W132–W138 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 22 April 2011*
doi:10.1093/nar/gkr247

miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments

Michael Hackenberg[1,2,*], Naiara Rodríguez-Ezpeleta[3] and Ana M. Aransay[3]

[1]Departmento de Genética, Facultad de Ciencias, Universidad de Granada, Campus de Fuentenueva s/n, 18071-Granada, [2]Laboratorio de Bioinformática, Centro de Investigación Biomédica (CIBM), PTS, Avda. del Conocimiento s/n, 18100-Granada and [3]Genome Analysis Platform, CIC bioGUNE, Parque Tecnológico de Bizkaia, 48160 Derio, Bizkaia, Spain

Received February 13, 2011; Revised March 26, 2011; Accepted April 5, 2011

ABSTRACT

We present a new version of miRanalyzer, a web server and stand-alone tool for the detection of known and prediction of new microRNAs in high-throughput sequencing experiments. The new version has been notably improved regarding speed, scope and available features. Alignments are now based on the ultrafast short-read aligner Bowtie (granting also colour space support, allowing mismatches and improving speed) and 31 genomes, including 6 plant genomes, can now be analysed (previous version contained only 7). Differences between plant and animal microRNAs have been taken into account for the prediction models and differential expression of both, known and predicted microRNAs, between two conditions can be calculated. Additionally, consensus sequences of predicted mature and precursor microRNAs can be obtained from multiple samples, which increases the reliability of the predicted microRNAs. Finally, a stand-alone version of the miRanalyzer that is based on a local and easily customized database is also available; this allows the user to have more control on certain parameters as well as to use specific data such as unpublished assemblies or other libraries that are not available in the web server. miRanalyzer is available at http://bioinfo2.ugr.es/miRanalyzer/miRanalyzer.php.

INTRODUCTION

Short non-coding RNA molecules such as microRNAs play important roles in the regulation of gene expression (1). They have been recognized as key players in many basic pathways, and their aberrant expression is implicated in numerous diseases such as cancer (2). With the advent of high-throughput sequencing (HTS) technologies, it is now possible to rapid and inexpensively measure the expression levels of known microRNAs and to improve the prediction of new microRNAs by including the expression data into the prediction models (3). Not in vain, the number of HTS experiments aiming to study microRNA expression has rapidly increased over the past few years. For example, the number of entries for '(Illumina Genome Analyzer) AND microRNA' in the GEO repository (4) are 12, 33 and 107 for 2008, 2009 and 2010, respectively. These numbers show a clear tendency that is expected to be even stronger when HTS techniques become cheaper and more accessible to everyone.

A number of algorithms have been developed in order to process these large amounts of data (5–8). Two years ago, we developed miRanalyzer (9), a tool for the detection of known and prediction of new microRNAs in HTS experiments. Here, we describe a new version of the tool, which has been completely redesigned and includes various new features. First, the alignments are now based on the ultrafast short read aligner Bowtie (10) that grants full colour space support, allows mismatches in the alignment of the read to the genome and is faster and more memory efficient than the previously implemented alignment algorithm. Second, the scope of the tool is extended to 31 species (including 6 plants) and allows to easily adding new ones. Third, the tool has no restriction on the number of input sequences for the prediction of new microRNAs, and the training of the prediction models takes into account differences between plant and animal microRNAs (11). Fourth, we have implemented a module, based on the DESeq package (12), to detect differential expression of microRNAs between two conditions. Additionally, taking advantage of the fact that

*To whom correspondence should be addressed. Tel: +34 958 243 261; Fax: +34 958 244 073; Email: mlhack@gmail.com

multiple samples are needed for this last module, we have also implemented the computation of the consensus sequences for predicted mature and precursor microRNAs. This will help assessing the reliability of the predictions, i.e. microRNAs predicted in different samples are more likely to be functional than those predicted in just one sample. Finally, we have prepared a standalone version of the miRanalyzer tool that works with an easily customized local file-based database.

miRanalyzer UPDATED

miRanalyzer workflow

Although some features have changed, the general workflow of the current version is broadly maintained (Figure 1). Two input formats are accepted: (i) read-count files (read sequences and counts tab separated), which can be generated from sequence or colour space fastq files using a provided perl script or generated by other means by the user, and (ii) multi-fasta files (see tutorial on the web page for more details). In a first step, the tool removes all reads with 'N' (or other irregular bases) and those shorter than 17 bases, and reads longer than 26 bases are trimmed and regrouped. The reads are then successively aligned to the corresponding species sequences in miRBase (to detect known microRNAs), the transcriptome (to detect mRNA contamination) and the genome (to predict new microRNAs). The mapping to miRBase is done in four substeps, aligning subsequently to mature, maturestar, unobserved maturestar and hairpin sequences. After each of these steps, the mapped reads are removed from the input file so they cannot erroneously be predicted as new microRNA. The reads that did not map to miRBase are successively aligned to transcriptome libraries (RefSeq and RFam). Among the aligned reads, those which map to more than N different entities within the same library are removed, i.e. will not be used in the following analysis steps. The parameter N is fixed to 5 in the web server, but can be modified in the stand-alone version. Finally, the remaining reads are mapped to the genome, and the alignments are used to predict new microRNAs following three steps: (i) clustering reads into putative mature microRNAs (see 'Data and Methods', 'Detection of read clusters' section); (ii) extracting candidate pre-microRNAs from the genome to select the energetically best candidate (see 'Data and Methods', 'Generating precursor candidates' section); and (iii) applying five different Random Forest models to calculate the probability that a given candidate is a microRNA (see 'Data and Methods', 'Prediction Models' section). The web server reports only those candidates having been predicted by at least three out of the five models. The predicted microRNAs can be viewed within a genome context by means of links to the UCSC Genome Browser and the NGSmethDB browser (13,14). Table 1 shows a summary of the miRanalyzer parameters.

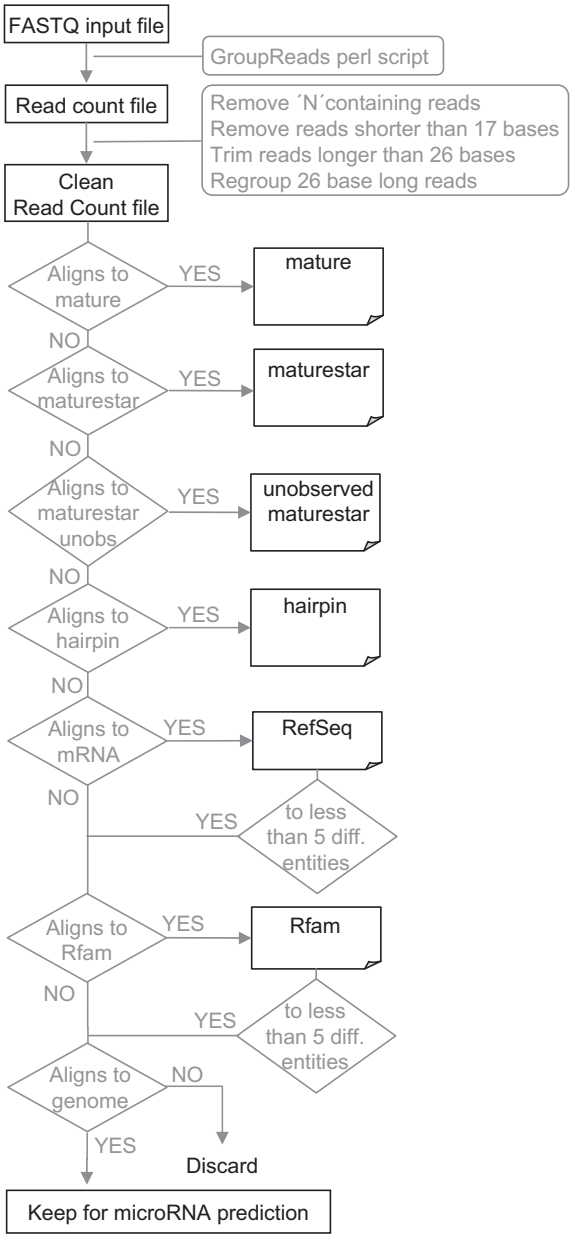

Figure 1. General workflow of miRanalyzer. The fastq file is transformed into a read count file, which is filtered to keep only sequences from 17 to 26 bases. These reads are successively mapped to several databases in order to identify known microRNAs, discard messenger RNA contaminations and select sequences for the microRNA prediction step.

Aligning the reads

The ultrafast short read aligner Bowtie (10) is used to align the reads to the different libraries and the genome, which allows, compared to the previous version, (i) the use of colour space sequences, (ii) a wider range of accepted mismatches and (iii) a gain in speed and memory efficiency. Bowtie requires a number of input parameters that define which alignments are legal and how many of them should be reported. Since no quality values do currently exist in the miRanalyzer input, the sum of the quality values at all mismatched read positions (−e/−maqerr) is set to an arbitrary value of 2000, which disables the quality values. Furthermore, we use the—best and—strata

Table 1. The default values of the parameters used in miRanalyzer are shown

General parameters				
Name	Description			Value
minLength	The minimum read length, all others will be removed			17
maxLength	The maximum read length, all reads will be trimmed to this length			26
Bowtie parameters				
Name	Description	miRBase	Trans. libraries	Genome
−k	Max. number of reported alignments	10	20	6
−l	The seed length	17	20	17
−n	The number of mismatches within the seed	1	1	1
Prediction parameters				
Name	Description			Value
Score	The posterior probability that the candidate is a true microRNA			0.9
minNoPositives	miRanalyzer predicts using five models (five different negative sets). This parameter determines the minimum number of models which predicts a candidate to be a new microRNA (default: 3).			3

The web server version allows the user to change the '−n' parameter. The standalone version allows manipulating all of them. We used −l 17 to detect known microRNAs and predict new microRNAs (align to the genome) as this is the shortest microRNA length in miRBase but −l 20 for the other libraries.

options to get only the best alignments, that is, those with least mismatches in the seed. Parameters −k (maximum number of reported alignments), −l (length of the seed) and −n (number of mismatches within the seed) are set to different values depending on the library used (Table 1). Note that in the case of alignments to the genome, we remove all reads with equal or more than −k valid best alignments. The reason is that reads with a high number of alignments are unlikely derived from microRNAs. When mapping against miRBase, transcriptome and Rfam databases, the −norc option (do not attempt to align against the reverse-complement reference strand) is used as the mappings to the reverse strand would be meaningless.

Bowtie detects the best alignments under a given seed length, to which the maximum number of allowed mismatches parameter is applied. Since microRNAs do not have a fixed size, a common seed cannot accommodate all possible length variants. In order to overcome this issue, we fix the seed to 17, the length of the shortest microRNA known, and post-process the Bowtie results to select the longest alignment that maintains the number of observed mismatches within the seed (see Figure 2 for details).

The range of input sequence length (17–26 nt) allows that reads can be longer than the reference sequence, i.e. the average length of known microRNAs is ~22 nt, something that Bowtie does not allow. In order to use Bowtie, a sequence of 25 'Gs' is artificially added to the known microRNA reference sequences. It might be that the seed alignment extends into this artificially added sequence. Those cases are detected and corrected (removing the matches between read bases and bases of the G-run) or removed (if after the correction the alignment is shorter than 17 nt).

Plant genomes

Although plant and animal microRNAs have a well-defined fold-back hairpin structure in common, plant pre-miRNAs have larger and more variable stem-loop structures (15). Therefore, the basic workflow is the same for plants and animals, but modifications have been introduced to take these differences into account when generating the prediction models (see 'Data and Methods', 'Prediction Models' section). In general, we found that the prediction models are much more accurate for animals compared to plants. In the training set, by means of a 10-fold cross-validation we obtain a mean recall (over all five models) and a mean precision of 0.923 and 0.892, respectively, for plants and of 0.978 and 0.965, respectively, for animals. Furthermore, given the wider range of pre-microRNA length in plants, the prediction for these organisms is more CPU intensive (many more secondary structures need to be calculated).

Differential expression

The differential expression module is based on the DESeq package, which is used to calculate the fold changes of the expression values and to assess its statistical significance (12). In order to use this module, all samples need to be processed first with miRanalyzer. The IDs assigned to each miRanalyzer job are then used to define the two groups to be compared. The DESeq input is a matrix where every entity (known and predicted microRNAs in our case) has an assigned read count for each sample. For known microRNAs, the names will be the same in each sample, which makes the matrix generation easy. However, the newly predicted microRNAs have arbitrarily assigned names such as Candidate_256 or Candidate_12, which normally do not coincide between samples, that is, Candidate_1 in sample 1 will not

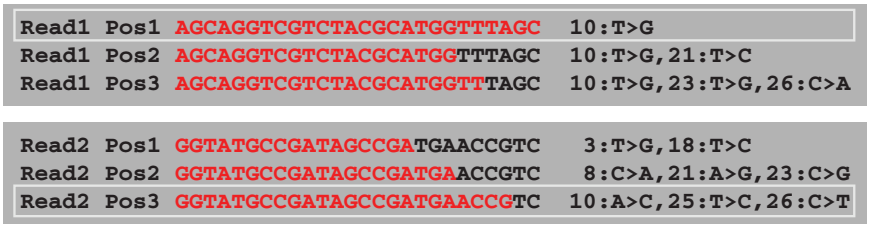

Figure 2. Selection of longest alignments performed by miRanalyzer. The example shows the best alignments for two reads obtained with Bowtie, and the one selected (light grey square). The 17 nt seed is outlined and the longest alignment maintaining the number of observed mismatches within the seed is highlighted in red. Note that for Read2, the chosen alignment is not the one that contains the least total number of mismatches.

forcedly correspond to Candidate_1 in sample 2. This means that candidates from different samples need to be grouped using the chromosomal coordinates and the sequence of the predicted mature microRNA. Precursor sequences from different samples correspond to the same predicted microRNA if they overlap in more than 80% and if their pairwise mature microRNA sequence identity is higher than 80%. The microRNA candidates detected in over 55% of all samples are selected. The module also calculates the consensus sequences for mature and precursor microRNAs using clustalw (16). Finally, for differentially expressed microRNAs, a new process that uses TargetSpy (17) can be launched to calculate putative target sites. We have chosen TargetSpy for three reasons: (i) it can easily detect targets for newly predicted microRNAs, (ii) it does not rely on cross-species comparison (conservation) and (iii) it has been shown to work very well in a broad range of different species, which is particularly important. The functional analysis (18) of the target genes is currently being redesigned and will be available soon.

Standalone version

The miRanalyzer standalone version needs a local file-based data base that holds miRBase, mRNA and RFam libraries, genome sequences, Bowtie indexes, prediction models and all other user defined libraries. miRanalyzer relies on three programs or packages that must be installed before: the Open Source Machine Learning Software Weka (19), the Vienna RNA Package (20) and Bowtie (10). There are several advantages when using the standalone version: parameter values can be changed, customized libraries can be added and not publically available assemblies can be included.

Conversion of fastq to RC format (read count)

A perl script to convert fastq format into read-count format is provided on the miRanalyzer web page. The script allows now (i) to process colour sequence data and SCARF format, (ii) to select a maximum read length and (iii) to force all reported sequences to be present in all of the analysed samples.

OUTLOOK

We present an updated version of miRanalyzer including many new features. The introduced improvements like full colour space support, differential expression analysis and plant genomes will allow to address the needs of a wider range of users. In the future, we intend to improve the accuracy of the alignments, by adding the possibility of using sequence quality values and the efficiency of the standalone version by introducing the possibility of parallelization. Other topics will be the incorporation of gene expression values in order to infer microRNA regulatory networks.

DATA AND METHODS

Sequence data

The detection of known microRNAs is currently based on the latest miRBase version 16 (21) and will be updated as new miRBase versions are released. In this version, we also distinguish between mature, maturestar, unobserved maturestar and hairpin microRNA sequences. Unobserved maturestar are those maturestar microRNAs that are theoretically possible but that are not present in miRBase (as a consequence of not having been experimentally observed). An updated list of all available species and assemblies can be found on the miRanalyzer tutorial page. The data were obtained from UCSC Genome Browser (13) with the exception of: silkworm (*Bombyx mori*) genome version 2 from SilkDB (22), *Arabidopsis thaliania* from the Arabidopsis Information Resource—TAIR (23), maize (*Zea mays*) version 1, vine grape (*Vitis vinifera*) version 12x and rice (*Oryza sativa*) version 6.1 from plantGDB (24) and *Medicago truncatula* version 3.0 from *Medicago truncatula* genome project (25). NCBI reference sequences (mRNA and 3′UTR) were used whenever available (26), and the mapping to other RNA families was carried out with the RFam database version 10 (27).

Training datasets

The size of the training set has been notably increased compared to the first version. Table 2 shows the data used for training the prediction models.

Detection of read clusters

Once aligned to the genome, the reads that may belong to the same candidate mature microRNA are clustered. Each cluster is defined by two coordinates: (i) the start and end coordinates, that is, the start and end positions of the most

Table 2. Data sets used to train the prediction models

Species	Tissues/Conditions	No. of microRNAs	References	GEO references
Animal				
H. sapiens	16	10 321	(29,30)	GSE19812, GSE20384, GSE21279, GSE20892
M. musculus	9	6201	(30,31)	GSE20384, GSE19473
D. melanogaster	9	587	(32)	GSE12462,GSE24314, GSE24608, SE24542, GSE24540
C. elegans	12	2091	(33,34)	GSE18634, GSE13339
D. rerio	2	695	(35,36)	GSE21503, GSE22068
B. mori	3	46	(37)	GSE17965
Plant				
A. thaliana	4	295	(38,39)	GSE20448, GSE16971
O. sativa	9	1302	(40,41)	GSE23217, GSE20748
Z. mays	3	193	(42)	GSE17339
V. vinifera	1	28	(43)	GSE18406

upstream and downstream reads, respectively, and (ii) the start position of the most expressed read; this latter is named the 'cluster anchor' and is used to decide if a read belongs to a cluster or not. Clusters are constructed following these two steps:

- Reads are sorted according to their read count from highest to lowest (most to least expressed).
- The most expressed read defines/opens the first cluster and the following reads are added to an existing cluster if the read (i) is located on the same strand as the pre-existing cluster and (ii) falls totally inside an already opened cluster, or its start coordinate lies within a window defined by (cluster anchor −2 bp, cluster anchor + 5 bp).

If a read is not found to belong to any pre-existing cluster, a new cluster is opened being this read the most expressed one (defining the cluster anchor) in the new cluster.

Generating precursor candidates

From the clustering process described above, genome positions of the putative mature microRNAs are obtained; however, the candidate precursor sequence on which many machine-learning features are based needs yet to be defined. Since we do not know neither the arm in which the mature microRNA is located nor the length of the precursor sequence, several candidate precursor sequences with different lengths for both, the hypothetical location in 3′ and 5′, are generated, and the one with best structural criteria and binding energy is kept.

The chromosome coordinates are given as:

For the 5′-arm (+ strand):

$$start^{precursor} = start^{mature} - Step * i$$
$$end^{precursor} = start^{mature} + 2 * len^{mature} - 1 + len^{loop} + Step * i$$

For the 3′-arm (+ strand):

$$start^{precursor} = start^{mature} - (len^{loop} + len^{mature} + Step * i)$$
$$end^{precursor} = end^{mature} + Step * i$$

Table 3. Features used for the Random forest prediction models

Feature	Used for kingdom
Number of bindings in read cluster sequence	Animal
Normalized mean free energy of precursor sequence	Plant and Animal
Number of bindings in precursor	Animal
Length of read cluster	Plant and Animal
The corresponding putative maturestar sequence is also present (binary value 0, 1)	Plant and Animal
Number of bindings in read cluster divided by the read cluster length	Plant
Number of reads in read cluster	Plant and Animal
Mean free energy of precursor sequence	Plant and Animal
Degree of bulb asymmetry in precursor	Animal
The number of bulbs in precursor secondary structure	Plant

We set the mature microRNA length to 20 nt and the loop to 15 nt for both, animal and plants. The parameters *Step* and *i* are set to 5 and 8 for animals and to 7 and 10 for plants. This is because plant precursor sequences can be longer than animal ones. Applying these values to the formulas above, we get that the minimum and maximum lengths are 65 and 135 for animals and 69 and 195 for plants.

Prediction models

For all candidate microRNAs generated in the step above, we calculate several features based on both, the secondary structure and expression derived properties [see Ref. (9) for a more detailed description]. In a first step, we discard a candidate if (i) its read cluster overlaps with the loop by more than 5 bp in the 5′-arm (on the 3′-arm no overlap is allowed), (ii) it has no hairpin, (iii) it has less than 19 bindings to the putative precursor sequence and (iv) it has less than 11 bindings to the region occupied by the read cluster (putative mature microRNA sequence). For the remaining candidates, the features described in Table 3 are calculated. These features have been selected out of a large pool of possible features applying the CfsSubsetEval algorithm in Weka. Finally, the training of five Random

forest models (28) for both animals and plants is performed.

ACKNOWLEDGEMENTS

We want to thank all miRanalyzer users for their valuable feedback helping us to improve the tool.

FUNDING

The Ministry of Innovation and Science of the Spanish Government (BIO2010-20219 to M.H.); the Junta de Andalucia (P07FQM3163 to M.H.); the 'Juan de la Cierva' fellowship (to M.H.); the Department of Industry, Tourism and Trade of the Government of the Autonomous Community of the Basque Country (Etortek Research Programs 2009/20011 to A.M.A.); from the Innovation Technology Department of the Bizkaia County (to A.M.A.). Funding for open access charge: Department of Industry, Tourism and Trade of the Government of the Autonomous Community of the Basque Country (Etortek Research Programs 2009/2011 to A.M.A.).

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Bushati,N. and Cohen,S.M. (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol.*, **23**, 175–205.
3. Li,L., Xu,J., Yang,D., Tan,X. and Wang,H. (2010) Computational approaches for microRNA studies: a review. *Mamm. Genome*, **21**, 1–12.
4. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
5. Pantano,L., Estivill,X. and Marti,E. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.
6. Zhu,E., Zhao,F., Xu,G., Hou,H., Zhou,L., Li,X., Sun,Z. and Wu,J. (2010) mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.*, **38**, W392–W397.
7. Huang,P.J., Liu,Y.C., Lee,C.C., Lin,W.C., Gan,R.R., Lyu,P.C. and Tang,P. (2010) DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.*, **38**, W385–W391.
8. 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.
9. 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**, W68–W76.
10. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
11. Millar,A.A. and Waterhouse,P.M. (2005) Plant and animal microRNAs: similarities and differences. *Funct. Integr. Genomics*, **5**, 129–135.
12. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.
13. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
14. Hackenberg,M., Barturen,G. and Oliver,J.L. (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res.*, **39**, D75–D79.
15. Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,B. and Bartel,D.P. (2002) MicroRNAs in plants. *Genes Dev.*, **16**, 1616–1626.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. Sturm,M., Hackenberg,M., Langenberger,D. and Frishman,D. (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*, **11**, 292.
18. Hackenberg,M. and Matthiesen,R. (2008) Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics*, **24**, 1386–1393.
19. Frank,E., Hall,M., Trigg,L., Holmes,G. and Witten,I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.
20. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*, **125**, 167–188.
21. 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.
22. Duan,J., Li,R., Cheng,D., Fan,W., Zha,X., Cheng,T., Wu,Y., Wang,J., Mita,K., Xiang,Z. *et al.* (2009) SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. *Nucleic Acids Res.*, **38**, D453–D456.
23. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
24. Duvick,J., Fu,A., Muppirala,U., Sabharwal,M., Wilkerson,M.D., Lawrence,C.J., Lushbough,C. and Brendel,V. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
25. Young,N.D., Cannon,S.B., Sato,S., Kim,D., Cook,D.R., Town,C.D., Roe,B.A. and Tabata,S. (2005) Sequencing the genespaces of Medicago truncatula and Lotus japonicus. *Plant Physiol.*, **137**, 1174–1181.
26. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
27. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–140.
28. Breiman,L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
29. Vaz,C., Ahmad,H.M., Sharma,P., Gupta,R., Kumar,L., Kulshreshtha,R. and Bhattacharya,A. (2010) Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. *BMC Genomics*, **11**, 288.
30. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.
31. Su,R.W., Lei,W., Liu,J.L., Zhang,Z.R., Jia,B., Feng,X.H., Ren,G., Hu,S.J. and Yang,Z.M. (2010) The integrative analysis of microRNA and mRNA expression in mouse uterus under delayed implantation and activation. *PLoS ONE*, **5**, e15513.
32. Ghildiyal,M., Xu,J., Seitz,H., Weng,Z. and Zamore,P.D. (2010) Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. *RNA*, **16**, 43–56.

33. de Lencastre,A., Pincus,Z., Zhou,K., Kato,M., Lee,S.S. and Slack,F.J. (2010) MicroRNAs both promote and antagonize longevity in C. elegans. *Curr. Biol.*, **20**, 2159–2168.
34. Kato,M., de Lencastre,A., Pincus,Z. and Slack,F.J. (2009) Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol.*, **10**, R54.
35. Cifuentes,D., Xue,H., Taylor,D.W., Patnode,H., Mishima,Y., Cheloufi,S., Ma,E., Mane,S., Hannon,G.J., Lawson,N.D. *et al.* (2010) A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science*, **328**, 1694–1698.
36. Shin,C., Nam,J.W., Farh,K.K., Chiang,H.R., Shkumatava,A. and Bartel,D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell*, **38**, 789–802.
37. Liu,S., Li,D., Li,Q., Zhao,P., Xiang,Z. and Xia,Q. (2010) MicroRNAs of Bombyx mori identified by Solexa sequencing. *BMC Genomics*, **11**, 148.
38. Li,Y., Zhang,Q., Zhang,J., Wu,L., Qi,Y. and Zhou,J.M. (2010) Identification of microRNAs involved in pathogen-associated molecular pattern-triggered plant innate immunity. *Plant Physiol.*, **152**, 2222–2231.
39. Moldovan,D., Spriggs,A., Yang,J., Pogson,B.J., Dennis,E.S. and Wilson,I.W. (2010) Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in Arabidopsis. *J. Exp. Bot.*, **61**, 165–177.
40. Li,T., Li,H., Zhang,Y.X. and Liu,J.Y. (2010) Identification and analysis of seven H2O2-responsive miRNAs and 32 new miRNAs in the seedlings of rice (Oryza sativa L. ssp. indica). *Nucleic Acids Res.*, **39**, 2821–2833.
41. Wu,L., Zhou,H., Zhang,Q., Zhang,J., Ni,F., Liu,C. and Qi,Y. (2010) DNA methylation mediated by a microRNA pathway. *Mol. Cell*, **38**, 465–475.
42. Wei,F., Stein,J.C., Liang,C., Zhang,J., Fulton,R.S., Baucom,R.S., De Paoli,E., Zhou,S., Yang,L., Han,Y. *et al.* (2009) Detailed analysis of a contiguous 22-Mb region of the maize genome. *PLoS Genet.*, **5**, e1000728.
43. Pantaleo,V., Szittya,G., Moxon,S., Miozzi,L., Moulton,V., Dalmay,T. and Burgyan,J. (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis. *Plant J.*, **62**, 960–976.