Published online 6 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W323–W328***
doi:10.1093/nar/gkp313

GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists

Alexey V. Antonov[1,]*, Sabine Dietmann[1], Philip Wong[1], Dominik Lutter[1] and Hans W. Mewes[1,2]

[1]Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg and [2]Department of Genome-Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Received January 30, 2009; Revised April 17, 2009; Accepted April 18, 2009

ABSTRACT

GeneSet2miRNA is the first web-based tool which is able to identify whether or not a gene list has a signature of miRNA-regulatory activity. As input, GeneSet2miRNA accepts a list of genes. As output, a list of miRNA-regulatory models is provided. A miRNA-regulatory model is a group of miRNAs (single, pair, triplet or quadruplet) that is predicted to regulate a significant subset of genes from the submitted list. GeneSet2miRNA provides a user friendly dialog-driven web page submission available for several model organisms. GeneSet2miRNA is freely available at http://mips.helmholtz-muenchen.de/proj/gene2mir/.

INTRODUCTION

MicroRNAs (miRNAs) are small non-coding RNAs that recognize and bind to (partially) complementary sites often in the 3′-untranslated regions of target genes in the cell and regulate protein production of the target transcript. Different combinations of miRNAs are expressed in different cell types (1,2). It has been demonstrated that miRNAs have roles in almost all biological processes in the cell by regulating gene activity. A number of miRNAs have also been linked to cell diseases (3,4).

High-throughput genomics technologies become a standard routine in many experimental laboratories. Sets of genes are delivered on a regular basis, whose measured states, such as mRNA or protein expression levels, gene-methylation status, gene copy number variations, loss of heterozygosity or homozygous deletions, are different between the explored cell states. It is logical to further expect that genes found at differential states can bear a signature of regulatory activity from miRNAs. To our knowledge there is no computational tool available, which can be exploited for this task.

A common step for analyses of gene list is an inference of biological processes that are statistically overrepresented among derived genes. A number of tools employing available gene-functional annotations (5,6) as well as pathway databases (7,8) have been developed (9–17). The advantages and limitations of most of these tools are reviewed in (8,18).

We have developed GeneSet2miRNA, a web tool which can identify significant subsets of genes from the given gene list which are the targets of a single or several miRNAs. As the number of experimentally validated gene targets for the most miRNAs is relatively small in comparison to the actual (expected) number, GeneSet2miRNA is using predicted targets (19). GeneSet2miRNA provides a user friendly dialog-driven submission web page and supports most available gene identifiers. GeneSet2miRNA supports automatic analyses for several model organisms: *Homo sapiens, Mus musculus* and *Rattus norvegicus*.

MATERIALS AND METHODS

miRNA cooperative model

Each miRNA f is associated with a set of genes S_f which were predicted to be the targets of f. To account for cooperative behavior of miRNAs we consider not only single miRNA but pairs, triplets and quadruplets of miRNAs. Each such pair, triplet or quadruplet, as well as each single miRNA, is referred further as a miRNA-regulatory model of degree 1 (single miRNA model), 2 (pair), 3 (triplet) and 4 (quadruplet). Each regulatory model r (f_1, f_2, ...) is associated with a set of genes S_r, where S_r is the set of genes that are regulated by all miRNAs from the model r ($f_1, f_2, \ldots$). Formally, S_r is computed as the intersection of sets (S_{f1}, $S_{f2}, \ldots$): $S_r = (S_{f1}$ **AND** S_{f2} **AND**).

There are relatively few experimentally validated miRNA—gene interactions in comparison to expected numbers. For example, the current release of the

*To whom correspondence should be addressed. Tel: +49 89 3187 2788; Fax: +49 89 3187 3585; Email: a.antonov@helmholtz-muenchen.de miRecords database (19) includes only 1135 records of validated miRNA-target interactions between 301 miRNAs and 902 target genes in seven animal species. For this reason, GeneSet2miRNA uses the *Predicted Targets* component of the miRecords database (19). The *Predicted Targets* component of miRecords stores predicted miRNA targets produced by 11 established miRNA target prediction programs.

The sets of predicted targets vary significantly between different programs. We consider a simple scoring system to arrange predictions by confidence levels: the more tools predict the interaction, the better the confidence. We use two thresholds to define for each miRNA f two models: soft and strict. In the first case, a gene is added to the set S_f if the interaction is predicted by at least four programs, and, in the second case the interaction should be predicted by at least five programs. It is clear that the strict model is a subset of the soft model.

GeneSet2miRNA

GeneSet2miRNA accepts a query list of genes (referred to as set A) and for each regulatory model r the number a_r genes in set A from the set S_r is counted. In the next step the null hypothesis H_0 (the set A is independent of set S_r) is tested using the number a_r, the size of set A, the size of set S_r and the total number of genes in the whole genome. Hypergeometric tests (accounted for multiple testing using Monte Carlo simulation procedure) is employed to identify miRNA models which are significantly intersected with the set A (18).

Consideration of all possible complex regulatory models (miRNA triplets and miRNA quadruplets) is computationally infeasible due to combinatorial complexity. For this reason a search algorithm is used based on greedy heuristics (9). Greedy heuristics does not guarantee the optimal solution is found in every case, but significantly reduce the computational complexity.

To adjust p-values for multiple testing GeneSet2miRNA employs the Monte-Carlo simulation approach. A set of genes of size A (equal to the size of the input list) are randomly sampled N times from the set B (the set of all genes from the genome). Each time the top enriched miRNA-regulatory models of degree 1 (2, 3 and 4) is identified based on hypergeometric test and the best p-value from such test is added to the corresponding distribution. After repeating this procedure N times we get four (for each degree a distribution) distributions of size N of the best p-values for the models inferred from a random gene list of size A. To estimate the corrected p-value the hypergeometric p-value for each regulatory model r (p_r) inferred from the input gene list is compared to the p-values (p_d) from the corresponding distribution (taking into account the degree of the model). Let us denote k to be the number of times p_r is equal to or inferior ($p_r \geq p_d$) than the p-values from the distribution. The estimate of the corrected p-value is given by formula $p_{corrected} = (k + 1)/N$. This corresponds exactly to the definition of an experiment–wise Westfall and Young p-value (9,12,20–22). For degree 1 (single miRNA model) N is fixed to 10 000 and thus the estimate of the lowest possible p-value is limited to 0.0001. For higher degrees (2, 3, 4) the N is fixed to 1000 and the estimate of the lowest possible p-value is limited to 0.001. More details on the searching algorithm and p-value adjustment can be found in (9,10,23). We need also to point out that p-values are estimated assuming that genes from the input list were sampled from the whole genome.

Automatically supported annotations and gene Ids

As input GeneSet2miRNA accepts several types of gene or protein identifiers. For example, for the human genome GeneSet2miRNA supports identifiers from 'Entrez Gene' (24), 'UniProt/Swiss-Prot', 'Gene Symbol' (24,25), 'UniGene' (24), 'Ensembl' (26), 'RefSeq Protein ID', 'RefSeq Transcript ID' (27) and 'Affymetrix probe codes' (28). Additionally a mixture of several identifier types is possible. In the first step user supplied gene Ids are mapped to 'Entrez Gene' identifiers. For this purpose files from NCBI and Affymetrix web sites are used. Detailed information on data sources used by GeneSet2miRNA is in Table 1.

The user gets full information on the mapping of the supplied gene ids. We would like to point out that protein and gene identifiers can be highly ambiguous (29) with multiple synonymous variants. For this reason the quality of the retrieved annotation can be different for different types of identifiers. To escape multiple mapping issues we recommend submitting 'Entrez Gene' identifies to GeneSet2miRNA.

GeneSet2miRNA application to experimental data

Here we present several examples of analyses of real data by GeneSet2miRNA to illustrate the potential utility the user can get from it.

Example 1. Proof of the principle

One of the concerns related to the current version of the tool might be that GeneSet2miRNA employs predicted targets for miRNAs. The next examples clearly demonstrate that this is not a limitation. To prove this we used several gene lists reported by recently published studies that used microarray analysis to reveal genes whose expression is affected by the presence of excess amount of different miRNAs. For example, in (30) the HepG2 liver cancer cells and A549 lung cancer cells were transected with synthetic miRNAs corresponding to let-7. Total RNA were isolated from the cells 72 h after transfection and hybridized to Affymetrix U133 arrays. The study (31) focused on two miRNAs: miR-1 and miR-124. miR-1 (miR-124, miR-373) RNA duplexes were transfected into HeLa cells, and mRNA was purified and profiled on microarrays. The number of reported differentially expressed genes in all cases is presented in Table 2.

We used five gene lists presented in Table 2 as an input for our tool. In all five cases the top enriched single miRNA model corresponds to the transfected miRNA. For example, in the third case (96 repressed genes in HeLa by miR-1 transfection) GeneSet2miRNA reports both soft and strict hsa-miR-1 models (hsa-miR-1.4 and hsa-miR-1.5) as top enriched ones. Note that the number

Table 1. Types of gene identifiers recognized by GeneSet2miRNA and data sources used for Id mapping

Type of Ids	File used
'Gene Symbol', 'Ensembl', 'LocusTag'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz
'RefSeq Protein ID', 'RefSeq Transcript ID'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2refseq.gz
'UniProt/Swiss-Prot'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_refseq_uniprotkb_collab.gz
'UniGene'	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2unigene
'Affymetrix probe codes'	http://www.affymetrix.com/ Annotation files

Table 2. Summary of five gene lists used to validate GeneSet2miRNA performance

No.	Article	Cell types	Transfected with miRNA	Number of reported genes	Source of gene list
1	(30)	HepG2 liver cancer cells	let-7	1334 (698 repressed and 636 up-regulated)	Supplementary Table S1
2	(30)	A549 lung cancer cells	let-7	629 (244 repressed and 385 up-regulated)	Supplementary Table S2
3	(31)	HeLa cells	miR-1	96 repressed genes	Supplementary Table S1
4	(31)	HeLa cells	miR-124	174 repressed genes	Supplementary Table S2
5	(31)	HeLa cells	miR-373	65 repressed genes	Supplementary Table S4

Table 3. Single enriched miRNA models reported by GeneSet2miRNA in the set of genes which respond to the treatment of the HeLa cells with the miR-1

No.	*p*-value, adjusted for multiple testing (Monte Carlo)	Hypergeometric distribution, *p*-value	SET A targets	SET A size	SET B targets	SET B size	Model
1	0.0001	1.8*e*–47	37	80	189	15 360	hsa-miR-1.5
2	0.0001	8.4*e*–58	52	80	474	15 360	hsa-miR-1.4
3	0.0001	1.6*e*–40	43	80	552	15 360	hsa-miR-206.4
4	0.0001	4*e*–19	15	80	75	15 360	hsa-miR-613.5
5	0.0001	1.3*e*–25	25	80	234	15 360	hsa-miR-206.5
6	0.0001	1.9*e*–25	28	80	355	15 360	hsa-miR-613.4
7	0.0001	5.4*e*–08	9	80	137	15 360	hsa-miR-183.5
8	0.0002	2.1*e*–07	12	80	347	15 360	hsa-miR-183.4

after the dot indicates the threshold used to generate the set S_f (the set of regulated genes). For a gene to be included into the set S_f the threshold specifies the minimal number of programs which predicts gene—hsa-miR-1 interactions. According to the miRecords database, 80 out of 96 repressed genes (Table 3) were predicted to be a target of at least one miRNA (the Column 'SET A size'). The column 'SET B size' specifies the total number of genes in the genome that are predicted to be regulated by at least one miRNA according to miRecords. Thirty-seven genes from the set A are predicted to be regulated by hsa-miR-1 by at least five programs (column 'SET A targets', model hsa-miR-1.5) and 52 genes are predicted to be regulated by hsa-miR-1 by at least four programs (column 'SET A targets', model hsa-miR-1.4). The column 'SET B targets' specifies the number of genes from the whole genome regulated by corresponding miRNA model. The *p*-values of both soft and strict hsa-miR-1 models estimated by Monte Carlo simulation are below 0.0001.

There are also several models significantly enriched related to different miRNAs (hsa-miR-206, hsa-miR-613, hsa-miR-183). Both, hsa-miR-206 and hsa-miR-613 are predicted to regulate very similar to hsa-miR-1 sets of the target genes (according to miRecords database). It is known that hsa-miR-206, hsa-miR-613 and hsa-miR-1 target the ACATTCCA octamer and it was shown that hsa-miR-206 and hsa-miR-1 were strongly expressed in different tissues, like, skeletal muscle and tongue (32). The situation with hsa-miR-183 is less clear as the targets sets of hsa-miR-183 and hsa-miR-1 are not as similar as in previous cases.

Links to the full results for all five-gene lists presented in Table 2 are available at our web site (http://mips.helmholtz-muenchen.de/proj/gene2mir/example/main.html). These examples demonstrate the practical proof that GeneSet2miRNA is able to reveal correctly the signature of miRNA activity in the gene lists. In this example, we knew in advance that the reported gene list must be overrepresented with genes regulated by particular miRNA. Though GeneSet2miRNA uses theoretically predicted miRNA–gene interactions, in each case it has successfully revealed significant traces of activity of the transfected miRNA in the experimentally derived gene list.

Table 4. Enriched miRNA models (pairs) reported by GeneSet2miRNA in the set of about 170 genes that reside in 174 homozygous deletions regions detected in a panel of 76 melanoma cell lines

	p-value, adjusted for multiple testing (Monte Carlo)	Hypergeometric distribution, *p*-value	SET A statistics	SET A size	SET B statistics	SET B size	Model
1	0.001	1.8*e*–08	10	140	89	15 329	[(hsa-miR-19a.4) AND (hsa-miR-520c-3p.4)]
2	0.001	8.6*e*–08	17	140	388	15 329	[(hsa-miR-520d-3p.4) AND (hsa-miR-520a-3p.4)]
3	0.001	9.2*e*–08	17	140	390	15 329	[(hsa-miR-302a.4) AND (hsa-miR-520d-3p.4)]
4	0.001	2*e*–07	17	140	413	15 329	[(hsa-miR-302a.4) AND (hsa-miR-520a-3p.4)]
5	0.001	2.1*e*–07	17	140	416	15 329	[(hsa-miR-520e.4) AND (hsa-miR-520a-3p.4)]
6	0.001	3.4*e*–07	16	140	379	15 329	[(hsa-miR-520e.4) AND (hsa-miR-520d-3p.4)]

Validation of statistical treatment

In previous examples we demonstrated that GeneSet2miRNA is capable of inferring correct models from a gene list known to be overrepresented by genes regulated by a given miRNA. Here we present a vice-versa validation, i.e. we demonstrate that GeneSet2miRNA is capable of recognizing a random gene list. In (30) Affymetrix U133 arrays were used for gene expression analyses. We used probes from this array (about 22 000 probes in total) to generate random gene lists. We generated 20 different random lists of size 100. We analyzed 20 generated lists using GeneSet2miRNA. Only in one case GeneSet2miRNA reported the *p*-value of the inferred model to be better than 0.05 (0.036), three times it was better then 0.1, four times it was better then 0.2. These results correspond to the definition of *p*-value: a probability to get the same quality model for a random list. In other words, if statistical treatment is correct then in *N* submission of different random lists only in one case on average the *p*-value of inferred models is expected to be better than $1/N$. We repeated the same test using random gene lists of size 25 and 1000. In both cases the results were similar to the previous one. Therefore, by these examples we demonstrated that GeneSet2miRNA provides correct estimates of *p*-values for the inferred models.

Example 2. Beyond the experimental scope

The choice of this example is aimed to demonstrate that GeneSet2miRNA can provide insight into the data that is beyond the scope of currently available experimental technologies. It is widely accepted that miRNAs play an important role in almost all biological process. However, the nature of such biological processes, like loss of heterozygosity (LOH) and copy number changes in a cancer cells, for example, make it almost impossible to detect experimentally the associated miRNA activity. The time scale of these processes and their stochastic nature makes experimental setup difficult. The only way to explore experimentally these biological phenomena, now, is to conduct analyses over many cell lines to detect chromosome regions which are on average more frequently subjected to variations. The output of such studies can be converted to genes that are located in the identified chromosome regions. As we show in the next example, GeneSet2miRNA is able to provide statistically validated hypotheses concerning which miRNAs can be related to these biological processes.

In a recently published study (33), Illumina 317K whole-genome single-nucleotide polymorphism arrays were used to define a comprehensive allelotype of melanoma based on loss of heterozygosity (LOH) and copy number changes in a panel of 76 melanoma cell lines. A total 174 homozygous deletions (HDs) were detected. Among those HDs, 52 HDs seemed to target a single locus, 87(50%) targeted more than one gene, and 35 of the HDs did not encompass the coding region of an annotated gene (human genome build hg17). Table 4 in the article lists the chromosomal regions showing HDs in one or more cell lines and about 170 genes that reside in those regions. We used these genes as an input for our tool.

According to the miRecord database, 140 of these genes were predicted to be a target of at least one miRNA by at least four prediction tools. Table 4 reports miRNA models (pairs) enriched in the considered gene list. As one can see, GeneSet2miRNA identified the traces of activity related to several miRNAs. Thus, GeneSet2miRNA provided statistically significant arguments that some miRNAs might be involved in the considered biological phenomena providing the biologist with novel possible experimental targets.

Interestingly, the hsa-miR-302s are found more abundantly in pluripotent ES-cells than in differentiate proliferating cells and have been shown to reprogram cancer cells to a more ES-like pluripotent state (34). hsa-miR-373 have been identified to be highly expressed in retinoblastoma (35) and are potential oncogenes involved in testicular germ cell tumors (36). hsa-miR-19a has been shown to be involved in myeloma (37).

CONCLUSION

Automatic functional profiling has become the ‘de facto’ approach for the secondary analysis of high throughput data (18). A number of tools employing available gene functional annotations have been developed. GeneSet2miRNA is a first web tool that employs profiling methodology to identify the signature of miRNA activity in a gene list. Similar ideas were proposed recently and implemented using the R software package (38). GeneSet2miRNA provides technical support to the user that corresponds to the best currently available standards in the field. It has a simple submission web page that covers several model organisms as well as the most popular gene/protein identifiers. These features make GeneSet2miRNA an attractive practical tool for biologists interpreting new experimental data.

FUNDING

Funding for open access charge: Helmholtz Zentrum Munich, German Research Center for Environmental Health.

Conflict of interest statement. None declared.

REFERENCES

1. Boehm,M. and Slack,F. (2005) A developmental timing microRNA and its target regulate life span in C. elegans. *Science*, **310**, 1954–1957.
2. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
3. Takamizawa,J., Konishi,H., Yanagisawa,K., Tomida,S., Osada,H., Endoh,H., Harano,T., Yatabe,Y., Nagino,M., Nimura,Y. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.*, **64**, 3753–3756.
4. He,L., Thomson,J.M., Hemann,M.T., Hernando-Monge,E., Mu,D., Goodson,S., Powers,S., Cordon-Cardo,C., Lowe,S.W., Hannon,G.J. *et al.* (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828–833.
5. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
6. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
7. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
8. Huang,d.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
9. Antonov,A.V. and Mewes,H.W. (2006) Complex functionality of gene groups identified from high-throughput data. *J. Mol. Biol.*, **363**, 289–296.
10. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.
11. Antonov,A.V., Dietmann,S. and Mewes,H.W. (2008) KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome Biol.*, **9**, R179.
12. Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
13. Goffard,N. and Weiller,G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res.*, **35**, W176–W181.
14. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
15. Khatri,P., Voichita,C., Kattan,K., Ansari,N., Khatri,A., Georgescu,C., Tarca,A.L. and Draghici,S. (2007) Onto-tools: new additions and improvements in 2006. *Nucleic Acids Res.*, **35**, W206–W211.
16. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler – a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
17. Reimand,J., Tooming,L., Peterson,H., Adler,P. and Vilo,J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36**, W452–W459.
18. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, **21**, 3587–3595.
19. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
20. Westfall,P.N. and Young,S.S.*Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York.
21. Antonov,A.V., Dietmann,S., Wong,P. and Mewes,H.W. (2009) TICL – a web tool for network-based interpretation of compound lists inferred by high-throughput metabolomics. *FEBS J.*, **276**, 2084–2094.
22. Antonov,A.V., Dietmann,S., Wong,P., Igor,R. and Mewes,H.W. (2009) PLIPS, an automatically collected database of protein lists reported by proteomics studies. *J. Proteome. Res.*, **8**, 1193–1197.
23. Antonov,A.V. and Mewes,H.W. (2008) Complex phylogenetic profiling reveals fundamental genotype-phenotype associations. *Comput. Biol. Chem.*, **32**, 412–416.
24. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2006) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **34**, D173–D180.
25. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
26. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al.* (2006) Ensembl 2006. *Nucleic Acids Res.*, **34**, D556–D561.
27. 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.
28. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
29. Draghici,S., Sellamuthu,S. and Khatri,P. (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.
30. Johnson,C.D., Esquela-Kerscher,A., Stefani,G., Byrom,M., Kelnar,K., Ovcharenko,D., Wilson,M., Wang,X., Shelton,J., Shingara,J. *et al.* (2007) The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer Res.*, **67**, 7713–7722.
31. Lim,L.P., Lau,N.C., Garrett-Engele,P., Grimson,A., Schelter,J.M., Castle,J., Bartel,D.P., Linsley,P.S. and Johnson,J.M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.
32. Clop,A., Marcq,F., Takeda,H., Pirottin,D., Tordoir,X., Bibe,B., Bouix,J., Caiment,F., Elsen,J.M., Eychenne,F. *et al.* (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.*, **38**, 813–818.
33. Stark,M. and Hayward,N. (2007) Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res.*, **67**, 2632–2642.
34. Lin,S.L., Chang,D.C., Chang-Lin,S., Lin,C.H., Wu,D.T., Chen,D.T. and Ying,S.Y. (2008) Mir-302 reprograms human skin cancer cells into a pluripotent ES-cell-like state. *RNA*, **14**, 2115–2124.
35. Zhao,J.J., Yang,J., Lin,J., Yao,N., Zhu,Y., Zheng,J., Xu,J., Cheng,J.Q., Lin,J.Y. and Ma,X. (2009) Identification of miRNAs associated with tumorigenesis of retinoblastoma by miRNA microarray analysis. *Childs Nerv. Syst.*, **25**, 13–20.

36. Voorhoeve,P.M., le,S.C., Schrier,M., Gillis,A.J., Stoop,H., Nagel,R., Liu,Y.P., van,D.J., Drost,J., Griekspoor,A. *et al.* (2007) A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Adv. Exp. Med. Biol.*, **604**, 17–46.
37. Pichiorri,F., Suh,S.S., Ladetto,M., Kuehl,M., Palumbo,T., Drandi,D., Taccioli,C., Zanesi,N., Alder,H., Hagan,J.P. *et al.* (2008) MicroRNAs regulate critical genes associated with multiple myeloma pathogenesis. *Proc. Natl Acad. Sci. USA*, **105**, 12885–12890.
38. Wu,X. and Watson,M. (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, **25**, 832–833.