Open Reading Frame (ORF) length and/or ORF conservation among different species. Usually, a minimum ORF cut-off is defined (normally 300 nt, may be lowered down to 150); below this threshold, the transcript is considered to be noncoding (Figure 1B). The different choice of this threshold has produced divergent estimates of the prevalence of ncRNAs in mammalian genomes, and in any case might lead to misannotations. Indeed, long-known regulatory ncRNAs, such as *Xist* and *H19*, contain by chance sufficiently long putative ORF to be erroneously annotated as protein-coding, while transcripts encoding short proteins might be incorrectly classified as ncRNAs [40]. Therefore, an additional criteria is searching for ORF homology to known proteins or domains, on the hypothesis that ORFs lacking cross-species conservation are more likely to occur randomly. In this way, RNAs with short, non-conserved ORF are most likely to represent *bona fide* ncRNAs. Few of these comparative methods, including *CSTminer* and *CRITICA*, work reasonably well also for genome-wide analyses.

Recently, two novel algorithms based on support vector machines (SVM), *CONC* and *Coding Potential Calculator (CPC)*, have been used to assess the coding potential of putative ncRNAs [41, 42]. In these algorithms, multiple distinct features of mRNAs are exploited by the machine learning methods to distinguish the ncRNAs from mRNAs. For instance, the *CONC* (for ‘coding or noncoding’) classifier considered features of native proteins such as peptide length, homology with known proteins, amino acid composition, secondary structure, solvent accessible surface area and sequence compositional entropy. Liu and colleagues trained this SVM using eukaryotic ncRNAs from the *RNAdb* and *NONCODE* databases [43, 44] and showed that protein features can be used to distinguish ncRNAs from mRNAs with 97% specificity and 98% sensitivity (Figure 1C). All the features contributed to the high classification accuracy, but the top-performing individual features were the number of database homologs and peptide length. The strong contribution of peptide length was probably a consequence of the shorter average length of the ncRNAs (526 nt) compared to the mRNAs (1746 nt). Instead, the presence of several protein homologs is clearly a strong support for the protein-coding potential of any RNA. The major weaknesses of the homology search were misclassification of RNAs coding for novel proteins or homology based hits of ncRNAs to mis-annotated hypothetical proteins. The inclusion of additional features, such as alignment entropy and amino acid composition, seemed to enhance the prediction accuracy.

A second SVM-based approach, named *CPC* [42] used fewer features than *CONC* for training (6 versus 180), but achieved comparable performance in significantly less time. A user-friendly web-based interface of *CPC* is also available (Table 2).

Table 2: Computational tools for non coding RNA research

Task	Tool	Source	Small RNAs	Long RNAs	Availability
Protein coding potential assessment	CST-miner	http://t.caspur.it/CSTminer/	N	Y	W
	ESTScan2	http://www.ch.embnet.org/software/ESTScan2.html	N	Y	W, D
	BlastX	ftp://ftp.ncbi.nih.gov/blast/ http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome	N	Y	W, D
	CRITICA	http://www.ttaxus.com/software.html	N	Y	D
	CONC	http://cubic.bioc.columbia.edu/~liu/conc/	N	Y	D
	CPC	http://cpc.cbi.pku.edu.cn/	N	Y	W, D
Generic structural RNA prediction and annotation	mfold	http://mfold.bioinfo.rpi.edu/	Y	Y	W, D
	RNAfold	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi	Y	Y	W, D
	QRNA	ftp://selab.janelia.org/pub/software/qrna/	Y	Y	D
	RNAz	http://www.tbi.univie.ac.at/~wash/RNAz/	Y	Y	W, D
	Evofold	http://www.soe.ucsc.edu/~jsp/EvoFold/	Y	Y	D
	Randfold	http://bioinformatics.psb.ugent.be/software/details/Randfold	Y	Y	D
	RNAstrand	http://www.bioinf.uni-leipzig.de/Software/RNAstrand/	Y	Y	D
Specific ncRNA classes prediction and annotation	MiPred	http://www.bioinf.seu.edu.cn/miRNA/	Y	N	W
	MirFinder	http://www.bioinformatics.org/mirfinder/	Y	N	D
	MirEval	http://tagc.univ-mrs.fr/mireval/	Y	N	W
	mirDeep	http://www.mdc-berlin.de/en/research/research_teams/systems_biology_of_gene_regulatory_elements/projects/miRDeep/	Y	N	D
	Snoscan	http://lowelab.ucsc.edu/snoscan/	Y	N	W, D
	snoGPS	http://lowelab.ucsc.edu/snoGPS/	Y	N	W, D
	snoSeeker	http://genelab.zsu.edu.cn/snoseeker/	Y	N	W, D
	snoReport	http://www.bioinf.uni-leipzig.de/Software/snoReport	Y	N	D
miRNA target prediction	miRanda	http://www.microRNA.org/ http://microrna.sanger.ac.uk/(miRBase targets)	Y	N	W, D
	TargetScan	http://www.targetscan.org/	Y	N	W
	PicTar	http://pictar.bio.nyu.edu/	Y	N	W
	DIANA-microT	http://diana.cslab.ece.ntua.gr/microT/	Y	N	W
	RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	Y	N	W, D
	mirTarget2	http://mirdb.org/miRDB/	Y	N	D
	miTarget	http://cbit.snu.ac.kr/~miTarget/	Y	N	W
	PITA	http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html	Y	N	W, D
	microTar	http://tiger.dbs.nus.edu.sg/microtar/	Y	N	D
snoRNA target prediction	snoTarget	http://hsc.utoledo.edu/depts/bioinfo/snotarget.html	Y	N	W, D

W = web-based resource, D = downloadable software or data.

In conclusion, by combining multiple discriminating features, SVM methods seem to outperform previous approaches and currently represent the forefront of protein-coding potential calculators.

One of the main problems for accurate protein-coding potential determination, independently from the specific method used, is the completeness of the input sequences. Indeed, reliable classification of novel transcripts into mRNAs or ncRNAs crucially depends on the full-length status of the input sequences. Several experimental variables, such as incomplete reverse transcription, internal priming of pre-mRNAs and genomic DNA contamination, can all result in the generation of spurious or truncate transcripts, many of which are likely to masquerade as ncRNAs.

It should also be noted that methods that assess the protein-coding potential of a transcript are based on the assumption that an RNA can be unequivocally annotated as protein-coding or noncoding, while in several cases RNAs might be bifunctional, that is they can be translated into proteins but also work independently as regulatory RNAs [40]. Therefore, these methods are very useful for selection of a stringent dataset of ncRNAs while performing genome-wide scans; however, when it comes to the functional characterization of single transcripts, the presence of an ORF should not exclude *a priori* the existence of additional regulatory functions at the RNA level, and vice versa.

Annotation based on primary sequence conservation and genomic context

While many small RNAs are evolutionary conserved (e.g. miRNAs, snoRNAs), long ncRNA genes are difficult to identify based on comparative primary sequence analysis with known regulatory RNAs, due to their sequence divergence across phyla. In fact, known functional ncRNAs (such as *Xist* or *Air*) are on the whole poorly conserved and display < 70% identity between mouse and human, similar to the level of conservation observed with introns. However, they often retain stretches of higher conservation within their overall sequence, suggesting the presence of functional domains necessary for the interaction with their molecular targets [45]. Despite the poor conservation of the primary sequence, it has been noted that some ncRNAs tend to maintain genomic equivalent position in different species [46], or conserve their genomic organization, as in the case of *Xist* in mammals [13]. In other words, the site of transcription is conserved between human and mouse genomes, even if the nucleotide sequence is not. Hence, at least in some cases, it might be possible to identify positional equivalents of ncRNAs by comparative genomics approaches.

Several approaches are based on the fact that long ncRNAs often originate from complex transcriptional loci, in which the ncRNAs are coordinately transcribed with their associated protein-coding transcripts. Therefore, it is possible to predict on a genome scale the localization and putative functional roles of ncRNAs on the basis of their genomic context and their relationships with neighbouring protein-coding genes. In this way, computational pipelines have been developed to detect specific ncRNA families, such as intronic ncRNAs, bidirectional transcripts and *cis*- and *trans*-antisense transcripts [46–49].

BIOINFORMATIC APPROACHES FOR SMALL ncRNA STUDY

Several distinct classes of small ncRNAs (<200 nt) play important regulatory roles in diverse cellular processes (for an overview, see [6]). The wider group of small RNAs, which include miRNAs, siRNAs and piRNAs, interacts with members of the Argonaute/Piwi (Ago/Piwi) protein family to form ribonucleoprotein complexes that silence gene expression either at the transcriptional or post-transcriptional level. Different classes of small RNAs bind to distinct Ago/Piwi family members and have distinguishing features, such as length, precursor structure and mechanism of biogenesis. However, in all cases, small RNAs guide the sequence-specific recognition of target nucleic acids by hybridizing to perfect, or nearly perfect, complementary sites [7].

miRNAs are the most abundant class of Ago/Piwi interacting small RNAs, and therefore the majority of computational methods developed for small ncRNA research is focused on the genome-wide prediction of miRNAs and their targets. Moreover, miRNA prediction has been facilitated by the fact that their precursors possess a definite length and secondary structure. Many of the mature miRNAs appear to be highly conserved across species and originate by a two-step endonucleolytic process (Figure 2) starting from long primary transcripts (pri-miRNA) that contain extensive regions of

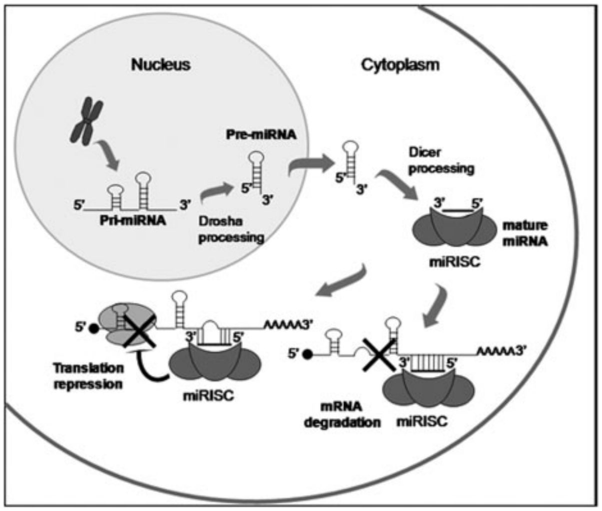

Figure 2: miRNA biogenesis and functions. Schematic representation of miRNA biogenesis in a eukaryotic cell. Pri-miRNA: miRNA primary transcript; Pre-miRNA: hairpin precursor miRNA; miRISC: miRNA induced silencing complex.

stem-loop structures. These stem-loops are excised by the RNaseIII enzyme Drosha in the nucleus to produce a ~70 nt hairpin precursor (pre-miRNA), which is then exported to the cytoplasm where it is further processed by Dicer to the mature miRNA [7]. During miRNA biogenesis, only miRNA derived from one strand of the RNA duplex is preferentially selected for entry into a silencing complex. The other strand, known as the miRNA*, has typically been assumed to be a carrier strand, but recent evidence demonstrated that, at least in *D. melanogaster*, miRNA* are often present at physiologically relevant levels and can associate with Argonaute proteins [50]. Once loaded into the silencing complex, miRNAs recognize and bind with a complex interaction or with perfect complementarity sequences within the 3′UTR or coding regions of their target mRNAs, blocking translation and/or inducing mRNA degradation [7].

Apart from Ago/Piwi interacting small RNAs, another large class of short regulatory RNAs is represented by snoRNAs, which direct the site-specific modification (2'-*O*-methylation and pseudouridylation) of ribosomal RNAs and other RNAs, and appear to be also involved in alternative splicing regulation [6, 51]. SnoRNAs recognize target sequences by formation of a guide RNA duplex and recruit associated proteins that catalyze the corresponding modification at the target site. Generally, snoRNAs range between 60–300 nt in length, but only short sequences participate in target recognition via antisense interactions. The two main classes of snoRNAs, called C/D box and H/ACA box, possess distinct sequence and structural motifs that can be used for computational prediction.

Finally, it should be stressed that the discovery of novel classes of small ncRNAs is progressing incessantly. For instance, whole-genome tiling arrays have identified short transcripts (20–200 nt), which preferentially map at the transcription starting sites (promoter-associated short RNAs, PASRs) or at the transcription termination site (termini-associated short RNAs, TASRs) of about half of the known protein-coding genes [20]. In addition, several independent classes of small RNAs were recently identified from cDNA libraries made from size-fractionated (20–40 nt) RNA [52]. Currently there are no established rules for the annotation of these RNAs. Moreover, it is likely that many small RNA classes still remain to be discovered, because some methods, such as CAGE (Cap-analysis of Gene Expression), PETs (Paired-End diTags), and full-length cDNAs, identify only capped RNAs, while tiling arrays exclude repeat elements [2]. Also, possible 5′ and 3′ chemical modification of small RNAs are a major issue for the cloning and discovery of novel families with currently used protocols. We expect that rapid advance in massively parallel sequencing technology will facilitate discovery of many novel classes of small RNAs, especially with concomitant advances in RNA purification methods, which would allow separation of modified RNA as well as depletion of major classes of known structural RNAs, such as tRNAs and short rRNAs.

We briefly review in the following sections the most common bioinformatic approaches for miRNA and snoRNA identification and target prediction.

Small RNA prediction

There are two mainstream directions for identification of new small ncRNA sequences: *ab initio* prediction methods and reverse strategies based on the inference of reliable candidate sequences starting from experimental data, usually deep sequencing of small RNA libraries.

A first challenge which faces the researcher interested in *ab initio* miRNA prediction is that the hairpin structure which is typical of the pre-miRNA precursor (Figure 2) is frequently found in the genome. However, the majority of the miRNA sequences clearly exhibit a folding free energy that is considerably lower than randomly shuffled sequences, indicating a high tendency towards a stable secondary structure [53]. The program *RandFold* (Figure 1D) is a downloadable software for ncRNA secondary structure prediction which could be included in a miRNA *ab initio* identification pipeline. An alternative approach, *MiPred*, uses a novel machine-learning technique to identify putative miRNA precursors and seem to provide elevated sensitivity and specificity [54]. However this method cannot scale up to the analysis of multiple sequences, so it is suitable only to work with a single putative miRNA precursor at a time. Recently, a large-scale clustering method was developed that allowed to determine the specific topological features of miRNA precursors and use them as a miRNA prediction tool, which can be used to screen thousands of putative stem-loop structures at a time [55]. In addition, this method has been implemented in a user-friendly online tool called *MirEval* that enables researchers with limited bioinformatics skills to conduct a thorough analysis of an input sequence for novel miRNAs [56]. The software *MiRFinder* [57], which can be freely downloaded, starts from pairwise genome comparison data and uses SVM to predict with good sensitivity and sensibility miRNA hairpin precursors from the raw genomic sequence. Finally, the frequent clustering of miRNA sequences in the genome has been successfully used as criteria for the detection of novel miRNAs in proximity of known ones, although the software is not downloadable [58].

Unlike miRNAs, both C/D box and H/ACA box snoRNAs, which direct two distinct types of chemical modifications of the target RNA molecules, have proved to be surprisingly difficult to find in genomic sequences. The first snoRNA-searching programs, namely *SNOSCAN* and *snoGPS* [59, 60], were essentially based on detecting guide snoRNAs, which target rRNAs or snRNAs. Recently, two different methods have been developed to look for all kind of snoRNAs, including 'orphan' snoRNAs, on a genome scale: *snoSeeker* [61] and *snoReport* [62]. *snoSeeker* is composed of two distinct programs, *ACAseeker* and *CDseeker*, and was designed to screen whole genomic alignments for putative snoRNA candidates as well as search for putative target sites. It is based on homology information and on the use of a number of probabilistic models to assess box elements, terminal stem pairing and complementary regions. Instead, *snoReport* used a combination of RNA secondary structure prediction and machine learning approaches to detect both C/D and H/ACA box snoRNAs, independently of the presence of alignment/homology information.

Small RNA discovery and annotation from deep-sequencing data

Massively parallel sequencing technologies have proved very promising for new small RNAs discovery, as demonstrated by recent landmark papers particularly focusing on miRNAs [63–65].

The correct identification of putative new miRNAs in the sea of small transcripts derived by one of such experiments, however, is a complex and almost 'artistic' procedure composed of many steps. Briefly, it requires careful primer and carrier removal, high stringency mapping of the mature miRNA to the genome (allowing for 3' editing), retrieval of the genome sequence corresponding to a putative pre-miRNA, secondary structure prediction of the putative precursor and filtering of other known ncRNAs (i.e. fragments of tRNAs or rRNAs, snoRNAs, piRNAs etc.). The sequences remaining after this 'cleaning' step are then clustered with known miRNAs. Clustering—with known miRNAs or only with novel sequences—is an additional evidence for a reliable new miRNA sequence, since about half of all miRNAs are clustered. All mapped reads are also related with existing genome annotation and are divided in intronic/exonic and conserved/nonconserved according to the overlap with phastCons conservation score. Finally, comparison with the two major miRNA databases (*miRBase* and *smiRNAdb*), allowing again for 3' editing, results in a reliable classification of the sequences in known, conserved and putative novel miRNAs. A general bioinformatic pipeline for the identification of new miRNA sequences from deep-sequencing data is illustrated in Figure 3.

An interesting 'probabilistic' approach to the identification of novel miRNA sequences from deep-sequencing data, freely downloadable, is *miRDeep* (Figure 1E). It is based on the differential count of reads deriving from the miRNA, the miRNA* or the loop parts of the precursors, together with the usual precursor secondary structure prediction, and has shown robustness in experimental validation [66]. It is also worth to mention the approach recently used by Lu and colleagues [67] to identify miRNAs from deep-sequencing data of *Drosophila* small RNA libraries: after a filtering

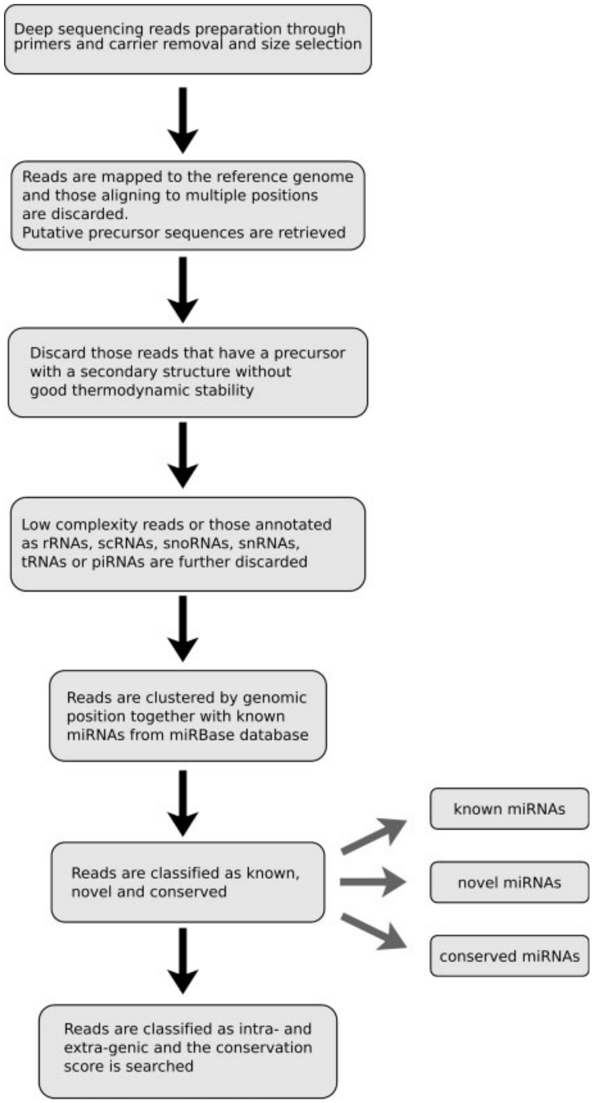
Deep sequencing reads preparation through primers and carrier removal and size selection
Reads are mapped to the reference genome and those aligning to multiple positions are discarded.
Putative precursor sequences are retrieved
Discard those reads that have a precursor with a secondary structure without good thermodynamic stability
Low complexity reads or those annotated as rRNAs, scRNAs, snoRNAs, snRNAs, tRNAs or piRNAs are further discarded
Reads are clustered by genomic position together with known miRNAs from miRBase database
Reads are classified as known, novel and conserved
known miRNAs
novel miRNAs
conserved miRNAs
Reads are classified as intra- and extra-genic and the conservation score is searched

Figure 3: Schematic description of a bioinformatic pipeline for *ab initio* identification of novel miRNAs from deep-sequencing data.

step of all reads matching known RNAs other than miRNAs (i.e. rRNAs, tRNAs, snoRNAs), the remaining reads and flanking sequences were subjected to secondary structure prediction combined with either of the three different parameter sets of increasing structural stringency to define putative hairpin precursors.

Small RNA target identification

An important step for the functional annotation of a small RNA is the identification of its physiological targets and ultimately the elucidation of entire biological pathways controlled by each small RNA. Most small RNAs contain sequence elements which are complementary to specific sites within their target RNAs. For instance, plant miRNAs hybridize to perfectly complementary target sequences within the coding sequence or 3'UTR of target mRNAs, while animal miRNAs require nearly perfect complementarity as well as a complex series of physical interactions and energetic constraints. In addition, the interaction of the miRNA with its functional targets is dependent on the local secondary structures of the target 3'UTR.

Many computational methods have been developed to predict miRNA targets [68] and are listed in Table 2. The basic target prediction is based on sequence complementarity and/or on favorable miRNA-target duplex thermodynamics. Additional criteria vary widely, but generally include (i) strong Watson-Crick base-pairing of the 5' seed of the miRNA (nt positions 2-8) to a complementary site in the 3'UTR of the mRNA; (ii) conservation of the miRNA binding site; and/or (iii) structural accessibility of the target. Although all these features are known to be important for effective miRNA-target interaction, the relative importance of each feature and how they contribute to function remains uncertain. Moreover, it is likely that other important parameters for functional miRNA-target interactions remain to be identified. For instance, binding cooperativity is currently an underestimated factor in most target prediction algorithms, but it would be increasingly important for improving the accuracy and efficiency of predictions. Moreover, target prediction of animal miRNAs should probably not be confined to 3'UTRs, as targeting of coding regions has been recently demonstrated in at least two different mammalian systems [69, 70].

Two of the most recent miRNA target predictors, *MicroTar* [71] and *PITA* [72], do not rely on evolutionary conservation of the binding site but do achieve a very good performance in specificity and sensitivity. This is particularly interesting in the case of tissue-specific miRNAs, which usually do not show conservation of the binding site. *MicroTar* calculates predicted free energies of unbound mRNA and putative mRNA-miRNA heterodimers, implicitly addressing the accessibility of the mRNA 3'UTR, while PITA uses a parameter-free model that computes the difference between the free energy gained from the formation of the miRNA-target duplex and the energetic cost of unpairing the target to make it accessible to the miRNA. *PITA* also takes into account the binding cooperativity and calculates a 'target score' for each miRNA, representing the combined effect of all predicted sites for that miRNA on the given UTR. Both programs can be downloaded and *PITA* is also available for single-query searches from a website, which includes a convenient target database for known miRNAs (Table 2).

As each miRNA has hundreds of putative mRNA targets, a promising computational field is the functional annotation of target genes of differentially expressed miRNAs. This could be achieved with tools and strategies similar to those employed for microarray analysis, or even with methods as simple as contingency analysis applied to gene ontology functional categories for detecting target gene enrichment [73]. Alternatively, an integrated website for functional annotation and profiling of known miRNA sequences is *miRGator* [74]. MiRNA function is inferred from the list of target genes predicted by a series of software, and statistical enrichment test of target genes in each term is performed for gene ontology, pathway and disease associations.

SnoRNAs target other RNAs for chemical modification through short stretches of sequence complementarity, which are located in definite positions of the snoRNA sequence [75]. Although snoRNAs predominantly target ribosomal RNAs and spliceosomal RNAs, the discovery of 'orphan' snoRNAs, which either have no known target or which target ordinary protein-coding mRNAs, suggests that they might play a diverse set of regulatory functions. For instance, a search with a recently developed computational web resource, *snoTARGET*, for possible guiding sites for orphan snoRNAs among the entire set of human and rodent exonic and intronic sequences, identified putative targets of HBII-85 C/D box snoRNAs within mRNAs, preferentially located in alternatively spliced exons [76].

DATABASES AND BIOINFORMATIC TOOLS FOR ncRNA ANNOTATION

Until recently, the non-protein coding portion of the genome was largely ignored by the public genome annotation repositories, mainly because these RNAs were considered as transcriptional noise. Consequently, genome-wide annotation of long ncRNAs in the most widely used genome browsers (i.e. UCSC genome browser or ENSEMBL) is present but still incomplete, and the information is not always easy to access. To overcome these limitations, and to complement the information present in the genome browsers, several databases specialized in the annotation of the non-protein coding portion of the transcriptome have been recently developed (Table 1):

(1) *RNAdb* collects data from different sources, including high confidence curated ncRNAs from literature and FANTOM3 ncRNA set, as well as computationally predicted conserved RNA structures [43]. This database can be queried in various ways: users can simply browse the collection or perform specific searches using keywords as well as by applying filters across nominated fields (i.e. species, disease-association, known or unknown function of the transcript). BLAST searches allow users to locate regions of similarity between sequences of interest and those stored in the database. NcRNA sets can be downloaded from *RNAdb* either as FASTA or XML files, for local viewing and data mining. In addition, entire datasets or search results can be returned as a custom track (BED file) to be directly loaded into the UCSC genome browser.

(2) *NONCODE* is a database of a wide variety of ncRNA classes (small and long ncRNAs) from 861 organisms covering all kingdoms of life (eukaryotes, eubacteria, archea and viruses) [44]. Data derive from three sources: (a) manual extracts from literature, (b) automatically filtered and manually confirmed Genbank sequences, and (c) experimental data from Chen's laboratory. The database can be browsed by species or ncRNA class, or searched using specific keywords. BLAST searches as well as a test version of UCSC Genome Browser for *NONCODE* are also available. The entire database can be downloaded as FASTA files for local viewing and data mining.

(3) Noncoding RNA Expression Database (*NRED*) is a brand new public repository of ncRNA expression, based on experimental results from various microarray and *in situ* hybridization platforms, including thousands of long ncRNAs in human and mouse [77]. The database can be queried by applying filters across nominated fields to select expression results based on probe characteristics and/or the values of the expression data. For instance, a user might extract a set of candidate ncRNAs which are significantly expressed in a tissue of interest. Search results can be viewed

online or downloaded as a table or a UCSC custom track. Results tables can be highly customized by including several information on the probe characteristics, genomic context of the target, sequence conservation and secondary structure prediction obtained by *RNAz*.

(4) The functional RNA Database 3.0 (*fRNAdb*) is a recent and a very interesting addition to the growing list of sequence-based comprehensive ncRNA databases [78]. Its main strength points are a vast and deeply annotated dataset (510075 entries in the current release) and a very neat interface with a local mirror of the UCSC Genome Browser. This is very useful in order to place a given ncRNA in its correct genomic context.

(5) Three specialized databases collect *cis*- and *trans*-natural antisense transcript predictions in several eukaryotic species: *NATsDB*, *Trans-SAMap* and *antiCODE* [49, 79, 80]. In this case, the focus is on the possible regulatory mechanism through which RNAs might act (i.e. sense–antisense base pairing), independently from the RNA coding potential. Therefore, these databases include both protein-coding and noncoding antisense transcript, and sense-antisense pairs are classified into three groups: coding-coding, coding-noncoding and noncoding-noncoding. The databases can be browsed according to species, overlapping pattern, coding potential and chromosome location, or searched using keywords; *antiCODE* also permits BLAST searches. Original datasets can be downloaded for local viewing and data mining.

(6) Known and cloned miRNA and snoRNA sequences are available not only from the above mentioned general ncRNA databases (*Rfam*, *RNAdb*, *NONCODE*, *fRNAdb*), but also from dedicated repositories (*miRBase*, *snoRNABase*, *Plant snoRNA database*) and genome annotation databases (NCBI, UCSC, etc.). In particular, *MiRbase* [81] is the central repository of mature miRNA sequences and related hairpins, and is well known in the miRNA research community. miRBase also includes the MicroCosm web resource, for searching computationally predicted miRNA targets across many species. Likewise, *snoRNABase* [82] and the *Plant snoRNA database* are dedicated databases containing sequences of C/D box and H/ACA box snoRNAs as well as their target sites on ribosomal and spliceosomal RNAs.

Recently, additional resources for the functional annotation of miRNAs have become available, due to the public release of high-throughput miRNA expression data, thus facilitating the use of information by biologists or bioinformaticians. Notable applications allow the user to associate expression profiles (in one or more tissues) of a subset of user-defined miRNAs to other information such as genome locations, host genes, predicted or validated target mRNAs. We briefly describe the principal features of few noteworthy databases:

(1) *SmiRNAdb* collects miRNA sequences and expression profile data obtained from cloning experiments by the Tuschl Lab, through a complex computational pipeline [83]. The database allows the search and visualization of the expression profiles of user-defined miRNAs in selected tissues. Results can be downloaded as a picture (pdf file) of the obtained heat map.

(2) MiRNA expression profiles are also available at the site www.microRNA.org [84], a resource which features attractive visualization either by heat maps or bar graphs, together with information about miRNA targets as predicted by the software *miRanda*.

(3) *Argonaute* allows the user to browse or search miRNAs on several criteria, including expression and association to human diseases [85]. Moreover, it provides links to published literature where information about miRNA expression data can be retrieved and compared with *de novo* generated profiles.

(4) *Tarbase* [86], a database of validated miRNA-mRNA interactions, includes the description of the validation method, the functional consequence of the interaction (cleavage or repression) and the source of information.

Concerning other small RNA classes, piRNA sequences can be retrieved from general ncRNA repositories, such as *RNAdb* and *NONCODE*, whereas short ncRNAs (e.g. PASR, TASR) derived from Affimetrix tiling array [20] are available as specific tracks and tables on the hg18 UCSC Genome Browser.

CONCLUSIONS AND PERSPECTIVES

Although the number and type of resources for ncRNA research are rapidly increasing and becoming more effective, there is currently no tool that allows the reliable annotation of all kinds of RNAs. On the bright side, however, many computational strategies are complementary, and it is usually possible and suitable to combine a set of methods to obtain reliable predictions.

Typically, the choice among different methods is also dependent on the explicit research aim. For instance, several specific tools are available for the study of known ncRNA classes, such as miRNAs or snoRNAs, and they are often preferable to general approaches, in order to maximize the speed and the sensitivity of the analysis. However, since it is likely that many unknown ncRNA families are still to be discovered, more general methods for searching ncRNAs are also desirable. In this case, several methods incorporating evolutionary models have recently achieved promising results.

Currently, small RNA research is more advanced than the study of long ncRNAs, thanks to the increasing number of biochemical studies elucidating the mechanisms of their function, and to novel deep-sequencing techniques, which have proved extremely useful for the analysis of this portion of the transcriptome. Nonetheless, we have only explored the tip of the iceberg: up to now, we have mainly looked at the most conserved and abundantly expressed small RNAs, while recent evidence suggests that many more are expressed in a tissue-specific manner and at a very low level. Therefore, the next computational challenge will be the reliable identification of novel small RNA candidates with low false positive rate and high sensitivity. Further progress is also important in the field of target recognition, as the general principles so far governing miRNA target recognition and mode of action are being progressively challenged by genetic and biochemical studies [87].

Finally, an increasingly important field in the bioinformatics of small RNAs is the integration of mRNA and miRNA expression data in order to understand the molecular networks in which each miRNA is involved.

In conclusion, ncRNA research is still at its infancy, as our knowledge about the diverse and complex ncRNA world is far from complete. Novel insights and speculations on this world may help us not only to increase our limited knowledge, but also to improve computational tools for further discoveries. At this stage, resources that are able to integrate different kind of data and information, either from *in silico* or experimental analyses, will become increasingly useful.

Key Points

- Noncoding RNAs are an emerging and diverse area of research, which challenge our current knowledge of eukaryotic transcriptomes and require novel computational approaches.
- Currently, prediction of long (>200 nt) and small (<200 nt) RNAs is better achieved with specific methods, as no tool allows the reliable identification of all ncRNA classes.
- Small ncRNAs are annotated on the basis of definite length, secondary structure, sequence motifs and conservation, while long ncRNAs annotation methods also rely on the discrimination between protein-coding and non-protein-coding transcripts.
- Computational approaches for small ncRNAs research are currently more advanced than those for long ncRNAs, due to a better understanding of their biology. Several tools are available for small RNA prediction, annotation, and functional characterization through target prediction.

References

1. Bompfünewerer AF, Flamm C, Fried C, *et al*. Evolutionary patterns of non-coding RNAs. *Theory Biosci* 2005;**123**: 301–69.
2. Carninci P, Yasuda J, Hayashizaki Y. Multifaceted mammalian transcriptome. *Curr Opin Cell Biol* 2008;**20**: 274–80.
3. Brannan CI, Dees EC, Ingram RS, *et al*. The product of the H19 gene may function as an RNA. *Mol Cell Biol* 1990; **10**:28–36.
4. Brockdorff N, Ashworth A, Kay GF, *et al*. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* 1992;**71**:515–26.
5. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993;**75**:843–54.
6. Mattick JS, Makunin IV. Small regulatory RNAs in mammals. *Hum Mol Genet* 2005;**14**:R121–32.
7. Farazi TA, Juranek SA, Tuschl T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* 2008;**135**:1201–14.
8. Bertone P, Stolc V, Royce TE, *et al*. Global identification of human transcribed sequences with genome tiling arrays. *Science* 2004;**306**:2242–6.
9. Johnson JM, Edwards S, Shoemaker D, *et al*. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* 2005;**21**:93–102.

10. Stolc V, Samanta MP, Tongprasit W, *et al*. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. *Proc Natl Acad Sci USA* 2005;**102**:4453–8.
11. Li L, Wang X, Stolc V, *et al*. Genome-wide transcription analyses in rice using tiling microarrays. *Nat Genet* 2006;**38**: 124–9.
12. Willingham AT, Gingeras TR. TUF love for "junk" DNA. *Cell* 2006;**125**:1215–20.
13. Prasanth KV, Spector DL. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev* 2007;**21**:11–42.
14. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* 2008; **9**:102–14.
15. Flynt AS, Lai EC. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat Rev Genet* 2008;**9**:831–42.
16. Feng J, Bi C, Clark BS, *et al*. The *Evf-2* noncoding RNA is transcribed from the *Dlx-5/6* ultraconserved region and functions as a *Dlx-2* transcriptional coactivator. *Genes Dev* 2006;**20**:1470–84.
17. Wang X, Arai S, Song X, *et al*. Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. *Nature* 2008;**454**:126–30.
18. Lanz RB, McKenna NJ, Onate SA, *et al*. A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* 1999;**97**:17–27.
19. Martianov I, Ramadass A, Serra Barros A, *et al*. Repression of the human dihydrofolate reductase gene by a noncoding interfering transcript. *Nature* 2007;**445**:666–70.
20. Kapranov P, Cheng J, Dike S, *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 2007;**316**:1484–8.
21. Wilusz JE, Freier SM, Spector DL. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 2008;**135**:919–32.
22. Hüttenhofer A, Schattner P, Polacek N. Noncoding RNAs: hope or hype? *Trends Genet* 2005;**21**:289–97.
23. Torarinsson E, Sawera M, Havgaard JH, *et al*. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* 2006;**16**:885–9.
24. Griffiths-Jones S. Annotating noncoding RNA genes. *Annu Rev Genom Hum G* 2007;**8**:279–98.
25. Gardner PP, Daub J, Tate JG, *et al*. Rfam: updates to the RNA families database. *Nucleic Acids Res* 2009;**37**:D136–40.
26. Athanasius F Bompfünewerer Consortium, Backofen R, Bernhart SH, *et al*. RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zoolog B Mol Dev Evol* 2007;**308**:1–25.
27. Machado-Lima A, del Portillo HA, Durham AM. Computational methods in noncoding RNA research. *J Math Biol* 2008;**56**:15–49.
28. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 2000;**16**:583–605.
29. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001;**2**:8.
30. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 2005; **102**:2454–9.
31. Pedersen JS, Bejerano G, Siepel A, *et al*. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2006;**2**:e33.
32. Meyer IM, Miklós I. Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res* 2005;**33**: 6338–48.
33. Okazaki Y, Furuno M, Kasukawa T, *et al*. Analysis of the mouse transcriptome based on functional annotation of 60,770 fulllength cDNAs. *Nature* 2002;**420**:563–73.
34. Frith MC, Bailey TL, Kasukawa T, *et al*. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA Biol* 2006;**3**:40–8.
35. Hatzigeorgiou AG, Fiziev P, Reczko M. DIANA-EST: a statistical analysis. *Bioinformatics* 2001;**17**:913–9.
36. Lottaz C, Iseli C, Jongeneel CV, *et al*. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* 2003;**19**(Suppl 2):ii103–12.
37. Gish W, States DJ. Identification of protein coding regions by database similarity search. *Nat Genet* 1993;**3**:266–72.
38. Mignone F, Grillo G, Liuni S, *et al*. Computational identification of protein coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res* 2003;**31**:4639–45.
39. Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* 1999; **16**:512–24.
40. Dinger ME, Pang KC, Mercer TR, *et al*. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comp Biol* 2008;**4**:e1000176.
41. Liu J, Gough J, Rost B. Distinguishing protein-coding from noncoding RNAs through support vector machines. *PLoS Genet* 2006;**2**:e29.
42. Kong L, Zhang Y, Ye ZQ, *et al*. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007;**35**: W345–9.
43. Pang KC, Stephen S, Dinger ME, *et al*. RNAdb 2.0 – an expanded database of mammalian noncoding RNAs. *Nucleic Acids Res* 2007;**35**:D178–82.
44. He S, Liu C, Skogerbø G, *et al*. NONCODE v2.0: decoding the noncoding. *Nucleic Acids Res* 2008;**36**:D170–2.
45. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* 2006;**22**:1–5.
46. Engstrom PG, Suzuki H, Ninomiya N, *et al*. Complex loci in human and mouse genomes. *PLoS Genet* 2006;**2**:e47.
47. Nakaya HI, Amaral PP, Louro R, *et al*. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol* 2007;**8**:R43.
48. Zhang Y, Liu XS, Liu QR, *et al*. Genome-wide in silico identification and analysis of *cis* natural antisense transcripts (*cis*-NATs) in ten species. *Nucleic Acids Res* 2006;**34**: 3465–75.
49. Li JT, Zhang Y, Kong L, *et al*. Trans-natural antisense transcripts including noncoding RNAs in 10 species:

implications for expression regulation. *Nucleic Acids Res* 2008;**36**:4833–44.

50. Okamura K, Phillips MD, Tyler DM, *et al*. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat Struct Mol Biol* 2008;**15**:354–63.

51. Kishore S, Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 2006; **311**:230–2.

52. Kawaji H, Nakamura M, Takahashi Y, *et al*. Hidden layers of human small RNAs. *BMC Genomics* 2008;**9**:157.

53. Bonnet E, Wuyts J, Rouzé P, *et al*. Evidence that microRNA precursors, unlike other noncoding RNAs, have lower folding free energies than random sequences. *Bioinformatics* 2004;**20**:2911–7.

54. Jiang P, Wu H, Wang W, *et al*. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res* 2007;**35**:W339–44.

55. Ritchie W, Legendre M, Gautheret D. RNA stem-loops: to be or not to be cleaved by RNAse III. *RNA* 2007;**13**: 457–62.

56. Ritchie W, Théodule FX, Gautheret D. Mireval: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics* 2008;**24**:1394–6.

57. Huang TH, Fan B, Rothschild MF, *et al*. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics* 2007;**8**:341.

58. Sewer A, Paul N, Landgraf P, *et al*. Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics* 2005;**6**:267.

59. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. *Science* 1999;**283**:1168–71.

60. Schattner P, Barberan-Soler S, Lowe TM. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA* 2006;**12**:15–25.

61. Yang JH, Zhang XC, Huang ZP, *et al*. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res* 2006;**34**:5112–23.

62. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* 2008;**24**:158–64.

63. Berezikov E, Thuemmler F, van Laake LW, *et al*. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* 2006;**38**:1375–7.

64. Berezikov E, van Tetering G, Verheul M, *et al*. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res* 2006;**16**:1289–98.

65. Takada S, Berezikov E, Yamashita Y, *et al*. Mouse microRNA profiles determined with a new and sensitive cloning method. *Nucleic Acids Res* 2006;**34**:e115.

66. Friedländer MR, Chen W, Adamidi C, *et al*. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008;**26**:407–15.

67. Lu J, Shen Y, Wu Q, *et al*. The birth and death of microRNA genes in Drosophila. *Nat Genet* 2008;**40**:351–5.

68. Mazière P, Enright AJ. Prediction of microRNA targets. *Drug Discov Today* 2007;**12**:452–8.

69. Forman JJ, Legesse-Miller A, Coller HA. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci USA* 2008;**105**:14879–84.

70. Tay Y, Zhang J, Thomson AM, *et al*. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* 2008;**455**:1124–8.

71. Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics* 2006;**7**: S20.

72. Kertesz M, Iovino N, Unnerstall U, *et al*. The role of site accessibility in microRNA target recognition. *Nat Genet* 2007;**39**:1278–84.

73. Gusev Y, Schmittgen TD, Lerner M, *et al*. Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. *BMC Bioinformatics* 2007;**8**(Suppl 7):S16.

74. Nam S, Kim B, Shin S, *et al*. miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res* 2008;**36**:D159–64.

75. Kiss T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell* 2002;**109**: 145–8.

76. Bazeley PS, Shepelev V, Talebizadeh Z, *et al*. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene* 2008;**408**:172–9.

77. Dinger ME, Pang KC, Mercer TR, *et al*. NRED: a database of long noncoding RNA expression. *Nucleic Acids Res* 2009;**37**:D122-6.

78. Mituyama T, Yamada K, Hattori E, *et al*. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res* 2009;**37**: D89–92.

79. Zhang Y, Li J, Kong L, *et al*. NATsDB: Natural Antisense Transcripts DataBase. *Nucleic Acids Res* 2007;**35**: D156–61.

80. Yin Y, Zhao Y, Wang J, *et al*. AntiCODE: a natural sense-antisense transcripts database. *BMC Bioinformatics* 2007;**8**:319.

81. Griffiths-Jones S, Saini HK, van Dongen S, *et al*. miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008;**36**: D154–8.

82. Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 2006;**34**:D158–62.

83. Landgraf P, Rusu M, Sheridan R, *et al*. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 2007;**129**:1401–14.

84. Betel D, Wilson M, Gabow A, *et al*. The microRNA.org resource: targets and expression. *Nucleic Acids Res* 2008;**36**: D149–53.

85. Shahi P, Loukianiouk S, Bohne-Lang A, *et al*. Argonaute – a database for gene regulation by mammalian microRNAs. *Nucleic Acids Res* 2006;**34**:D115–8.

86. Papadopoulos GL, Reczko M, Simossis VA, *et al*. The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* 2009;**37**:D155–8.

87. Brodersen P, Voinnet O. Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* 2009;**10**:141–8.