Nucleic Acids Research Advance Access published May 27, 2011

Nucleic Acids Research, 2011, 1–6
doi:10.1093/nar/gkr357

DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments

Mario Fasold[1,2], David Langenberger[1,2], Hans Binder[1,2], Peter F. Stadler[1,2,3,4,5,6,7] and Steve Hoffmann[1,2,*]

[1]Interdisciplinary Center for Bioinformatics and Bioinformatics Group, Department of Computer Science, University Leipzig, Härtelstrasse 16-18, 04107 Leipzig, [2]LIFE, Leipzig Research Center for Civilization Diseases, University Leipzig, Philipp-Rosenthal-Strasse 27, 04107 Leipzig, [3]Max-Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany, [4]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, 04103 Leipzig, Germany, [5]Department of Theoretical Chemistry, University of Vienna, Währinger Straße 17, 1090 Wien, Austria, [6]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark and [7]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received February 25, 2011; Revised April 20, 2011; Accepted April 26, 2011

ABSTRACT

Small non-coding RNAs (ncRNAs) such as microRNAs, snoRNAs and tRNAs are a diverse collection of molecules with several important biological functions. Current methods for high-throughput sequencing for the first time offer the opportunity to investigate the entire ncRNAome in an essentially unbiased way. However, there is a substantial need for methods that allow a convenient analysis of these overwhelmingly large data sets. Here, we present DARIO, a free web service that allows to study short read data from small RNA-seq experiments. It provides a wide range of analysis features, including quality control, read normalization, ncRNA quantification and prediction of putative ncRNA candidates. The DARIO web site can be accessed at http://dario.bioinf.uni-leipzig.de/.

INTRODUCTION

High-throughput sequencing (HTS) using a small RNA preparation protocol (small RNA-seq) was primarily designed to measure the expression of microRNAs. Closer inspection of the resulting sequence libraries, however, revealed that many other ncRNA types are chopped into RNA molecules of microRNA-like length, and are hence detectable in the sequencing data as well (1). Some of the non-miRNA sources of short RNA sequences include tRNAs (tRNA-derived fragments) (2–4), snoRNAs (snoRNA-derived small RNAs) (5), 21U-RNAs (6) or snRNAs (1). Recently, small RNA sequencing has helped to identify new RNA species such as microRNA offset RNAs (moRs), which derive from miRNA precursors. Although they have first been described in the simple chordate *Ciona intestinalis* (7), they could be verified in mammalian transcriptomes (8) and have later been linked to Kaposi's sarcoma-associated Herpesvirus (9,10).

Hence, small RNA-seq data contain a plethora of processing and maturation products potentially including yet unknown RNA species. Despite this fact, many small RNA-seq data analysis tools such as miRanalyzer (11), miRDeep (12) or miRNAkey (13) focus on microRNAs—largely neglecting other types of RNAs. In addition, these programs are often restricted to specific sequencing platforms due to embedded mapping algorithms. Other tools such as deepBase do not allow the upload of own experimental data (14).

In addition to finding new RNA species, the expression levels of ncRNAs have been shown to be associated with a number of different phenotypes. Various forms of neoplastic diseases such as colorectal cancer (15), for instance, show changes in miRNA expression levels. Likewise, differential snoRNA expression has been found in a study with menigioma cells (16). RNA quantification is possible using tools such as rQuant.web (17) or RSEQTools (18); however, they are not readily applicable to small ncRNA analysis as annotation data must be collected from different sources.

We have combined a ncRNA prediction method (1,8) with tools to quantify ncRNAs in a completely platform independent and easy to use web tool. DARIO performs RNA-seq quality controls and quantifies RNA

*To whom correspondence should be addressed. Tel: +49 341 9716711; Fax: +49 341 9716679; Email: steve@bioinf.uni-leipzig.de

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

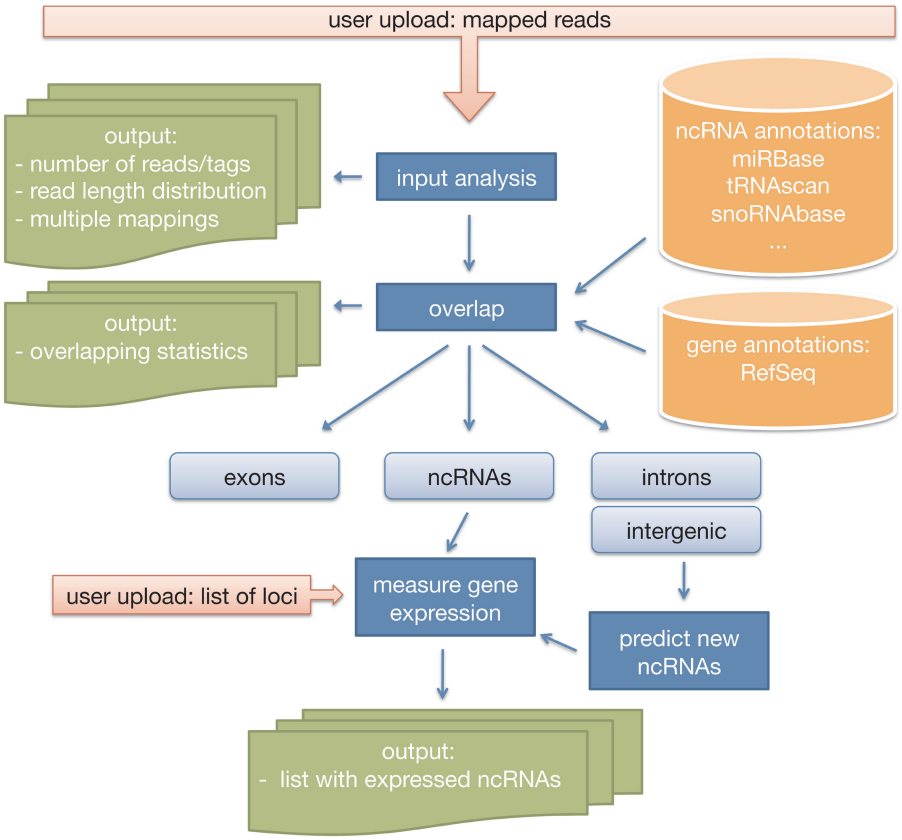


Figure 1. Simplified workflow of a DARIO computation. After the user upload, the data are run through some quality checks with regard to read lengths distributions and multiple mappings. Subsequently, the mapping loci are overlapped with ncRNA annotation data for gene expression measuring. A random forest classifier predicts new ncRNAs. The results of the analysis are easily accessible from a summary web page.

expression based on annotated ncRNAs from different ncRNA databases. The expression data and ncRNA predictions can be downloaded in the standardized BED format. We provide a script to locally convert SAM files and other mapping files to the BED format. The script is optimized to greatly reduce the amount of data that has to be uploaded to the DARIO server.

MATERIALS AND METHODS

Workflow

The DARIO web service requires previously mapped reads stored in compressed or uncompressed files in BAM or BED format. The uploaded file is uncompressed, if necessary, and examined for validity. A first analysis of the input data provides measures for quality control. The reads are then overlapped with various gene models of the selected species relevant for the analysis of small ncRNAs. Mapping loci overlapping with exonic regions are excluded from further analysis. Mapping loci overlapping with introns and intergenic regions are used to predict non-annotated ncRNAs. Finally, the results are summarized in HTML pages and data tables. A simplified workflow of the DARIO web service is depicted in Figure 1.

Sequence and annotation data

Genome assemblies of six supported species were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html): *Homo sapiens* (hg18, NCBI 36.1 and hg19, GRCh37), *Macaca mulatta* (rheMac2, MGSC Merged 1.0), *Mus musculus* (mm9, NCBI37), *Danio rerio* (danRer7, Zv9), *Drosophila melanogaster* (dm3, BDGP Release 5) and *Caenorhabditis elegans* (ce6, WUSTL School of Medicine GSC and Sanger Institute version WS190). For each assembly, we retrieved the UCSC Known Genes Track using the UCSC Table Browser in order to generate intron/exon lists.

ncRNA annotation was collected from several databases. While miRNA annotation was obtained from the miRBase v16 (19), most of the other ncRNA loci were downloaded from the UCSC Genome Browser. For human ncRNA data sets, we additionally included tRNA track (20), wgRNA track (21) for snoRNAs and the rnaGene track for other ncRNAs. For mouse, the tRNA track was used. For fly, our annotation encompasses the flyBaseNoncoding track from FlyBase (22). The sangerRnaGgene track containing WormBase annotations (23) is provided for Worm ncRNA data analysis. Where necessary, annotations were lifted to alternative assemblies with the UCSC tool liftover

(http://hgdownload.cse.ucsc.edu/downloads.html). Additional ncRNA annotations were collected from the Mouse Genome Database (24) as well as from Ensembl/BioMart for zebrafish (25). If tRNA or snoRNA annotations were not available, we predicted candidates using tRNAscan-SE (26) or snoReport (27), respectively.

Webserver implementation

The web site and the HTML results are created by a set of Python scripts and the Mako template engine. The jobs are scheduled in a queued fashion and distributed over a set of active machines. Upon completion, the results are transferred to the web server and available under a personalized link for 4 weeks. Mapping loci are merged to blocks based on their genomic positions and assembled to regions of blocks using blockbuster v1.0 (8) with default parameters. These are then classified using the random forest method in WEKA v3.6 (1,28,29). Graphics are created using R (30) and the ggplot2 graphics package (31). RNAz Version 1.0 (32) has been used to screen all supported assemblies for potential functional RNA structures. Predicted ncRNA candidates are overlapped with these screenings to provide RNAz support.

RESULTS AND DISCUSSION

The DARIO web site provides a simple web form that allows the user to specify and upload input data. The web site currently supports seven assemblies of six species: human (hg18, hg19), rhesus monkey (rheMac2), mouse (mm9), fruit fly (dm3), worm (ce6) and zebrafish (danRer6). After file upload, a job is created and queued for computation. The user may supply an email address to be notified upon job completion. A single job typically takes between 5 and 30 min. The results are summarized on a single web page containing job details, quality control measures and figures, ncRNA quantification and classification. All results can be downloaded for further analysis.

Input format

DARIO uses mapped sequences as input. The alignments may be provided in the common BAM or BED formats (http://genome.ucsc.edu/FAQ/FAQformat.html). The BED files require the fields for sequence identifier, strand and need to provide the read count in the score field. This format allows to collapse reads occurring multiple times into unique sequence tags, dramatically reducing space requirements of sequencing data. DARIO allows upload of (g)ZIPed files.

We provide a small, no-dependency perl script to convert SAM and SOAP format files into the BED input format. Virtually, all common mapping tools (segemehl, BWA, SOAP, Bowtie, etc.) can write their output alignment to either of these formats.

Using genome loci of previously mapped reads, and thus decoupling read alignment and analysis, has a number of advantages over using raw sequence reads. First, DARIO has no dependencies to any sequencing platform or mapping tool. Thus, read data originating from any sequencing platform and aligned with any mapping program can be used. Second, this greatly reduces the required amount of data to be uploaded to the server (e.g. 1GB SAM file → 15MB compressed BED file).

Quality control

There are numerous errors and biases that can occur during sample handling, library preparation and sequencing in a small RNA-seq experiment, rendering an assessment of the experiments quality a necessity (33–35). A basic set of figures (Figure 2) gives the researcher a first impression of the quality of the experiment. This includes the read length distribution, the number and occurrence of multiple mapped reads, the fraction of reads mapping to different genomic loci (exon, intron or intergenic) and ncRNA classes (miRNA, tRNA, snoRNA, etc.). Other measures include the number of mappable reads and the number of tags.

RNA quantification

For expression analysis, mapping loci are overlapped with annotated ncRNAs from a variety of sources. To handle multiple mappings, the number of reads for each sequence tag is divided by the number of its mapping loci. This normalized expression value is assigned to each mapping locus. These expression values are additionally normalized based on the absolute number of mappable reads (RPM), to allow subsequent differential expression analysis. Note that these measures do not necessarily reflect precursor ncRNA abundance as RNA processing and sequencing protocol lead to a non-uniform read distribution across the precursor RNA.

A list of expressed ncRNAs, itemized by ncRNA classes, is generated (Figure 3). The user obtains information about the normalized expression, the number of mapped reads (raw and multimap normalized), as well as a link to the UCSC genome browser for each expressed locus. The UCSC link helps the experimenter to quickly scan the data for new types of ncRNAs, e.g. microRNA-offset-RNAs (moRs) or vault RNAs, and to get a deeper understanding of the processing of these poorly understood ncRNA classes.

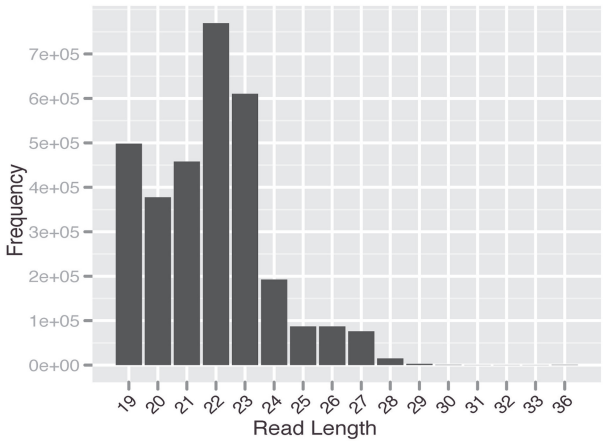
The web interface allows the upload of own annotation tracks. The specified regions are included in all downstream analysis. Predicted RNAs from previous DARIO runs can directly be used as user annotation.

Classification

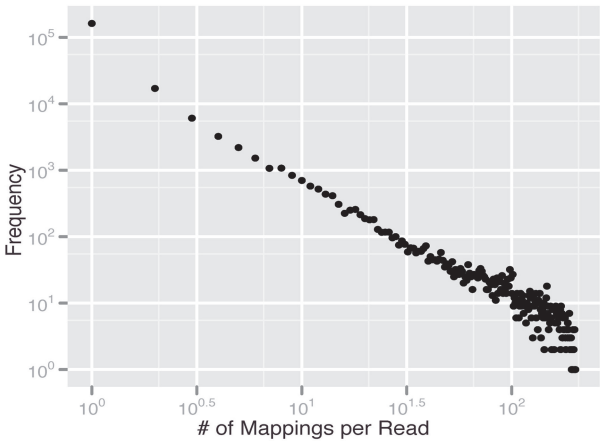
DARIO predicts new ncRNAs using a previously published machine learning approach (1). This method relies on characteristic read patterns exhibited by different classes of ncRNA. The classifier achieves positive predictive values (PPVs) and recall rates of 0.8. With recall rates varying from 0.6 to 0.7 and PPVs between 0.7 to 0.8, snoRNA predictions mark the lower bound of the classification [cf. Table 2 in (1)]. Reciever operator characteristic curves for all predicted ncRNAs in a number of species is given in the Supplementary Figure S1. For each candidate, a prediction score is given along with a RNAz

Quality Control

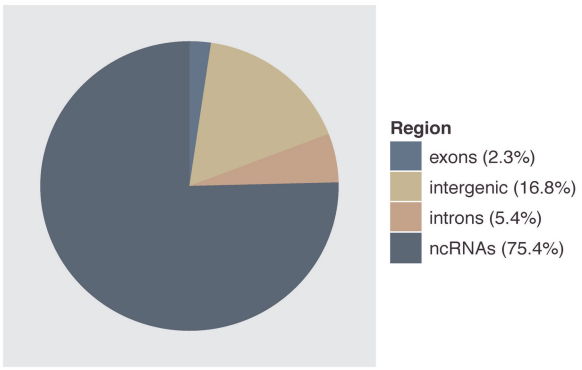
The following figures indicate whether problems might have occured during sample or library preparation.



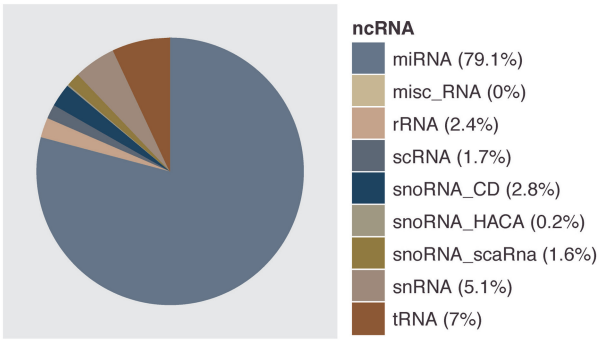
Read Length Distributions



Multiple Mappings Distribution



Read fractions mapping into genes



Read fractions mapping into ncRNAs

Figure 2. The DARIO web server provides a set of graphics for quality control. The figures show the read length distribution, the number of multiple mappings, the distribution of mapping loci across the genome and the annotated non-coding RNAs. The user may immediately check the success of his short RNA sequencing run in terms of capturing the ncRNA of interest.

ncRNA Class	Reads	Reads (normalized)	# of Genes	Table
miRNA	1,141,865	816,175.5	661	View List
snoRNA_CD	50,991	29,031.09	181	View List
snoRNA_HACA	2,125	2,019.533	85	View List
tRNA	433,796	72,332.05	537	View List

Chromosome	Start Loci	End Loci	Strand	ID	RPM	Reads	Reads (normalized)	Visualization
chr10	70184934	70185001	+	HBII-419	5.08e-02	5	5.00	View at UCSC
chr11	10779589	10779731	-	U97	1.07e+00	210	210.00	View at UCSC
chr11	17052775	17052867	-	U14B	1.76e-01	34	22.58	View at UCSC
chr11	17053900	17053991	-	U14A	3.55e+00	483	451.08	View at UCSC
chr11	46740514	46740625	-	HBII-166	5.24e-02	9	8.50	View at UCSC
chr11	62376957	62377083	-	U22	3.94e-01	72	69.50	View at UCSC
chr11	62377372	62377443	-	U31	3.48e-01	36	34.50	View at UCSC
chr11	62377710	62377780	-	U30	1.65e-01	16	16.00	View at UCSC
chr11	62377951	62378016	-	U29	2.21e-01	20	20.00	View at UCSC
chr11	62378668	62378743	-	U28	1.54e+01	1.682	1.607.00	View at UCSC

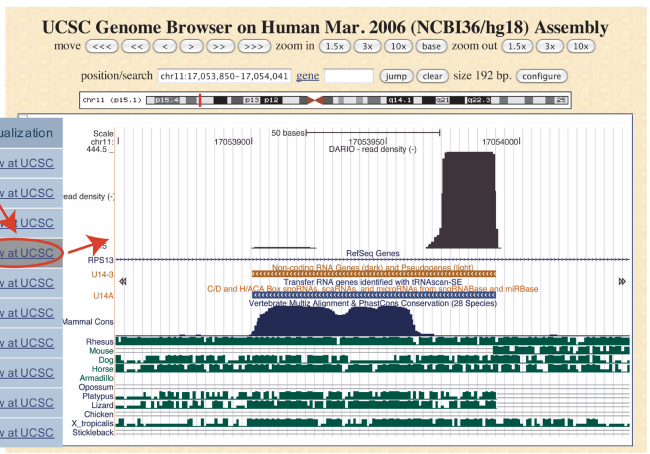


Figure 3. The DARIO analysis output is partitioned into different ncRNA classes. For each ncRNA class, a list that may be sorted by location, name or expression criteria is provided. A link to the UCSC genome browser allows the instantaneous inspection of the ncRNAs, in this case a snoRNA including available ncRNA annotation tracks and conservation.

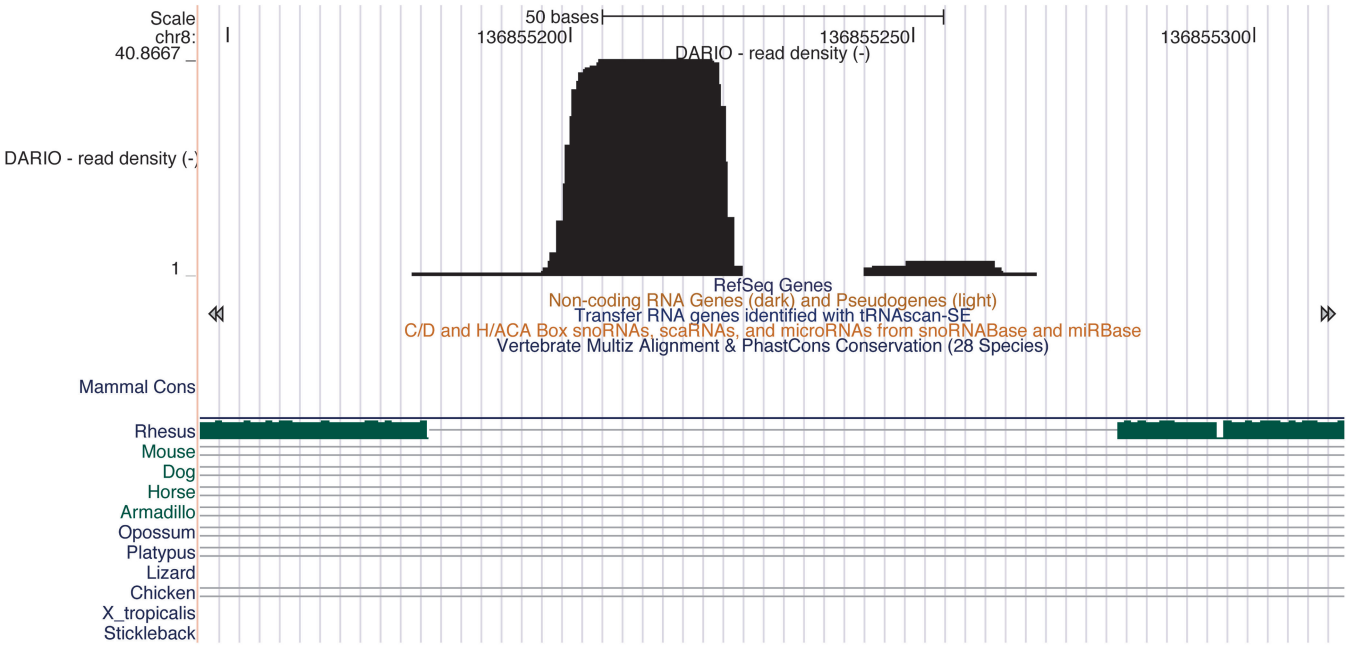


Figure 4. Example for a DARIO prediction for a miRNA. The integrated random forest classifier predicts a miRNA on the human chromosome 8 in an intergenic region. The expression pattern shows a typical miR and miR* processing product constellation. Interestingly, the UCSC browser reports neither annotations nor conservation at this position.

classification (32), if available. One of the candidate miRNAs predicted on the human chromosome 8 using the DARIO platform is shown in Figure 4. With the links to the UCSC genome browser, it is possible to instantaneously inspect the prediction by loading multiple different annotation tracks.

CONCLUSION

HTS offers wide-ranging possibilities for analyzing ncRNAs in an unprecedented way. However, deciphering the world of non-coding RNAs in HTS data requires tools that allow integrated analysis in a user-friendly way. We have developed the first integrated tool for the analysis and prediction of various small ncRNAs on user-provided RNA-seq data. The web service allows researchers to quickly grasp and assess the success of a short RNA-seq experiment. The web server overlaps the mapping loci with ncRNA genes from a number of ncRNA classes and annotation databases in order to quantify RNA abundance with different expression measures. Reads that do not map to annotated ncRNA genes are identified and classified. DARIO provides an easy to use web interface and thus greatly facilitates both initial evaluation and downstream analysis of read data originating from arbitrary sequencing platforms. Further versions of DARIO will allow to directly compare sets of small RNA transcriptomes to evaluate differences in expression levels of ncRNAs.

Availability and requirements

DARIO can be accessed freely via the web browser using the URL http://dario.bioinf.uni-leipzig.de/. There are no restrictions on use and no login requirement. It has been tested with several browsers and works with Safari, Firefox and Internet Explorer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Andreas Gruber for the inital web site template, Alexander Donath und Fabian Externbrink for their contribution to the backend of the web site and Christian Otto for his help with R routines.

FUNDING

This publication is supported by LIFE - Leipzig Research Center for Civilization Diseases, Universität Leipzig. European Social Fund and the Free State of Saxony. Funding for open access charge: LIFE Center for civilization Diseases funded by the State of Saxony and the European Union.

Conflict of interest statement. None declared.

REFERENCES

1. Langenberger,D., Bermudez-Santana,C., Stadler,P. and Hoffmann,S. (2010) Identification and classification of small RNAs in transcriptome sequence data. *Pac. Symp. Biocomput.*, **15**, 80–87.

2. Haussecker,D., Huang,Y., Lau,A., Parameswaran,P., Fire,A. and Kay,M. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, **16**, 673.
3. Lee,Y., Shibata,Y., Malhotra,A. and Dutta,A. (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, **23**, 2639.
4. Cole,C., Sobala,A., Lu,C., Thatcher,S., Bowman,A., Brown,J., Green,P., Barton,G. and Hutvagner,G. (2009) Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, **15**, 2147.
5. Taft,R., Glazov,E., Lassmann,T., Hayashizaki,Y., Carninci,P. and Mattick,J. (2009) Small RNAs derived from snoRNAs. *RNA*, **15**, 1233.
6. Ruby,J., Jan,C., Player,C., Axtell,M., Lee,W., Nusbaum,C., Ge,H. and Bartel,D. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, **127**, 1193–1207.
7. Shi,W., Hendrix,D., Levine,M. and Haley,B. (2009) A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.*, **16**, 183–189.
8. Langenberger,D., Bermudez-Santana,C., Hertel,J., Hoffmann,S., Khaitovich,P. and Stadler,P.F. (2009) Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, **25**, 2298–2301.
9. Umbach,J.L. and Cullen,B.R. (2010) In-depth analysis of Kaposi's sarcoma-associated herpesvirus microRNA expression provides insights into the mammalian microRNA-processing machinery. *J. Virol.*, **84**, 695–703.
10. Lin,Y., Kincaid,R., Arasappan,D., Dowd,S., Hunicke-Smith,S. and Sullivan,C. (2010) Small RNA profiling reveals antisense transcription throughout the KSHV genome and novel small RNAs. *RNA*, **16**, 1540.
11. Hackenberg,M., Sturm,M., Langenberger,D., Falcon-Perez,J.M. and Aransay,A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, 68–76.
12. Friedländer,M., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
13. Ronen,R., Gan,I., Modai,S., Sukacheov,A., Dror,G., Halperin,E. and Shomron,N. (2010) miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, **26**, 2615–2616.
14. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.*, **38**, D123–D130.
15. Lanza,G., Ferracin,M., Gafa,R., Veronese,A., Spizzo,R., Pichiorri,F., Liu,C.G., Calin,G.A., Croce,C.M. and Negrini,M. (2007) mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Mol. Cancer*, **6**, 54.
16. Chang,L.S., Lin,S.Y., Lieu,A.S. and Wu,T.L. (2002) Differential expression of human 5S snoRNA genes. *Biochem. Biophys. Res. Commun.*, **299**, 196–200.
17. Bohnert,R. and Ratsch,G. (2010) rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, **38**, W348–W351.
18. Habegger,L., Sboner,A., Gianoulis,T., Rozowsky,J., Agarwal,A., Snyder,M. and Gerstein,M. (2011) RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics*, **27**, 281.
19. 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.
20. Chan,P.P. and Lowe,T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, **37**, D93–D97.
21. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
22. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.*, **35**, D486–D491.
23. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
24. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A. and Eppig,J.T. (2003) MGD: the Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
25. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
26. Schattner,P., Brooks,A.N. and Lowe,T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**, W686–W689.
27. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
28. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
29. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I. (2009) The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18.
30. R Development Core Team. (2008) R: a language and environment for statistical computing. *R Found. Stat. Comput.*, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.
31. Wickham,H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York Inc.
32. Washietl,S., Hofacker,I. and Stadler,P. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454.
33. Dohm,J.C., Lottaz,C., Borodina,T. and Himmelbauer,H. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**, e105.
34. Linsen,S.E., deWit,E., Janssens,G., Heater,S., Chapman,L., Parkin,R.K., Fritz,B., Wyman,S.K., deBruijn,E., Voest,E.E. *et al.* (2009) Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods*, **6**, 474–476.
35. Hansen,K.D., Brenner,S.E. and Dudoit,S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.