A comparison of the Oligomap and TargetScan algorithms for miRNA target analysis

Badreddin Edris
BIOC 218 Final Project (Win-09)
Stanford University

Introduction

MicroRNAs (miRNA) are a class of single-stranded, non-coding RNA molecules that range from 21 to 23 nucleotides in length. These small RNAs regulate a number of cellular and genetic functions by forming stem-loop structures that are complementary to one or more messenger RNA (mRNA) molecules, generally resulting in a down-regulation of gene expression through transcript degradation or destabilization (Bartel, 2004).

Since their introduction into the biological vernacular in 2001, a veritable explosion of miRNA research has swept the scientific community, with over 4,000 articles and nearly 1,000 reviews listed on the U.S. National Library of Medicine's PubMed search engine (as of March, 2009). These articles highlight the expanding nature of miRNA research as knowledge of the role and aberrant expression of these small RNAs in cancer (Lu et al, 2005, Luo and Chegini, 2008, and Iorio et al., 2008), diabetes (Tang et al., 2008), cardiovascular disorders (Divakaran and Mann, 2008), neurological diseases (Singh, 2007), tissue-specific gene expression (Liu et al., 2008), and developmental processes (Sun and Tsao, 2008) continues to grow. For these reasons, it is crucial that the development of computational algorithms to annotate and compare miRNA target profiles keeps pace with the data being generated by experimentalists in order for a more complete understanding of the diverse roles of miRNAs to take shape (Pene et al., 2009).

Here, a newly developed algorithm, Oligomap, is critically compared to one of the original miRNA target prediction programs, TargetScan, to illustrate the improvements that have taken place in the last half decade in the realm of computational methods for miRNA target profiling.

Formation and processing of miRNAs

Initially, miRNAs are transcribed as primary transcripts (pri-miRNAs) that are much longer than mature miRNAs. These pri-miRNAs, which have a 5' cap and a polyadenylated ("poly-A") tail, form a stem-loop structure and their conserved structural regions are recognized and eventually processed by a protein "microprocessor" complex containing Drosha, a nuclease, and DiGeorge syndrome critical region gene 8 (DGCR8), a double-stranded RNA binding protein that is produced by a gene deleted in DiGeorge syndrome, a disorder characterized by cardiovascular or immune birth defects. These interactions take place in the nucleus and result in the removal of the 5' cap and the poly-A tail from the pri-miRNA before the molecule is transported to the cytoplasm (Boyd, 2008) (Figure 1).

Once in the cytoplasm, Dicer (an endonuclease) cleaves the stem-loop structure of the molecule (now referred to as a pre-miRNA) resulting in the formation of two short RNA molecules. One of these molecules, termed the guide strand, is then incorporated by the Argonaute protein (an RNase) into the RNAi-Induced Silencing Complex (RISC). At this point, bases 2-8 of the miRNAs ("miRNA seed region") are able to pair with a complimentary sequence at the 3' untranslated region (3' UTR) of an mRNA molecule in order to inhibit translation, induce mRNA degradation, or otherwise post-transcriptionally regulate gene expression at the mRNA level (Bartel, 2009) (Figure 1).

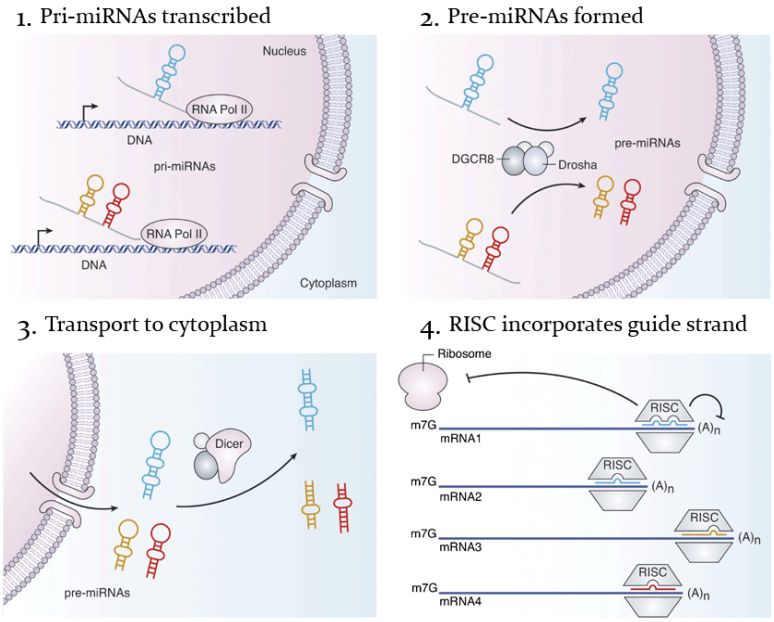

Figure 1. A schematic representation of the miRNA processing pathway, beginning at the (1) pri-miRNA stage and progressing to (2) the pre-miRNA stage, (3) processing in the cytoplasm, and (4) incorporation into the RISC protein complex (adapted from Boyd, 2008).

The miRBase Sequence database

In order to properly annotate and map miRNAs, the miRBase Sequence database was created by the Wellcome Trust Sanger Institute and the University of Manchester in the United Kingdom. First released in 2003, the database (online at *http://microrna.sanger.ac.uk*) provides a centralized and searchable database for all miRNA sequences ("miRBase Sequences"), predicted mRNA targets of miRNAs ("miRBase Targets"), and a registry where novel miRNAs can be submitted by researchers after manuscript acceptance but prior to publication ("miRBase Registry"). In addition to being available online, the database can also be downloaded for use with sequence similarity software and algorithms, such as BLAST and its variants (commonly used, described below) and Oligomap (relatively new, also described below). This makes miRBase a very powerful repository of biological information and an excellent computational tool for researchers interested in small RNAs (Griffiths-Jones et al., 2008 and Landgraf et al., 2007).

miRNAs are annotated in a similar fashion to genes and amino acids (see Figure 1 in the Appendix for a screen shot from the miRBase website). Typically, each miRNA entry in the miRBase sequence database contains:

1. An accession number.
2. An ID number in the following format: [first letter of genus][first two letters of species]-mir-[sequential number]. For example, human (*Homo sapiens*) miRNA 143 would be listed as "has-mir-143". Orthologous miRNAs (from different organisms) are generally given the same number.
3. A sequence for both the pre-miRNA and the mature miRNA.
4. Sources of experimentally verification for each miRNA (with a literature reference).
5. Location on chromosome.
6. Related miRNAs.
7. Additional literature references.

Taken together, the wealth of information contained within miRBase allows for the accurate identification, annotation, and comparison of miRNAs from the growing collection of high-

throughput sequencing data being generated by researchers around the globe who are interested in the biology of small RNA species. As will be discussed below, miRBase is an invaluable resource for miRNA annotation and target analysis.

Initial methods for miRNA annotation

Basic Local Alignment Search Tool (BLAST)

Throughout the majority of the brief history of miRNAs, most analyses have been performed using microarray technology (Yin et al., 2008 and Kong et al., 2009). However, the advent of high-throughput sequencing technology has provided the ability to analyze experimental samples by providing thousands or millions of sequence reads from a single sequencing run (Nelson et al., 2008). Libraries of miRNAs have been generated for high-throughput sequencing, with a general protocol consisting of isolating the total RNA from a sample of interest, doing a size-based separation to obtain smaller RNA fragments, ligating an adaptor to Dicer products, concatamerization of these fragments, and finally ligating this concatamer into a vector for sequencing (Hafner et al., 2008).

To annotate the sequences generated from these data to miRNAs of known function in a given genome, most approaches used a variant of the BLAST algorithm. BLAST utilizes a heuristic approach to search a sequence of interest ("query") against an organismal genome ("target sequence database") by creating staggered "words" that are then organized into a search tree to determine alignments (Altschul et al., 1990).

Limitations of the BLAST algorithm

While the BLAST algorithm provided an excellent framework for finding strong matches between queries and targets, it was initially designed for mapping longer DNA, RNA, or polypeptide sequences to a database. This allowed the algorithm to use much longer "word" sizes (typically, 8-character words are used for DNA searches, depending on the version of the algorithm being used) while still being able to ensure that 1-error hits were returned from a given query (Berninger et al., 2008).

However, when adapting this approach to query a database of small RNAs, such as miRNAs, a number of issues cause the BLAST algorithm to be a less-than-ideal choice. In particular:

1. Sequencing errors in short RNAs (generally ranging from 18-30 nucleotides in length, or 19-21 nucleotides in the case of miRNAs) can greatly reduce the returned alignment scores. In order to remedy this, a much shorter "word" length (of approximately 3 characters long) must be used in order to return an accurate alignment when performing a miRNA search against a genome of interest using a BLAST algorithm.
2. While the reduced "word" length is an adequate solution for returning good alignments for miRNAs, it is not amenable to use with high-throughput sequencing data; when hundreds of thousands of sequence reads are being mapped, combined with a short "word" length and the general complexity of the BLAST algorithm, searching a single large miRNA-derived library can take well over a day to complete (Berninger et al., 2008).

For these reasons, a special-purpose algorithm designed especially to annotate miRNAs and to map short RNA sequences to a target databases was developed by Philipp Berninger and colleagues at the University of Basel and the Swiss Institute for Bioinformatics. This algorithm, termed Oligomap, was published in 2008 and provides an exhaustive miRNA search algorithm that is much lighter on resources and much faster to execute than BLAST and its variants. The advantages of Oligomap for small RNA annotation will be discussed in detail in the sections that follow.

Initial methods for miRNA target analysis and prediction

TargetScan

Originally published in 2003 by Benjamin Lewis and his colleagues at the Massachusetts Institute of Technology and the Whitehead Institute, TargetScan is an algorithm used to predict targets of miRNAs in vertebrates. The algorithm takes into account thermodynamic modeling of RNA-to-RNA duplex interactions and comparative sequence analysis to find miRNA targets that are conserved across more than one genome.

To do this, TargetScan utilizes the following iterative algorithm once a miRNA sequence is inputted:

1. A search of the 3' untranslated region (3' UTR), where miRNAs putatively bind, is utilized to uncover segments where the seed region (bases 2-8) on the miRNA are complementary to the genome of interest.
2. Areas of complementation are deemed “seed matches”.
3. Any seed matches are extended with additional bases to get as much pairing with the miRNA as is possible. This extension takes place in both the 3’ and the 5’ directions and stops when mismatches are found.
    a. Then, uses RNAfold, a secondary RNA structure prediction program (Hofacker et al., 1994; online at *http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi*) to extend the seed matching as far possible in the 3' direction of the miRNA.
    b. The program then assigns a "folding free energy" to each miRNA-target interaction using RNAeval, a free energy evaluator of RNA molecules with a fixed secondary structure (Hofacker et al., 1994; online at *http://www.tbi.univie.ac.at/RNA/RNAeval.html*).
4. Finally, a Z score is assigned to each UTR, using the following equation:

$$Z = \sum_{k=1}^{n} e^{-G_k/T}$$

where n is the number of seed matches in the 3' untranslated region, G_k is the calculated free energy (units: kcal/mol) of the interaction between the miRNA its target for the k^{th} target evaluated in the previous step, and T is a parameter that influences the relative weighting of UTRs as a function of the affinity and abundance of their target sites; T values are assigned by a trial-and-error method involving training sets of miRNAs until reasonable values were determined.

5. This process is iterated for each 3’ untranslated region of each of the organisms whose comparative genomes are being used in the study. The process stops when the Z score

reaches a value higher than a pre-determined cut-off. A rank is assigned to each UTR match determined by the algorithm.

A visual representation of the seed matches (red), seed extension (blue), and their calculated *Z* scores gives a schematic display of the algorithm's function (Figure 2).

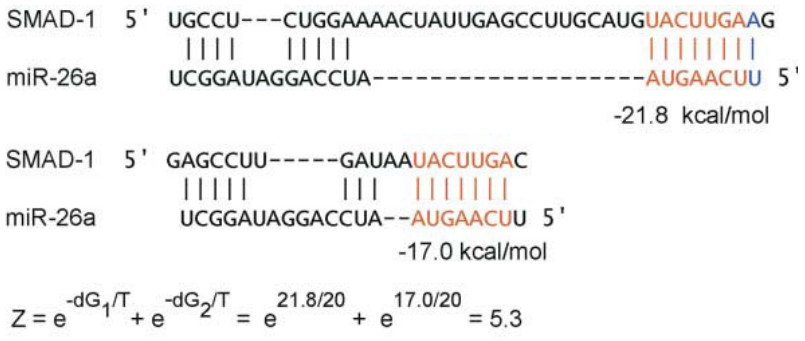

Figure 2. Predicted structures and calculated *Z* scores for miR-26A complex with the gene SMAD-1 (adapted from Lewis et al., 2003).

Limitations of TargetScan

While TargetScan has proven to have a good signal-to-noise ratio in mapping small RNAs to their predicted targets (Lewis et al., 2003), there exist a few inherent limitations in the algorithm.

The first of these is the fact that the *T* values calculated by TargetScan influence the internal weighting of the scores generated by the algorithm without having any thermodynamic basis for their calculation; this is in contrast to the *n* and *G* values that are calculated based on a structure prediction algorithm and free energy calculation. In this way, the TargetScan algorithm produces *Z* scores that are based in part on reproducible calculations (that should not vary from dataset to dataset) and in part on internally adjusted weighting variables. Ideally, an algorithm should use equations or heuristics that minimize variability across different datasets.

Another limitation inherent to the TargetScan program as a whole is the fact that datasets used with the algorithm must be pre-annotated and any non-annotated miRNAs must be input separately when using the TargetScan web interface. While this is certainly doable with low-throughput datasets, the advent of high-throughput sequencing technology

has created the need for an algorithm and miRNA target analysis program that can input a raw dataset and output predicted miRNA targets robustly and with minimal memory and computing power.

As will be described below, the Oligomap algorithm is able to improve upon BLAST-derived sequence annotation programs as well as TargetScan's prediction algorithms to provide an excellent tool for computational and experimental small RNA scientists who seek to use the latest sequencing technology to inform their research.

Advantages of the Oligomap algorithm over TargetScan and BLAST

The Oligomap algorithm

As delineated above, the Oligomap algorithm was designed to rapidly identify small RNA sequences and their targets without having to undertake the complex computational algorithms of BLAST, TargetScan, and their variants. Oligomap is able to identify all sequences that map exactly or contain a single error (which can be a mismatch, an insertion, or a deletion, Figure 3) to a target database.

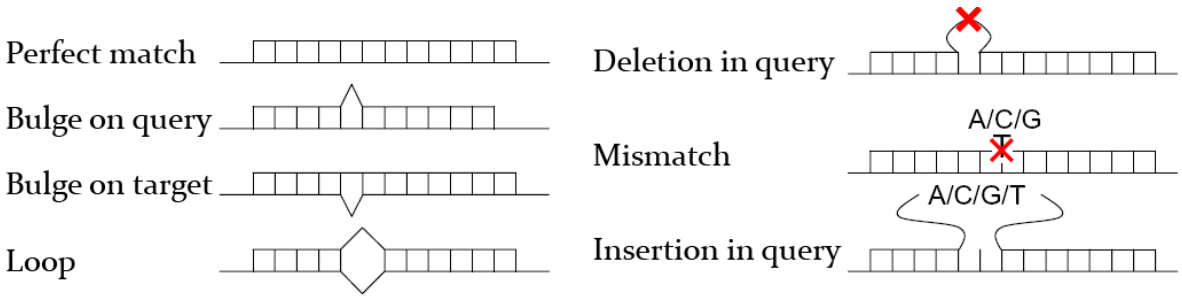

Figure 3. The types of alignments that can be generated by the Oligomap algorithm. In each instance, the query is represented by the top strand and the target is represented by the bottom strand (adapted from Berninger et al., 2008).

The algorithm utilizes a tree-building computational model with the following parameters: each miRNA query is treated as a path that starts at the root of a tree where every node in the tree corresponds to one of the four possible nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T). Therefore, each miRNA can follow a path with four possibilities

at each subsequent nucleotide in its sequence, with an ID code tagged onto the each node of the tree that identifies the small RNA (Figure 4). A suffix of the target sequence ending at a given target in the database is termed a "walker" and includes the following information for each given search:

1. The location in the genome where the alignment began (coordinate).
2. The position where the "walker" was created.
3. The type of match encountered by the walker (perfect match or mismatch).
4. A pointer to a node on the nucleotide tree.

As a "walker" continues along a given sequence and matches it to the nucleotide nodes of the tree, one of two things can happen: a match occurs between a known target sequence and the match is reported, or more than one mismatch occurs and the "walker" corresponding to the given sequence is discarded from the data set (Figure 5).

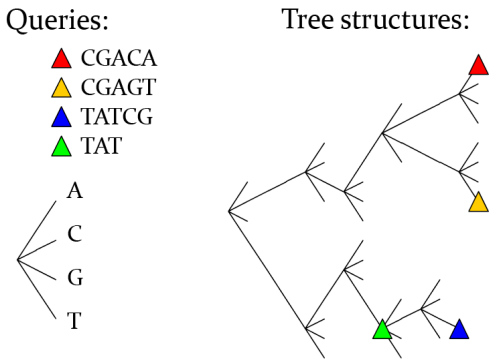

Figure 4. An example of four possible search queries and their resulting locations on the tree structure using a particular order of nucleotide nodes (adapted from Berninger et al., 2008).

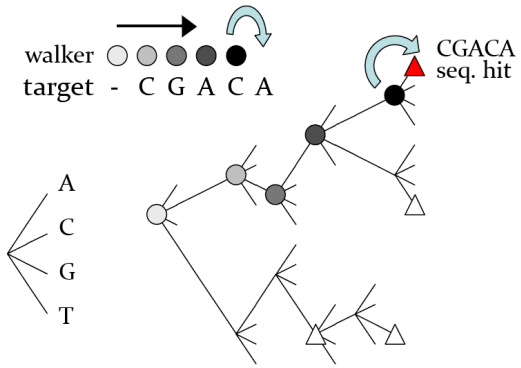

Figure 5. A "walker" scans a given sequence (corresponding to Figure 2) along the nucleotide tree and generates a hit when a query matches a target (adapted from Berninger et al., 2008).

In this way, Oligomap is not only able to provide exhaustive and accurate sequence matches of a given query in a target database of interest but it is also able to accomplish this task while utilizing an extremely simple algorithm that uses very little in the way of

computational and time resources. Also, unlike the TargetScan algorithm, Oligomap does not rely on internally calibrated *T* scores that can change with each dataset used, thus making the results generated using Oligomap's algorithm more robust and reproducible across different experimental datasets.

Additionally, Oligomap is designed to work from high-throughput datasets, whereas TargetScan was initially made before these technologies became pervasive. In this way, not only is Oligomap's target prediction algorithm simpler and light on computational resources, but it is also able to accomplish the task of two algorithms (BLAST for annotation, TargetScan for target prediction) while utilizing high-throughput sequencing technology.

Computational and time resource requirements for Oligomap

As noted earlier, one of the most attractive features of Oligomap is its fast runtime and relatively small memory footprint. With high-throughput data sets generated from small RNAs allowing for up to five million reads per length, the resource management and speed of a given algorithm used to query a database becomes extremely important.

Using a 2.2 GHz AMD Operton computer, Berninger and colleagues were able to show that they could map one million short RNA sequences in just less than three hours and while using approximately 1 GB of memory (Figure 6). Given the ever-increasing speed of commercially available computer processors and desktop and laptop machines with more and more random access memory (RAM) being produced, this shows that Oligomap is relatively light on resources.

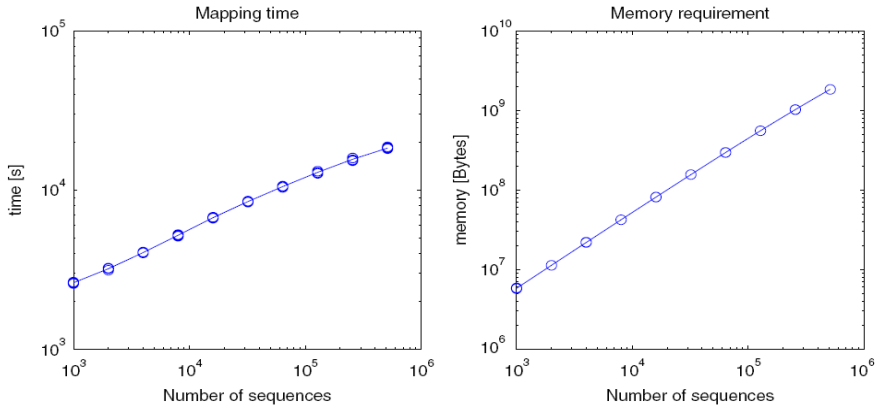

Figure 6. Mapping time and memory usage versus number of sequences (adapted from Berninger et al., 2008).

Additional analyses with Oligomap: automated annotation

In addition to simply finding database matches, Oligomap can also automatically annotate the small RNAs that it finds using its algorithm. To do this, Oligomap utilizes miRBase (or an alternate database for non-miRNA sequences) and proceeds with the following protocol:

1. Perfect and 1-error matches are mapped to the genome of interest (available from the University of California, Santa Cruz Genome Browser: *http://genome.ucsc.edu*).
2. The locus/loci for each of the small RNAs are identified, minimizing the number of mismatch, insertion, or deletion errors.
3. Too distant mappings (which have < 92.5% identity) are filtered out.
4. The remaining small RNAs are mapped to miRBase; perfect and 1-error matches are identified.
5. Too distant mappings (which have < 92.5% identity) are again filtered out.
6. A functional category is assigned to each miRNA based on database annotations.
7. If no in-species matches are found, but an out-of-species match is found, the out-of-species annotation is used for a given small RNA.
8. In instances where small RNAs map equally well to multiple miRNA precursors, the small RNAs are counted toward each of the precursors and is weighted in an inverse fashion to the number of a given precursor.
9. Final counts for each small RNA species in the sample are generated.

Berninger and colleagues have provided the algorithms and programs described above in a software package that can be freely downloaded and used to analyze your own samples. These applications are available at *http://www.mirz.unibas.ch/smiRNAdb*.

Taken together, it is clear that Oligomap provides an excellent algorithm and set of applications to identify, categorize, and predict miRNA targets derived from high-throughput sequencing data. This is because the small RNA-specific nature of the Oligomap algorithm,

when compared to TargetScan and its variants, has the benefit of being much simpler and utilizing much less computational resources to effectively generate useful data from miRNA sequencing experiments. Given this, it would seem to be a compelling option for miRBase and its future iterations to utilize an algorithm such as Oligomap in order to more accurately annotate and predict miRNA targets.

Future biological and computational challenges in the miRNA world

Much is being learned on a daily basis about the biology of small RNAs and their roles in normal and pathological development. However, a number of outstanding challenges lie ahead for computational and experimental biologists, particularly in the realm of miRNA target prediction (Broderson and Voinnet, 2009).

Among these challenges is the limitation of comparative sequence analysis for accurately determining gene targets being regulated by miRNAs. Owing the particularly small nature of miRNAs and the lack of complete knowledge of how and where miRNAs bind mRNA transcripts to regulate them, preferentially conserved sites are often difficult to distinguish from those conserved by chance. Naturally, computational algorithms used for target prediction must set some cutoffs, but this leads to setting computational thresholds that result in some real targets being eliminated without removing all false positives. Therefore, additional factors that determine target specificity of miRNAs need to be determined so that they can be incorporated into more accurate miRNA target prediction algorithms (Didiano and Hobert, 2008).

Another significant challenge facing scientists interested in developing better target-prediction algorithms for miRNAs is how targeting specificity takes place in instances where other regulatory processes are taking place in the cell. Alternative cleavage or polyadenalation of a transcript, which are common phenomena in proliferating cells, can alter the regulatory sites present in an mRNA and, as a result, change miRNA target efficacy (Sandberg et al., 2008). Additionally, cells under stress have altered miRNA-to-mRNA transcript binding profiles (Bhattacharyya et al., 2006).

While these and other issues will certainly pose continuous challenges to computational and experimental biologists in the coming years, Oligomap nonetheless provides an excellent algorithmic framework from which to analyze and annotate miRNA samples. As our functional knowledge of miRNAs grow and the databases and data sets available to scientists continue to increase, Oligomap and its future iterations will prove to be invaluable in creating meaningful information on which our understanding of mechanistic biology can grow and our miRNA target prediction algorithms can improve.

References

- Altschul SF, W Gish, W Miller, EW Myers, DJ Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**(3): 403–410.
- Bartel DP. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**(2): 281-297.
- Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2): 215-233.
- Berninger P, D Gaidatzis, E van Nimwegan, M Zavolan. 2008. Computational analysis of small RNA cloning data. *Methods* **44**(1): 13-21.
- Bhattacharyya SN, R Habermacher, U Martine, EI Closs, w Filipowicz. 2006. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell* **125**: 1111–1124.
- Boyd SD. 2008. Everything you wanted to know about small RNA but were afraid to ask. *Lab Invest* **88**(6): 569-578.
- Brodersen P, O Voinnet. 2009. Revisiting the principles of microRNA target recognition and mode of action. *Nature Rev Mol Cell Biol* **10**(2): 141-148.
- Didiano D, O Hobert. 2008. Molecular architecture of a miRNA-regulated 3' UTR. *RNA* **14**(7): 1297-1317.
- Divakaran V, DL Mann. 2008. The emerging role of microRNAs in cardiac remodeling and heart failure. *Circ Res* **103**(10): 1072-1083.
- Griffiths-Jones S, HK Saini, S van Dongen, AJ Enright. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**(Database Issue): D154-D158.
- Hafner M, P Landgraf, J Ludwig, A Rice, T Ojo, C Lin, D Holoch, C Lim, T Tuschl. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**(1): 3-12.
- Hofacker IL, W Fontana, PF Stadler, S Bonhoeffer, M Tacker, P Schuster. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie* **125**: 167-188.
- Iorio MV, P Casalini, E Tagliabue, S Menard, CM Croce. 2008. MicroRNA profiling as a tool to understand prognosis, therapy response and resistance in breast cancer. *Eur J Cancer* **44**(18): 2753-2759.
- Kong W, JJ Zhao, L He, JQ Cheng. 2009. Strategies for profiling microRNA expression. *J Cell Physiol* **218**(1): 22-25.
- Landgraf P, M Rusu, R Sheridan, A Sewer, N Iovino, A Aravin, S Pfeffer, A Rice, AO Kamphorst, M Landthaler, C Lin, ND Socci, L Hermida, V Fulci, S Chiaretti, R Foà, J Schliwka, U Fuchs, A Novosel, RU Müller, B Schermer, U Bissels, J Inman, Q Phan, M

Chien, DB Weir, R Choksi, G De Vita, D Frezzetti, HI Trompeter, V Hornung, G Teng, G Hartman, M Palkovits, R Di Lauro, P Wernet, G Macino, CE Rogler, JW Nagle, J Ju, FN Papavasiliou, T Benzing, P Lichter, W Tam, MJ Brownstein, A Bosio, A Borkhardt, JJ Russo, C Sander, M Zavolan, T Tuschl. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**(7): 1401-1414.

- Lewis BP, I Shih, MW Jones-Rhoades, DP Bartel, CB Burge. 2003. Prediction of mammalian microRNA targets. *Cell* **115**: 787-798.
- Liu CG, R Spizzo, GA Calin, CM Croce. 2008. Expression profiling of microRNA using oligo DNA arrays. *Methods* **44**(1): 22-30.
- Lu J, G Getz, EA Miska, E Alvarez-Saavedra, J Lamb, D Peck, A Sweet-Cordero, BL Ebert, RH Mak, AA Ferrando, JR Downing, T Jacks, HR Horvitz, TR Golub. 2005. MicroRNA expression profiles classify human cancers. *Nature* **435**(7043): 834-838.
- Luo X, N Chegini. 2008. The expression and potential regulatory function of MicroRNAs in the pathogenesis of leiomyoma. *Semin Reprod Med* **26**(6): 500-514.
- Nelson PT, WX Wang, BR Wilfred, G Tang. 2008. Technical variables in high-throughput miRNA expression profiling: much work remains to be done. *Biochim Biophys Acta* **1779**(11): 758-765.
- Pene F, E Courtine, A Cariou, JP Mira. 2009. Toward theragnostics. *Crit Care Med* **37**(1 Suppl): S50-S58.
- Sandberg R, JR Neilson, A Sarma, PA Sharp, CB Burge. 2008. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* **320**: 1643–1647.
- Singh SK. 2007. miRNAs: from neurogeneration to neurodegeneration. *Pharmacogenomics* **8**(8): 971-978.
- Sun BK, H Tsao. 2008. Small RNAs in development and disease. *J Am Acad Dermatol* **59**(5): 725-737.
- Tang X, G Tang, S Ozcan. 2008. Role of microRNAs in diabetes. *Biochim Biophys Acta* **1779**(11): 697-701.
- Yin JQ, RC Zhao, KV Morris. 2008. Profiling microRNA expression with microarrays. *Trends Biotechnol* **26**(2): 70-76.

Appendix

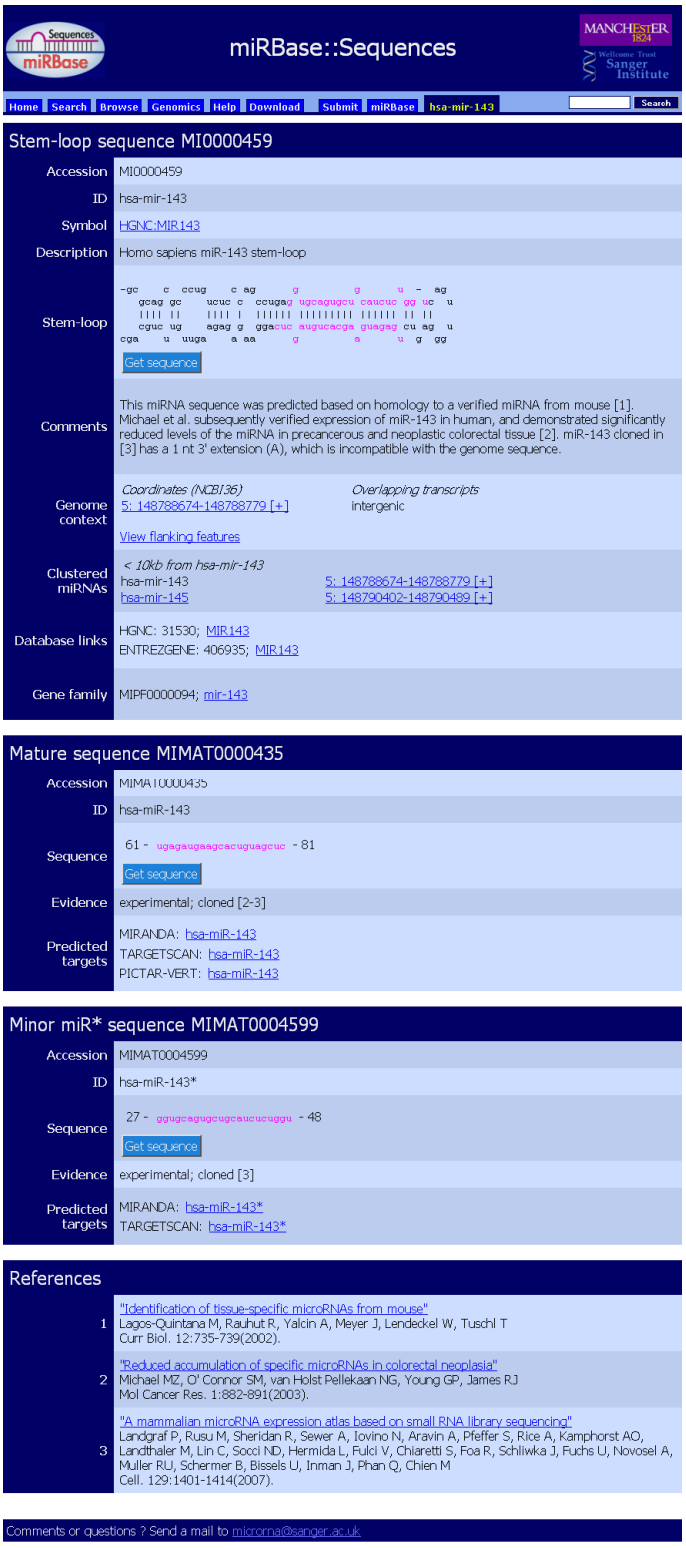

Figure 1. Screen shot of miRBase entry for hsa-mir-143 with accession M10000459 (accessed March 10, 2009).