

# miROrtho: computational survey of microRNA genes

Daniel Gerlach<sup>1,2</sup>, Evgenia V. Kriventseva<sup>1</sup>, Nazim Rahman<sup>1</sup>,  
Charles E. Vejnár<sup>1,2</sup> and Evgeny M. Zdobnov<sup>1,2,3,\*</sup>

<sup>1</sup>Department of Genetic Medicine and Development, University of Geneva Medical School, <sup>2</sup>Swiss Institute of Bioinformatics, 1 Rue Michel-Servet, 1211 Geneva, Switzerland and <sup>3</sup>Imperial College London, South Kensington Campus, SW7 2AZ London, UK

Received August 19, 2008; Revised September 26, 2008; Accepted September 29, 2008

## ABSTRACT

MicroRNAs (miRNAs) are short, non-protein coding RNAs that direct the widespread phenomenon of post-transcriptional regulation of metazoan genes. The mature ~22-nt long RNA molecules are processed from genome-encoded stem-loop structured precursor genes. Hundreds of such genes have been experimentally validated in vertebrate genomes, yet their discovery remains challenging, and substantially higher numbers have been estimated. The miROrtho database (<http://cegg.unige.ch/mirortho>) presents the results of a comprehensive computational survey of miRNA gene candidates across the majority of sequenced metazoan genomes. We designed and applied a three-tier analysis pipeline: (i) an SVM-based *ab initio* screen for potent hairpins, plus homologs of known miRNAs, (ii) an orthology delineation procedure and (iii) an SVM-based classifier of the ortholog multiple sequence alignments. The web interface provides direct access to putative miRNA annotations, ortholog multiple alignments, RNA secondary structure conservation, and sequence data. The miROrtho data are conceptually complementary to the miRBase catalog of experimentally verified miRNA sequences, providing a consistent comparative genomics perspective as well as identifying many novel miRNA genes with strong evolutionary support.

## INTRODUCTION

MicroRNAs (miRNAs) represent an abundant class of short non-protein coding RNAs that direct post-transcriptional regulation of metazoan genes through repression of mRNA translation or transcript degradation. Since their initial discovery in

*Caenorhabditis elegans*, the roles of miRNAs have been recognized as a widespread phenomenon, implicated in processes such as cell differentiation and cancer (1–6). Intensive studies have begun to unravel the mechanisms and characteristics of these single-stranded, ~22-nt long RNA molecules that are processed from genome-encoded precursor genes with a defining stem-loop RNA structure. Nevertheless, the discovery and characterization of novel miRNA genes have proved to be challenging both experimentally and computationally, and the miRNA gene repertoire therefore remains largely unexplored. The human genome tops the fast growing number of miRNA genes, with several hundreds now cataloged in the miRBase database of published miRNA sequences (7) and many more estimated (8,9).

The high-throughput experimental approaches usually identify only the short mature segments of the miRNA genes along with other types of endogenous small RNAs (10,11) and degradation products of mRNAs or structural RNAs. Robust computational post-processing of the experimentally derived sequences is therefore essential to identify the underlying miRNA genes. The widely applied discriminatory requirement of a characteristic stem-loop structure for the putative precursor is, however, insufficient as hairpin structures are common in eukaryotic genomes and are not a unique feature of miRNAs (12). Nonetheless, the rapid accumulation of genome-wide sequencing data provides another line of evolutionary evidence from comparative sequence analyses.

Computational screening methods that rely heavily on sequence conservation criteria, such as MirScan (13), were among the first to appear. These characteristically exhibit high specificity [e.g. predicting 35 new miRNA candidates in *C. elegans* (13) and 107 in human (14), many of which were experimentally confirmed], but their sensitivity, the ability to predict novel or divergent homologs in other organisms, is low. Methods that relax sequence conservation requirements in favor of conservation patterns specific to miRNAs (such as a more diverged loop sequence and a more conserved hairpin stem) gained

\*To whom correspondence should be addressed. Tel: +41 22 379 59 73; Fax: +41 22 379 57 06; Email: [evgeny.zdobnov@unige.ch](mailto:evgeny.zdobnov@unige.ch)

substantially higher sensitivity, e.g. Snarloop has been used to predict 214 candidate miRNAs in *C. elegans* (15) and miRSeeker (16) to predict 48 candidate miRNAs in *Drosophila melanogaster*. A similar approach was proposed that takes into account the shapes of conservation patterns of known miRNAs, e.g. phylogenetic shadowing (17,18). The first 7 nt from the second position of the 5'-end of the mature miRNA, termed the seed sequence, are presumed to be critical for the interaction between the miRNA and its targets (19–22). The intra-species abundance or inter-species conservation of such potential seeds have also been proposed as alternative starting points for miRNA gene hunting (23,24).

Secondary structure thermodynamic stability is another important characteristic that can be used to distinguish miRNAs from other hairpins (25). The recently developed software RNAz combines thermodynamic stability and conservation of secondary structure to predict non-coding RNAs (26) from multiple alignments of orthologous regions. Methods relying on phylogenetic conservation of miRNA structure and sequence are by definition restricted in terms of their predictive power. To overcome this limitation, several groups have developed *ab initio* approaches (12,27–32) to predict novel, non-conserved genes. However, these approaches often suffer from high rates of false positives.

Aiming to fuel further studies of microRNA'omes, we present here the database of computationally derived miRNA gene candidates using a novel comparative genomics approach coupled with machine-learning techniques that we consistently applied to a comprehensive set of available metazoan genomes. The three-tier pipeline consists of: (i) a custom designed SVM-based *ab initio* predictor, plus screening for known miRNA homologs, (ii) an orthology delineation procedure and (iii) an SVM-based classifier of the multiple sequence alignments of the putative orthologs. These data are conceptually complementary to the miRBase catalog of experimentally verified miRNA sequences (7). High-throughput experimental exploration of small RNAs requires rigorous follow-up bioinformatic analyses to claim evidence of microRNA genes. Decoupling experimental and bioinformatics approaches, the miOrtho data effectively provide independent supporting evidence for the numerous ongoing experimental interrogations of microRNA'omes.

## MATERIAL AND METHODS

### *Ab initio* predictors

The first tier of our analysis pipeline is a novel *ab initio* miRNA prediction procedure. We scanned the genomic sequences using RNALfold (33) for locally stable hairpins characteristic of miRNA precursors, requiring a length of 60–120 nt, a minimum free-folding energy less than  $-15$  kcal/mol, a stem of 20–60 base pairs, a maximal interior loop size of 8 nt, and a maximum bulge loop size of 5 nt. The loop, however, was allowed to include short stem-loops e.g. hsa-let-7b. Those properties accommodate the vast majority of experimentally validated miRNAs

(although there are exceptions, e.g. dme-mir-31b and dme-mir-1017). As stem-loop structures are abundant and not exclusive to miRNA genes, this step yields hundreds of millions of candidates: 1.3 million for the  $\sim 170$  Mb genome of fruitfly *Drosophila melanogaster*. The availability of many experimentally validated miRNAs revealed that although there are biases in biophysical properties of miRNA stem-loops in comparison to non-miRNA sequences, such as higher thermodynamic stability (25), no clear discriminatory features have yet been identified. We investigated a number of the most discriminating features, such as the minimum free-energy index (34) or the mean base pair distance in the ensemble of structures, and trained an SVM (support vector machine) classifier using LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>). The total number of features used for this first SVM was 253. The radial basis function kernel (RBF) was used on 1000 experimentally verified animal pre-miRNAs from miRBase (7) and a negative set of 3000 potent stem-loops from other confirmed ncRNAs [Rfam (35)]. Optimal parameters for the RBF kernel (C-SVC  $c = 2.0$ ,  $\gamma = 0.03125$ ) were estimated using a heuristic approach implemented in grid.py, which is a part of the LIBSVM package. A non-redundant training dataset was compiled using CD-Hit-EST (36) at a cutoff of 90% sequence identity. We tested the performance of the SVM on a test set of 237 miRNA sequences and 568 non-miRNA stem-loops which were not used for training the SVM model. Using the SVM posterior probability cutoff at 0.5, the accuracy was estimated to be 95.03%, the area under the ROC curve (receiver operating characteristic) was 0.984, corresponding to a sensitivity and specificity of 0.84 and 0.97, respectively. Using a 10-fold cross-validation procedure on the training data, we received an average AUC (area under the ROC curve) of 0.982. If the potent hairpins had  $>70\%$  sequence overlap at the same locus, the one with the lower SVM score was discarded.

This single sequence SVM filter allows the space of likely candidates to be reduced by about 95%, yet still yields rather high numbers of gene candidates: 42 000 for *D. melanogaster*. The miRNA structure itself is likely to contribute to these elevated numbers: miRNAs have complementary arms in their stem-loop structure and the reverse complement of a precursor often also folds into a stable RNA hairpin. Nevertheless, we did not explicitly require a choice between the sense and the anti-sense candidates (if both of them passed the other filters) as there is evidence of miRNA loci with both strands yielding a functional miRNA, e.g. dme-mir-iab-4 and dme-mir-iab-4as.

### Homology-based predictor

Screening for homologs of currently known miRNAs (miRBase 11.0) captures putative miRNAs that either did not pass the stem-loop screen, e.g. 13 (8%) of known *D. melanogaster* miRNAs, or failed the *ab initio* SVM classification, another 19 (13%). Our procedure initially performs a WU-BLAST (<http://blast.wustl.edu>) search using the default parameters, plus the DUST

filter and the `hspsepSmax = 30` option, which defines the maximal separating distance between two high score pairs to allow for a varying loop while still matching the better conserved 5' and 3' arms. Next, blast hits longer than 20 nt are extended at both ends to match the length of the query sequence. These hits are further filtered using a minimum free energy filter ( $\leq -15$  kcal/mol) and a RANDFOLD (25) filter ( $P \leq 0.05$  on 100 sequence randomizations). We investigated the RNASHAPES (37) filter, which predicts the probability of a sequence to fold into a simple stem-loop like structure, but it was not employed as several known miRNAs, e.g. *hsa-let-7a-1*, would not pass the filter. The candidate miRNAs were then aligned to the query sequence using MAFFT (38) and the conservation of the seed region was calculated by mapping the known mature miRNA region on the query miRNA to the alignment. The hits were then tested for the following criteria: a 100% conserved seed region, >90% conservation of the putative mature part, and a total hairpin identity >65%. As close paralogs (like *hsa-let-7*, *mmu-let-7*, etc) can map to the same locus when searched again one genome (e.g. the chimp), the matches were then clustered using GALAXY (<http://main.g2.bx.psu.edu>) and choosing one representative with the lowest *e*-value of all queries.

### Orthology delineation

Groups of likely orthologous genes were automatically identified using a strategy employed previously for protein-coding genes (39) based on all-against-all sequence comparisons using the ParAlign algorithm (40) with NT2 substitution matrix; followed by clustering of best reciprocal hits (BRHs) from highest scoring ones to  $10^{-6}$  *e*-value cutoff for triangulating BRHs or  $10^{-10}$  cutoff for unsupported BRHs, and requiring a sequence alignment overlap of at least 20 nt across all members of a group. Furthermore, the orthologous groups were expanded by genes that are more similar to each other within a genome than to any gene in any of the other species, and by very similar copies that share over 97% sequence identity, which were identified initially using CD-Hit (36). The orthology filter allowed us to reduce the space of the miRNA candidates by a further 92%. Passing the orthology filter provides evolutionary support for the predicted miRNAs; however, detailed inspection highlighted the need for further rigorous sequence classification to remove questionable predictions.

### Multi-species conservation classifier

We further analyzed the R-COFFEE (41) multiple sequence alignments of orthologous groups of putative miRNA sequences. From the alignments we gathered the 13 most descriptive features for conservation properties of sequence, energy and structures such as: GC content, number of taxa, mean pairwise sequence identity, number of consistent mutations, conservation of the mature part, etc. Those descriptors were chosen among a larger set of features, in order to optimally describe the typical conservation profile of a miRNA gene family and to reduce false positive predictions. Alignments that mapped to at least one known miRNA from miRBase

11.0 were used as the positive training and testing sets (344 and 100 alignments, respectively). Among those alignments which did not map to any known miRNA family, we randomly selected (with manual checking) the negative training and testing sets (344 and 100 alignments, respectively). The GIST SVM software package (<http://www.cs.columbia.edu/compbio>) was used for training, testing and classification using the default parameter. The final set of newly predicted miRNAs based on the alignment SVM was selected from all alignments which had SVM score  $\geq 0.5$ , a 100% conserved seed, a mature part >90% conserved and having representatives in at least four taxa. Performance estimation of the alignment SVM on the independent test set showed an accuracy of 91%, with the area under the ROC curve (AUC) of 0.97, and sensitivity and specificity of 0.9 and 0.92, respectively. The AUC for the 10-fold cross validation using the training data averaged to 0.998. The alignment SVM filter allowed us to reduce the space of the miRNA candidates by a further 98%, followed limited manual curation of novel miRNA candidates. We further analyzed the multiple alignments of novel miRNAs (without known homologs) to predict the mature part using a sliding 23-nt long sliding window and scanning for the region with the highest information content in the 5' or the 3' arms. The predictions, however, should be taken with caution without further experimental support.

### DATABASE CONTENT

The miROrtho database (<http://cegg.unige.ch/mirortho>) presents computationally predicted putative miRNA genes for a comprehensive set of sequenced animal genomes (selection of genomes in Table 1), employing an in-house developed pipeline combining SVM-based classifiers and orthology delineation procedure adapted from OrthoDB (39). The alignments shown on the website were calculated using R-COFFEE (41), which combines MUSCLE (42), Probcons4RNA (43), MAFFT (38) and the secondary structures predicted by RNAplfold (33). Based on these alignments consensus secondary structures color-coded according to consistent/compensatory mutation were calculated using RNAalifold (44) which incorporates a ribosome scoring matrix suited for aligned RNA sequences. The database aims to provide a comprehensive comparative perspective on the animal repertoire of miRNA genes with direct reference to the putative ortholog multiple alignments, RNA secondary structure conservation, etc. As there seem to be numerous lineage specific miRNAs and miRNA-like sequences that are difficult to differentiate without experimental evidence, we see miROrtho as complementary to miRBase, the repository of experimentally verified miRNA sequences. Overall, miROrtho contains 7887 putative miRNA genes that are homologous to known miRNAs in miRBase 11.0, and 1437 confident predictions that are as yet without experimental support or homology to known miRNAs. Most experimental surveys provide support for mature miRNA sequences, while the identities of the underlying miRNA precursor genes remain somewhat uncertain.

**Table 1.** Analyzed genomes

Species name	Abbreviation	Size (Mb)	Number of miRNA genes			Source
			Homologs <sup>a</sup>	New <sup>b</sup>	miRBase 11.0	
<i>Aedes aegypti</i>	Aaeg	1384	58	1	0	AaegL1
<i>Anopheles gambiae</i>	Agam	273	55	1	45	AgamP3
<i>Apis mellifera</i>	Amel	235	60	1	54	Amel_4.0
<i>Bombyx mori</i>	Bmor	397	33	0	21	SW_scaffold_ge2k
<i>Caenorhabditis elegans</i>	Cele	100	149	0	154	WB170
<i>Canis familiaris</i>	Cfam	2532	383	138	203	CanFam 2.0
<i>Ciona intestinalis</i>	Cint	173	25	0	34	JGI2
<i>Danio rerio</i>	Drer	1626	324	22	337	ZFISH6
<i>Drosophila ananassae</i>	Dana	230	108	12	0	CAF1
<i>Drosophila erecta</i>	Dere	152	136	16	0	CAF1
<i>Drosophila grimshawi</i>	Dgri	200	110	13	0	CAF1
<i>Drosophila melanogaster</i>	Dmel	129	153	15	152	CAF1
<i>Drosophila mojavensis</i>	Dmoj	194	98	14	0	CAF1
<i>Drosophila persimilis</i>	Dper	188	108	16	0	CAF1
<i>Drosophila pseudoobscura</i>	Dpse	153	106	15	76	CAF1
<i>Drosophila sechellia</i>	Dsec	167	139	16	0	CAF1
<i>Drosophila simulans</i>	Dsim	142	131	15	0	CAF1
<i>Drosophila virilis</i>	Dvir	206	101	14	0	CAF1
<i>Drosophila willistoni</i>	Dwil	237	112	12	0	CAF1
<i>Drosophila yakuba</i>	Dyak	169	135	16	0	CAF1
<i>Gallus gallus</i>	Ggal	1100	168	49	149	WASHUC2
<i>Gasterosteus aculeatus</i>	Gacu	462	320	12	0	BROAD S1
<i>Homo sapiens</i>	Hsap	3665	626	151	678	NCBI36
<i>Macaca mulatta</i>	Mmul	3097	530	145	464	MMUL_1
<i>Monodelphis domestica</i>	Mdom	3606	205	82	119	monDom5
<i>Mus musculus</i>	Mmus	2661	505	117	472	NCBIM36
<i>Ornithorhynchus anatinus</i>	Oana	2073	207	57	0	Oana-5.0
<i>Pan troglodytes</i>	Ptro	3524	546	147	100	PanTro 2.1
<i>Rattus norvegicus</i>	Rnor	2719	440	110	287	RGSC 3.4
<i>Strongylocentrotus purpuratus</i>	Surc	907	13	0	0	Spur_y2.1
<i>Takifugu rubripes</i>	Trub	393	250	13	131	FUGU4
<i>Tetraodon nigroviridis</i>	Tnig	402	282	14	132	TETRAODON7
<i>Tribolium castaneum</i>	Tcas	200	37	1	0	Tcas_2.0
<i>Xenopus tropicalis</i>	Xtro	1511	351	24	184	JGI4.1

<sup>a</sup>Homologs to miRBase 11.0 miRNAs.<sup>b</sup>New predictions that do not show any homology to any annotated miRNA.


In contrast, computational procedures rely on recognizing characteristic sequence and structural properties of the precursors, where even approximate prediction of mature miRNAs is rarely possible. This complementarity extends further, where computational predictions at different stringencies can either be used to prioritize experimental verification, or as direct independent support of miRNAs identified through high throughput experimental screens. Although miRBase accepts annotation of very close homologs of experimentally supported miRNAs, the comparative perspective is heavily biased towards favorite experimental model species. Such a bias is avoided in miROrtho through the consistent application of the same procedures across all the available genomes, delineating groups of orthologous miRNAs over distantly related organisms. The miROrtho methodology has also been applied to the task of miRNA gene annotation in a number of ongoing initial genome analyses, and this database will provide the supporting information for these predictions.

It should be noted that there is still no defining feature that clearly discriminates between *bona fide* miRNA precursors and other abundant genomic sequences capable

of similar hairpin folding. Classification filters will therefore inevitably suffer from false negatives and false positives (see Materials and Methods section for estimates), leading to errors at each step along the pipeline. Even the most inclusive initial screen for locally stable stem-loop structures misses some miRNAs reported in miRBase as experimentally validated (e.g. dmemir-1017). Despite the strict 97% specificity of our *ab initio* SVM, the abundance of false positives is clear and overloads the orthology filter. Computational methods developed for miRNA gene discovery are constantly improving, and will continue to do so as our knowledge of experimentally validated miRNAs grows.


## WEB INTERFACE

The miROrtho database presents all predicted miRNA genes within the context of family groups of orthologous miRNAs. For each such family, we provide (Figure 1): (i) a table of annotated miRNA names and genomic coordinates, (ii) a multiple alignment of the miRNA sequences displaying RNA structure conservation, (iii) the minimum




**Zdobnov's Computational Evolutionary Genomics group**

[Home](#) [Data](#) [Services](#) [Links](#)



---



# http://cegg.unige.ch/mirortho


miOrtho: the catalogue of animal microRNA genes  
Your query returned 1 orthologous group(s)

Home Search by Genomic Location Browse Blast Search Help

Species	Internal ID	Name	Group ID	Family	miRBase	Chromosome	Start	End	Strand	UCSC
Agam	<a href="#">223776</a>	agam_223776	<a href="#">41740</a>			CM000357.1	22427276	22427349	+	
Aaeg	<a href="#">223782</a>	aaeg_223782	<a href="#">41740</a>			CH477494.1	120066	120145	-	
Cpip	<a href="#">223788</a>	cpip_223788	<a href="#">41740</a>			DS233046.1	55494	55573	-	
Tcas	<a href="#">223794</a>	tcas_223794	<a href="#">41740</a>			gij 74477186 gb CM000277.1	12045734	12045819	-	
Nvit	<a href="#">223800</a>	nvit_223800	<a href="#">41740</a>			SCAFFOLD26	605703	605787	+	
Amel	<a href="#">223806</a>	amel_223806	<a href="#">41740</a>			gnl Amel_4.0 Group4.16	137092	137186	+	

```

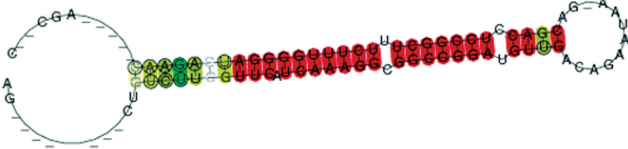
.....((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((.....
Agam_223776      CCGGCCAGGGGDUUCUCGGCCCTACAGCAUUGACCCUUGAGGCGCGCAAAACURUUSCUCCG-----U
Aaeg_223782      U--CGA-----AGGAGAGGGGDUUCUCGGCCCTACCGGCAUUGAAGAAUUUGAGGCGCGCAAAACURUUSCUCCGUC-----GU
Cpip_223788      A--CAU-----CCGAAUAGGGGDUUCUCGGCCCTACCGGCAUUGAAGAAUUUGAGGCGCGCAAAACURUUSCUCCGUC-----GU
Tcas_223794      G--UGA--CGCAAAGUAGGGGDUUCUCGGCCCTACCGGAAUUAUAAAGAUUGAGGCGCGCAAAACURUUSCUCCGUC-----GU
Nvit_223800      C--CGG--CAACAAAGUAGGGGDUUCUCGGCCCTACAGCU--AAGACAGCUGUAGGCGCGCAAAACURUUSCUCCGUC-----CGA
Amel_223806      GAACGUAGCGAUAAGUAGGGGDUUCUCGGCCCTACAGCA--ACGAUGCUGUAGGCGCGCAAAACURUUSCUCCGUC-----CGA
.....10.....20.....30.....40.....50.....60.....70.....80.....90.....
    
```



- Sequence conservation is represented by grey bars
- Mature miRNAs are underlined
- Data:
  - [Alignment reliability estimation \(core index\)](#)
  - [Groups with same seed](#)
  - [Alignment](#)
  - [Fasta Sequences](#)
  - [RNAalifold output | RNAalifold \(stochastic backtracking\)](#)
  - [RNAstrand output](#)

Types of pairs

	1	2	3	4	5	6
Incompatible pairs						
Compatible pairs						



[Disclaimer](#) [Statistics](#) [Funding](#)

**Figure 1.** miOrtho screenshot showing a novel miRNA gene family. The results page consists of three parts: (i) a table with detailed information about the individual miRNAs; (ii) a multiple sequence alignment with the consensus secondary structure displayed above in dot-bracket format and conservation profile bars displayed below, with the sequence of the mature miRNAs underlined; (iii) the consensus secondary structure of the orthologous sequences. Both alignment and consensus secondary structure are color-coded according to consistent and compensatory base changes.

energy consensus miRNA hairpin fold, (iv) FASTA sequences and multiple alignment files. Color coding of the alignments and the depicted folds enables clear visualization of compensatory and consistent mutations within a given miRNA family. The mature miRNA sequences are underlined: as annotated in miRBase for known miRNAs or as predicted for novel families. Furthermore, we provide detailed folding information of individual pre-miRNAs including minimum free energy folding, the partition function folding and the centroid structure of the stem-loop. Three images show the secondary structure of a single pre-miRNA with the mature part annotated in red, color-coded according to base pairing probabilities and positional entropy per position. The data can be browsed by the species tree, or can be queried by

annotation such as known families (e.g. let-7), identifiers or chromosomes. The predictions can be also searched by sequence homology using WU-BLAST (<http://blast.wustl.edu>).

### ACKNOWLEDGEMENTS

We thank R.M. Waterhouse for help with the article, and the Vital-IT facility (<http://www.vital-it.ch/vitalit-intro.htm>). We would also like to acknowledge the sequencing centers that made the genome sequences that were used for this study, available before publication: The Baylor College of Medicine ([www.hgsc.bcm.tmc.edu](http://www.hgsc.bcm.tmc.edu)), the Washington University School of Medicine

(genome.wustl.edu), the Broad Institute (www.broad.mit.edu), the J. Craig Venter Institute (www.jcvi.org), the DOE Joint Genome Institute (www.jgi.doe.gov), the Sanger Center (www.sanger.ac.uk), the Institute for Genomic Research (www.tigr.org), Celera Genomics (www.celera.com), and Genoscope (www.genoscope.cns.fr).

## FUNDING

Swiss National Science Foundation (SNF PDFMA3-118375 and 3100A0-112588). Funding for open access charges: Swiss National Science Foundation (SNF 3100A0-112588).

*Conflict of interest statement.* None declared.

## REFERENCES

- Ambros, V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
- Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
- Du, T. and Zamore, P.D. (2005) microPrimer: the biogenesis and function of microRNA. *Development*, **132**, 4645–4652.
- Calin, G.A. and Croce, C.M. (2006) MicroRNA signatures in human cancers. *Nat. Rev. Cancer*, **6**, 857–866.
- Zhang, B., Pan, X., Cobb, G.P. and Anderson, T.A. (2007) microRNAs as oncogenes and tumor suppressors. *Dev. Biol.*, **302**, 1–12.
- Barbarotto, E., Schmittgen, T.D. and Calin, G.A. (2008) MicroRNAs and cancer: profile, profile, profile. *Int. J. Cancer*, **122**, 969–977.
- Griffiths-Jones, S., Saini, H.K., van Dongen, S. and Enright, A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
- Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B. and Rigoutsos, I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203–1217.
- Berezikov, E., van Tetering, G., Verheul, M., van de Belt, J., van Laake, L., Vos, J., Verloop, R., van de Wetering, M., Guryev, V., Takada, S. et al. (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.*, **16**, 1289–1298.
- Kim, V.N. and Nam, J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
- Aravin, A. and Tuschl, T. (2005) Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.*, **579**, 5830–5840.
- Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E. et al. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.
- Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B. and Bartel, D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.
- Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. and Bartel, D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
- Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G. and Kim, J. (2003) Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell*, **11**, 1253–1263.
- Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of *Drosophila* microRNA genes. *Genome Biol.*, **4**, R42.
- Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. and Rubin, E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.
- Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H. and Cuppen, E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
- Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
- Doench, J.G. and Sharp, P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.*, **18**, 504–511.
- Brennecke, J., Stark, A., Russell, R.B. and Cohen, S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
- Stark, A., Brennecke, J., Russell, R.B. and Cohen, S.M. (2003) Identification of *Drosophila* MicroRNA targets. *PLoS Biol.*, **1**, E60.
- Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
- Weaver, D.B., Anzola, J.M., Evans, J.D., Reid, J.G., Reese, J.T., Childs, K.L., Zdobnov, E.M., Samanta, M.P., Miller, J. and Elisk, C.G. (2007) Computational and transcriptional evidence for microRNAs in the honey bee genome. *Genome Biol.*, **8**, R97.
- Bonnet, E., Wuyts, J., Rouze, P. and Van de Peer, Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
- Washietl, S., Hofacker, I.L. and Stadler, P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
- Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E. and Zavolan, M. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267.
- Xue, C., Li, F., He, T., Liu, G.P., Li, Y. and Zhang, X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
- Nam, J.W., Kim, J., Kim, S.K. and Zhang, B.T. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res.*, **34**, W455–W458.
- Helvik, S.A., Snove, O. Jr. and Saetrom, P. (2007) Reliable prediction of Droscha processing sites improves microRNA gene prediction. *Bioinformatics*, **23**, 142–149.
- Ng, K.L. and Mishra, S.K. (2007) De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, **23**, 1321–1330.
- Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X. and Lu, Z. (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339–W344.
- Hofacker, I.L., Priwitzer, B. and Stadler, P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
- Zhang, B.H., Pan, X.P., Cox, S.B., Cobb, G.P. and Anderson, T.A. (2006) Evidence that miRNAs are different from other RNAs. *Cell Mol. Life Sci.*, **63**, 246–254.
- Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. (2006) RNASHAPES: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
- Katoh, K. and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.*, **9**, 286–298.
- Kriventseva, E.V., Rahman, N., Espinosa, O. and Zdobnov, E.M. (2008) OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.*, **36**, D271–D275.
- Saebo, P.E., Andersen, S.M., Myrseth, J., Laerdahl, J.K. and Rognes, T. (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.*, **33**, W535–W539.

41. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-Coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Res.*, **36**, e52.
42. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
43. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.
44. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.