BIOINFORMATICS DISCOVERY NOTE

Vol. 28 no. 8 2012, pages 1058–1061
doi:10.1093/bioinformatics/bts114

Genome analysis

Advance Access publication March 8, 2012

Defining and providing robust controls for microRNA prediction

William Ritchie[1,*,†], Dadi Gao[1,†] and John E. J. Rasko[1,2]

[1]Gene and Stem Cell Therapy Program, Centenary Institute, University of Sydney, Sydney, New South Wales and [2]Cell and Molecular Therapies, Royal Prince Alfred Hospital, Camperdown, 2050, Australia

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: microRNAs are short non-coding RNAs that regulate gene expression by inhibiting target mRNA genes. Next-generation sequencing combined with bioinformatics analyses provide an opportunity to predict numerous novel miRNAs. The efficiency of these predictions relies on the set of positive and negative controls used. We demonstrate that commonly used positive and negative controls may be unreliable and provide a rational methodology with which to replace them.

Contact: w.ritchie@centenary.org.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 4, 2011; revised on February 7, 2012; accepted on March 3, 2012

1 INTRODUCTION

microRNAs (miRNAs) are short, ~22 nucleotide long RNAs that reduce gene expression, usually by binding to the 3′ untranslated region of mRNAs. They are involved in many biological processes including cell proliferation and differentiation, development and disease (Alvarez-Garcia and Miska, 2005). In mammals, the primary miRNA transcript is processed by Drosha into a precursor (pre-miRNA) of ~70 nt with a distinctive stem-loop structure. This precursor is exported to the cytoplasm and further processed into a mature miRNA by Dicer (Ying, *et al*., 2006).

A plethora of algorithms have been created to predict novel miRNAs using the primary sequence and predicted folding structure of pre-miRNAs (Ritchie, *et al.*, 2007; Xue, *et al*., 2005). More recently, predictive algorithms have included next-generation sequencing (NGS) data to discover novel miRNAs (Friedlander, *et al*., 2008). These algorithms are crucial for discovering tissue- or species-specific miRNAs that are not present in widely-used databases. To evaluate these predictive tools and assist the creation of novel algorithms, it is essential to define a high-confidence set of positive and negative controls. These will not only be used to benchmark algorithms but also to identify distinguishing features of miRNAs.

We suggest that the most commonly used reference set of positive controls, miRBase (Kozomara and Griffiths-Jones, 2010), requires further refinement to create a high-confidence set appropriate for use as positive controls. In order to improve the dynamic range of the predictions, we introduce a novel method to create a set of robust negative controls. We provide both sets as Supplementary Material.

2 METHODS

Murine miRBase v17 entries were separated into either a ‘high-confidence’ or ‘low-confidence’ dataset (Supplementary Material 1) depending on the amount of published data that examined their function. The high-confidence list was created by automatically retrieving highly cited miRNAs from the GeneCards website (http://www.genecards.org/). Each miRNA for which GeneCards displayed at least three citations that did not include the words ‘sequencing’, ‘library’ or ‘miRBase’ in the title were considered high-confidence. Such studies were discarded because they are unlikely to focus on the functional aspect of individual miRNAs. Subsequent curation of this list guaranteed that each of the 139 miRNAs had been experimentally reported as exhibiting *bone fide* activity as a miRNA either through direct target interaction or a change in target expression subsequent to a change in microRNA expression. The low-confidence list was generated by randomly selecting 139 miRNAs from miRBase that were not in the high-confidence list and that were added in the database in the past year.

Our negative control data consisted of transcripts for which there was no evidence of processing by Dicer. Processing of RNAs can be detected in small RNA sequencing data by searching for a high number of aggregated reads in a specific sub-region of a larger contig. Sequences generated from miRNAs, for example, will cover the pre-miRNA transcript but will aggregate at much higher levels at the mature miRNA coordinates conditional on sufficient depth and correct library preparation. We searched for contigs of at least 70 nt with a coverage of at least five reads that contained such aggregated reads in any of the 42 murine experiments downloaded from the Gene Expression Omnibus (Barrett and Edgar, 2006) (Supplementary Material 2). We first aligned reads from these experiments using Bowtie 0.12.7 on a masked version of the mm9 genome. This produced a total of 257 million mapped reads across all samples with a minimum of 2 million mappable reads for SRR042460 (neutrophils) and a maximum of 11.4 million for SRX018984 (progenitors). We then assembled reads from these experiments that overlapped by at least 10 nt into contigs based on their genomic coordinates using BioSamTools (http://search.cpan.org/~lds/Bio-SamTools/lib/Bio/DB/Sam.pm). We then searched within these contigs for a sudden increase in coverage of at least 5-fold across at least 10 nt. This sudden increase is indicative of small RNA processing or degradation from longer transcripts. Contigs for which we never found aggregated reads in any of the 42 experiments were classified as ‘Expressed but never processed’ (ENPs). We found 12 329 ENPs none of which overlap miRBase coordinates. In every case where a pre-miRNA contig from miRBase was detected in one of the 42 experiments, we systematically found the aggregated reads corresponding to a mature miRNA.

To independently assess the quality of the positive control groups, we analyzed 42 murine RNAseq NGS experiments (Supplementary Material 2) to identify miRNAs that exhibited variation from their reference miRBase entry. These non-reference nucleotides (NRNs) result in the expression of isomirs that are widespread and can vary in proportion in different samples (Morin, *et al*., 2008). Isomirs can arise as a consequence of RNA editing,

[*]To whom correspondence should be addressed.

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

a change in Dicer cleavage position, or the addition of bases at the ends of mature miRNAs. We examined the first and last 10 nt of mature miRNAs from miRBase as well as 5 nt adjacent to each end for the presence of NRNs. Each of the 42 experiments were mapped to the mouse genome mm9 using Bowtie 0.12.7 (Langmead, *et al*., 2009). The frequency of each NRN was calculated with the Bio::DB::Sam scripts (http://search.cpan.org/∼lds/Bio-SamTools/lib/Bio/DB/Sam.pm) for each of the 42 experiments and averaged across all the tissues where the pre-miRNA is expressed. This normalization is important because the high confidence microRNAs were expressed more widely and at much higher levels than the low-confidence dataset (data not shown).

To evaluate the efficiency of our controls, we assessed the effect of a Dicer knockout on the expression of microRNAs predicted with our controls (controls 1) versus commonly used controls (controls 2). Controls 1 consisted of our set of ‘high-confidence’ miRBase entries as positive controls and our ENP dataset as negative controls. Controls 2 consisted of all murine miRBase (v17) sequences as positive controls and randomly selected exonic sequences as negative controls. Both control sets were then filtered for predicted structure with at least 15 bp in the predicted stem and a terminal loop. A support vector machine (SVM) from the svm-light implementation (http://svmlight.joachims.org/) was trained on both controls 1 and 2 using the inductive classification method (Christmann, *et al*., 2002) with default parameters to produce classifiers SVM1 and SVM2, respectively. The training features for both SVM1 and SVM2 consisted of 36 commonly used miRNA structural features (Supplementary Material 3). SVM1 and SVM2 were then tested on putative pre-miRNA transcripts. These were defined by the three following criteria:

- size >70 nt;
- expressed (>50 Reads per Kilobase of Mappable Reads) in either of two NGS Dicer knockout experiments performed in the Cortex and Cerebellum (GSE21090) or in Fibroblasts (GSE22760) available at http://www.ncbi.nlm.nih.gov/sra;
- predicted structure with at least 15 bp in the predicted stem and a terminal loop.

We then compared the expression of these putative pre-miRNA transcripts before and after Dicer knockout using a Bayesian method for comparing digital counts (Audic and Claverie, 1997). We defined for each of the two knockout experiments the following measurements:

- True positives (TPs): number of transcripts predicted to have a pre-miRNA structure that show a significant drop in expression in the knockout sample ($P < 0.01$).
- False positives (FPs): number of transcripts predicted to have a pre-miRNA structure that show no significant drop in expression in the knockout sample ($P > 0.1$).
- True negatives (TNs): number of transcripts predicted to not have a pre-miRNA structure that show no significant drop in expression in the knockout sample ($P > 0.1$).
- False negatives (FNs): number of transcripts predicted to not have a pre-miRNA structure that show a significant drop in expression in the knockout sample ($P < 0.01$).

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

It is worth noting that our measure of TN and FN assume that all pre-miRNA like structures that show decreased expression after Dicer knockout are miRNAs. This approximation may affect the measure of accuracy but does not affect precision.

3 RESULTS

3.1 Positive controls

miRBase (Kozomara and Griffiths-Jones, 2010) is the most widely used repository of miRNA sequences. It has set the standard for defining novel miRNAs and has introduced a nomenclature used by the majority of researchers. The number of sequences it contains has doubled over the past 2 years in part because of the advent of NGS. Although the evidence required for depositing novel miRNA sequences is rigorous, no proof of function is required. Recent studies have shown that miRNAs are not the only molecule in their size category to regulate genes. Numerous endogenous small RNAs and ectopically expressed RNAs of similar size as miRNAs regulate transcription initiation and alternative splicing (Aartsma-Rus and van Ommen, 2007; Taft, *et al*., 2010). Researchers must therefore be cautious in assigning miRNA functionality to small transcripts derived from sequencing experiments.

Most published miRNA prediction algorithms use sequences obtained from miRBase as a positive control to define miRNA-specific characteristics and to test their performance (Ng and Mishra, 2007; Ng Kwang Loong and Mishra, 2007). To investigate whether all miRBase sequences are suitable for this role, we divided them into a ‘high-confidence’ dataset for which there exists published evidence of their function and a ‘Low-confidence’ dataset for which there is little or no evidence. We then tested these groups for the presence of NRN. NRNs are commonly reported for miRNAs and occur with higher frequency near both extremities of the mature miRNA where Dicer cleaves the pre-miRNA (Berezikov, *et al*., 2011; Lee, *et al*., 2010). Because miRBase entries are not assessed for NRNs, we used this miRNA-specific property to independently measure possible differences between high-confidence miRBase entries and low-confidence entries (Fig. 1). The high-confidence dataset contained many NRNs around both ends of the mature sequence, which is in agreement with the observation that numerous miRNAs are expressed as isomirs. We could not detect NRNs in the seed region of these miRNAs, consistent with the crucial role of this region for target site recognition. The same analysis on the low-confidence dataset gave strikingly different results. Almost no NRNs were expressed in or adjacent to the low-confidence mature coordinates. Because the frequency of NRN’s is averaged across all tissues where the pre-miRNA is expressed, the discrepancy in the two datasets is not due to differences in expression level or tissue specificity. This result indicated that low-confidence RNAs from miRBase are either processed with much more fidelity than the high-confidence miRNAs, are processed through a different mechanism or are not likely to be *bona fide*.

To further investigate the differences between high-confidence and low-confidence sequences, we compared their predicted secondary structures using RNAfold (Hofacker, 2009) and focused on pre-miRNA features such as minimal free energy, triplet structure, loops and bulges. Of the 36 commonly used structural features that define pre-miRNAs, we found that 16 were significantly different between the two groups (Supplementary Material 3). Among these, was the size of the external loop known to be crucial for miRNA activity (Liu, *et al*., 2008).

The NRN analysis and structural comparison do not prove that the low-confidence sequences are not functional miRNAs. It is, however, clear that these two sets can be distinguished based on the

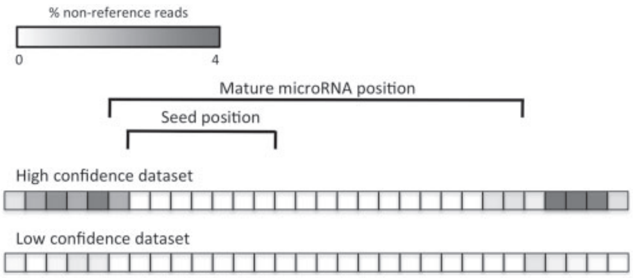

Fig. 1. Percentage of reads containing NRNs at specific positions surrounding mature miRNAs averaged across all sequences in the high-confidence dataset (top) and low-confidence dataset (bottom) of miRBase sequences.

processing of pre-miRNAs and on the hairpin structure and should therefore at least be considered separately.

3.2 Negative controls

The predicted folding of pre-miRNAs allows researchers to loosely approximate the 3D structure of RNA that is recognized by the enzyme Dicer. For this criterion to be used efficiently, it is important to define a set of negative controls that are not recognized by Dicer. The latter should consist of a set of transcripts that are expressed in the same cellular compartment as Dicer but are not recognized by it. Instead of this, most algorithms use random genomic sequences or exonic sequences (Ritchie, *et al*., 2007; Xue, *et al*., 2005). These sequences are poor negative controls because there is no evidence that, if transcribed as small RNAs, these sequences would not be recognized by Dicer and processed into mature miRNAs. To improve upon these controls, we searched for small RNAs (>70 nt) that were expressed but never processed in 42 tissue types. We discovered 12 329 such transcripts we termed ENPs. Because these are expressed at detectable levels in tissues where Dicer is active but are never processed it is likely that certain features of these sequences are preventing them from being recognized. None of the ENP sequences overlap with miRBase entries and none of the pre-miRNAs from miRBase were expressed without an aggregation of reads at the mature coordinates. It is therefore unlikely that adding more tissues to our analysis would have changed our classification of ENPs. When compared with the commonly used exonic sequences, we discovered that ENPs had distinct structural features such as a lower minimal free energy and smaller bulges (Supplementary Material 5). These differences may have a major impact on how miRNA detection algorithms define their structural criteria and emphasize the importance of using an appropriate negative dataset.

3.3 Validation of the control sets

To evaluate the efficiency of our controls, we compared the effect of a Dicer knockout on miRNAs predicted using our novel controls or using commonly used controls. TP predictions should be expressed at lower levels after the Dicer knockout whereas FPs should mostly remain unchanged. We downloaded sequencing data from two Dicer knockout experiments in the Cortex and Hippocampus and in Fibroblasts. We then compared the expression level of predicted mature miRNAs before and after knockout (Supplementary Material 6) using a well established protocol for comparing digital expression counts between two conditions (Audic and Claverie, 1997). The results of this analysis, summarized in Table 1, show that both sets of controls give similar levels of accuracy but that the level of precision, which represents the number of true predictions among all the structures predicted to be miRNA hairpins is considerably higher when applying our set of controls to the same prediction algorithm. This analysis means that our controls produce less FPs, which is crucial in studies where predicted miRNAs will be further analyzed in the wet lab.

Table 1. Accuracy and precision of microRNA hairpin predictions using our novel controls (novel) versus commonly used controls (common) measured in two knockout experiments (fibroblast and brain).

	Dicer KO Fibroblast		Dicer KO Brain	
	Novel	Common	Novel	Common
Accuracy (%)	72.0	69.3	74.0	73.5
Precision (%)	60.0	53.3	68.9	60.0

4 DISCUSSION

The recent discovery of numerous small functional RNAs through NGS experiments has revitalized the field of miRNA gene prediction. The quality of these predictions relies heavily on the positive and negative datasets used to create and test them. We demonstrate that commonly used datasets are heterogeneous and may contain RNAs from different families or miRNAs with distinct functions. The methods we describe in this article use NGS data to characterize RNAs based on their processing patterns and functional evidence taken from the literature. Our aim in providing these rigorous positive and negative controls was to improve the predictive accuracy of existing algorithms, leading to their refinement. Although this study was performed using murine samples and microRNA sequences, the methods described here are applicable to other mammalian species. Experimental confirmation of any predicted miRNA remains as an important standard in the field.

Funding: Australian National Health and Medical Research Council (Training Fellowship 571156), Cancer Council NSW (Project Grant 1006260), the Rebecca L. Cooper Foundation and the Cell and Gene Trust (Cure The Future).

Conflict of Interest: none declared.

REFERENCES

Aartsma-Rus,A. and van Ommen,G.-J.B. (2007) Antisense-mediated exon skipping: a versatile tool with therapeutic and research applications, *RNA*, **13**, 1609–1624.

Alvarez-Garcia,I. and Miska,E.A. (2005) MicroRNA functions in animal development and human disease. *Development*, **132**, 4653–4662.

Audic,S. and Claverie,J.M. (1997) The significance of digital gene expression profiles. *Genome Res*., **7**, 986–995.

Barrett,T. and Edgar,R. (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Meth. Enzymol*, **411**, 352–369.

Berezikov,E., *et al.* (2011) Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res*., **21**, 203–215.

Christmann,A. *et al.* (2002) Classification based on the support vector machine, regression depth, and discriminlant analysis. In *Proceedings in Computational Statistics (Compstat 2002)*, Berlin, Germany, pp. 225–230.

Friedlander,M.R. *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

Hofacker,I.L. (2009) RNA secondary structure analysis using the Vienna RNA package. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit12 12.

Kozomara,A. and Griffiths-Jones,S. (2010) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

Langmead,B*., et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,L.W. *et al.* (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, **16**, 2170–2180.

Liu,G. *et al.* (2008) Pre-miRNA loop nucleotides control the distinct activities of mir-181a-1 and mir-181c in early T cell development. *PLoS One*, **3**, e3592.

Morin,R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

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.

Ng Kwang Loong,S. and Mishra,S.K. (2007) Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification, *RNA*, **13**, 170–187.

Ritchie,W. *et al*. (2007) RNA stem-loops: to be or not to be cleaved by RNAse III. *RNA*, **13**, 457–462.

Mol. Biol., **17**, 1030–1034.

Taft,R.J. *et al.* (2010) Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.*, **17**, 1030–1034.

Xue,C. *et al.* (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.

Ying,S.Y. *et al.* (2006) The microRNA: overview of the RNA gene that modulates gene functions. *Methods Mol. Biol.*, **342**, 1–18.