Methods for microRNA profiling and discovery using massively parallel sequencing

by

Ryan David Morin

B.Sc., Simon Fraser University, 2003

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

The Faculty of Sciences

(Bioinformatics)

The University of British Columbia
October 2007

©Ryan David Morin, 2007

ABSTRACT

MicroRNAs (miRNAs) are emerging as important, albeit poorly characterized, regulators of gene expression. Here, I review the current knowledge of miRNAs in humans, including their biogenesis, modes of action and various methods for studying them. To fully elucidate the various functions of miRNAs in humans, we require a more complete understanding of their numbers and expression changes amongst different cell types. This document includes a description of a new method for surveying the expression of miRNAs that employs the new Illumina sequencing technology. A set of methods is presented that enables identification of sequences belonging to known miRNAs as well as variability in their mature sequences. As well, a novel system for miRNA gene discovery using these data is described. Application of this approach to RNA from human embryonic stem cells (hESCs) obtained before and after differentiation into embryoid bodies (EBs) revealed the sequences and expression levels of 362 known plus 170 novel miRNA genes. Of these, 190 known and 31 novel microRNA sequences exhibited significant expression differences between these two developmental states. Owing to the increased number of sequence reads, these libraries currently represent the deepest miRNA sampling in any human cell type spanning nearly six orders of magnitude of expression. Predicted targets of the differentially expressed miRNAs were ranked to identify those that are likely under cooperative miRNA regulation in either hESCs or EBs. The predicted targets of those miRNAs enriched in either sample shared common features. Included amongst the high-ranked predicted gene targets are those implicated in differentiation, cell cycle control, programmed cell death and transcriptional regulation. Direct validation of these predicted targets or global discovery of miRNA targets should reveal the functions of these sequences in the differentiation of hESCs.

TABLE OF CONTENTS

Abstract ii
Table of Contents iv
List of Tables v
List of Figures vi
Acknowledgements vii
Dedication viii
Co-authorship statement ix
1 An introduction to microRNAs 1
1.1 MicroRNA biogenesis 1
1.2 MicroRNA function 1
1.3 Methods for microRNA target prediction 3
1.4 MicroRNA discovery, detection and profiling 4
1.5 Perspectives and research goals 7
1.6 References 8
2 Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells 11
2.1 List of authors 11
2.1.1 Manuscript Status (footnote) 11
2.2 Introduction 11
2.3 Results 15
2.3.1 Sequencing and annotation of small RNAs 15
2.3.2 Variability in microRNA processing 15
2.3.3 Enzymatic modification of microRNAs 17
2.3.4 MicroRNAs differentially expressed between hESCs and EBs 18
2.3.5 Novel microRNA genes 19
2.3.6 Targets of differentially expressed microRNAs 19
2.4 Discussion 22
2.5 Methods 27
2.5.1 Small RNA library preparation 27
2.5.2 Small RNA genome mapping and quantification 30
2.5.3 Differential expression detection 30
2.5.4 Small RNA annotation 31
2.5.5 Novel miRNA detection 33
2.5.6 Cooperative miRNA target prediction 35
2.5.7 Gene ontology analysis 35
2.6 References 47
3 Concluding remarks 52
3.1 Goals accomplished 52
3.2 Future directions 54
3.3 References 56
Appendices 57
Appendix A: Supplementary Tables 57
Appendix B: Supplementary Methods and accompanying data 70

LIST OF TABLES

Table 2.1: Top 20 miRNAs differentially expressed between the hESC and EB libraries..........36

Table 2.2: Putative novel miRNAs with significant differential expression..........37

Table 2.3: Top 25 statistically over-represented seeds of miRNA sequences found in hESCs or EBs..........38

LIST OF FIGURES

Figure 2.1: Distribution of sequence counts for the different classes of small RNAs..39-40

Figure 2.2: The repertoire of isomiRs and 3' modifications of hsa-mir-191...........42

Figure 2.3: Clustering of over-represented Gene Ontology classes in predicted targets of differential microRNAs...44-45

ACKNOWLEDGEMENTS

I am grateful to my rotation supervisors, Dr. Peter Unrau and Dr. Francis Ouellette, for assisting me in developing my scientific skills and providing guidance throughout the duration of my degree.

I extend my deepest gratitude to my supervisor, Dr. Marco Marra, for providing me with an excellent work environment and thesis project.

This work would not have been possible without the ongoing moral support of my loving wife, Sarah.

DEDICATION

To my parents, grandparents,
wife, Sarah,
and daughter, Natalie

CO-AUTHORSHIP STATEMENT

The manuscript included in this document is a culmination of work conducted by both laboratory scientists and bioinformaticians. Drs. Michael O'Connor and Connie Eaves produced the human embryonic stem cell and embryoid body RNA that was used for small RNA sequencing. Dr. Martin Hirst, Anna-liisa Prabhu, Yongjun Zhao, Helen McDonald and Thomas Zeng prepared small RNA libraries and sequenced the material. Malachi Griffith and Drs. Florian Kuchenbauer and Allen Delaney provided bioinformatics support and helpful suggestions for strategies and improvements to the data analysis and presentation. Ryan Morin and Marco Marra co-conceived the project. Marco Marra provided methodological, conceptual, emotional and financial support. Ryan Morin performed all database and algorithm design, bioinformatic analyses, wrote the manuscript and produced all figures. Rhonda Oshanek edited the document for content and grammar.

1 AN INTRODUCTION TO MICRORNAS

1.1 microRNA biogenesis

MicroRNAs (miRNAs) are short RNA molecules, 19-25 nucleotides (nt) in length that derive from stable fold-back sub-structures of larger transcripts (Cai et al. 2004). MiRNA genes can exist in the introns of spliced mRNAs, either as solitary non-coding transcripts, or within polycistronic transcripts containing multiple miRNA genes (Bartel 2004). In most cases, these primary miRNA transcripts (pri-miRNAs) are cleaved by a complex of Drosha and its cofactor DGCR-8 producing one or more miRNA precursor (pre-miRNA) hairpins with a 2-nt overhang at their 3' ends (Lund et al. 2004). Pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin-5, after which they are further cleaved by Dicer, releasing short RNA duplexes with a 2-nt 3' overhang at both ends (Lund et al. 2004). The recently discovered 'mirtrons', which have been observed in both *Drosophila melanogaster* and *Caenorhabditis elegans* are an exception to this mechanism as they undergo nuclear export and Dicer cleavage without Drosha processing (Ruby et al. 2007). In either case, after cleavage by Dicer, one of the strands of the remaining RNA duplex is thought to be preferentially incorporated into the RNA-induced silencing complex (RISC), leaving an inactive molecule (miRNA*) that is subsequently degraded (O'Toole et al. 2006).

1.2 microRNA function

Of the miRNA/miRNA* duplex, which comprises a short strand from each arm of the pre-miRNA hairpin, only one strand is generally thought to become an active miRNA molecule, hence assembled within RISC (Bartel 2004). The active strand is identified based on the thermodynamic properties at both ends of the duplex, including the identity of the two nucleotides in the 3' overhang (O'Toole et al. 2006). However, for an increasing number of miRNA genes (currently 79) (http://microrna.sanger.ac.uk), mature sequences from both hairpin arms (i.e., both strands of the miRNA/miRNA* duplex), are known to accumulate in approximately equal abundance in the cell and are both thought to be biologically active. The activity of miRNA and miRNA* sequences is often presumed based on cloning frequencies or homology to known miRNAs. However, large-scale sequencing projects are leading to a steady increase in the number of miRNA genes known to produce two functional miRNAs. As the miRNA genes of this type accumulate, a new notation is being implemented that may replace the limited miRNA/miRNA* notation. This system, which is used throughout the remainder of the text, disambiguates miRNAs arising from a single pre-miRNA using their positioning 5' (5p) or 3' (3p) within the pre-miRNA hairpin (Griffiths-Jones et al. 2006).

In addition to the mature miRNA, the RISC protein complex can contain a variety of proteins. A key component of RISC is the member of the Argonaute protein family, of which there are four in human (EIF2C1, EIF2C2, EIF2C3 and EIF2C4) (Tolia et al. 2007). Once assembled within RISC, miRNAs can elicit post-transcriptional repression of target genes. In animals, miRNA-target interactions occur through semi-complementary base pairing to sites usually within the 3'-UTR of the target transcript (Bartel 2004). Strong base pairing at the 5' end of the miRNA, generally termed the seed region, is thought to be a requisite for miRNA-target interaction. If the target site contains a perfect complement to the miRNA seed sequence, relaxed hybridization is tolerated along the remainder of the miRNA/mRNA duplex. If, however, there is a lack of complementarity in the seed, stronger complementarity along the remainder of the miRNA/mRNA duplex is required for a strong interaction (Bartel 2004).

Hybridization of miRNAs (within RISC) to their target mRNAs can result in post-transcriptional regulation by two separate mechanisms: transcript degradation or translational repression. MiRNAs that anneal with extensive complementarity to their target sequences and are accompanied by a RISC containing EIF2C2 are thought to direct the cleavage of target messenger RNAs. This process is analogous to the action of short interfering RNAs (siRNAs) (Liu et al. 2004; Meister et al. 2004). In animals, the more widespread mechanism of miRNA action is that of translational repression. This process is mediated by RISCs containing EIF2C1 and generally involves miRNAs with reduced complementarity to their target sites (Bartel 2004). The binding of multiple RISCs to the same transcript can enhance this effect considerably (Grimson et al. 2007; Doench et al. 2003).

1.3 Methods for microRNA target prediction

That mammalian cells tolerate relaxed complementarity between miRNAs and their targets poses many problems in computational 'target prediction'. Hence, creating accurate target prediction algorithms is an active field of study, with numerous emerging approaches that rely on differing methodologies and assumptions. These algorithms use different methods to assess the relative complementarity between miRNAs and their potential targets (John et al. 2004). Presently, the three leading programs for mammalian miRNA target prediction are PicTar (Krek et al. 2005), MiRanda (John et al. 2004) and TargetScanS (referred to as TargetScan throughout the remainder of this document) (Lewis et al. 2005). To enhance their specificity, all three methods enforce some degree of evolutionary conservation within the region of candidate target sites. These methods also consider pairing within the seed region of the miRNA to be more important than in the remainder of the sequence. Though all three approaches use similar inputs, they result in moderate to poor agreement between their target predictions, with the strongest agreement between the outputs of PicTar and TargetScan (Sethupathy et al. 2006). A common drawback amongst all these programs is that they lack predictions for recently discovered miRNAs, since their database updates do not occur as frequently as updates to miRBase. TargetScan was used in the analyses described herein because it provides the ability to perform 'custom' target predictions with novel miRNA sequences.

1.4 microRNA discovery, detection and profiling

In addition to current deficiencies in target identification, the field of miRNA research also lacks a complete list of the human miRNA genes. Additionally, evaluation of expression levels of the known miRNAs amongst the numerous tissue and cell types is a field still in its infancy (Landgraf et al. 2007). Both of these deficiencies can be attributed to a lack of a robust methodology for profiling the expression level of each miRNA in a sample. The current commercially available high-throughput methodologies rely on primers or probes designed to detect each of the current reference miRNA sequences residing in miRBase, which acts as the central repository for all known miRNAs (Griffiths-Jones 2006). These systems, which can be purchased as individual probes or in a microarray format (Castoldi et al. 2007), allow concurrent profiling of the expression of many miRNAs (Zhao et al. 2006). The recent incorporation of Locked Nucleic Acid (LNA) nucleotide analogs into probes promises a great enhancement in the robustness of these technologies, potentially facilitating discrimination between miRNAs with single nucleotide differences (Vester et al. 2004). MiRNA microarrays offer good reproducibility and facilitate clustering of samples by similar miRNA expression profiles (Davison et al. 2006; Porkka et al. 2007). However, these methodologies are restricted to detecting and profiling of only the known miRNA sequences previously identified by sequencing or homology searches, or querying candidate regions of the genome that have been predicted to produce novel miRNAs (Berezikov et al. 2006a).

An emerging alternative method for miRNA detection and profiling involves generating cDNA libraries from small RNA cellular fractions, followed by deep sequencing using a variety of approaches (Lu et al. 2007). Previously, sequencing-based strategies for miRNA detection have been hindered by the requirements of concatamerization and cloning as well as the expense of capillary DNA sequencing (Cummins et al. 2006; Pfeffer et al. 2005). Nevertheless, small RNA sequencing has several advantages over hybridization-based methodologies. Discovery of novel miRNAs need not rely on querying candidate regions of the genome for expression using microarrays. Rather, novel miRNAs can be discovered by direct observation of their mature sequences, followed by validation of the folding potential of flanking genomic DNA using various approaches (Berezikov et al. 2006a; Cummins et al. 2006). Small RNA sequencing also offers the potential to detect variation in mature miRNA length and enzymatic modification of miRNAs such as RNA editing (Kawahara et al. 2007) and 3' nucleotide additions (Aravin et al. 2005; Landgraf et al. 2007).

In contrast to capillary sequencing, recently developed 'next-generation' sequencing technologies offer the prospect of a vast yet inexpensive increase in throughput, therefore providing a more complete view of the miRNA transcriptome. With the depth of sequencing now possible, we have an opportunity to identify low-abundance miRNAs or those exhibiting modest expression differences between samples, which may not be detected by hybridization-based methods such as microarrays. Next-generation miRNA profiling has already been accomplished in a variety of organisms (Fahlgren et al. 2007; Kasschau et al. 2007; Rajagopalan et al. 2006; Yao et al. 2007; Berezikov et al. 2006b) using, at first, Massively Parallel Signature Sequencing (MPSS) methodology (Nakano et al. 2006) and more recently the Roche/454 platform (Berezikov et al. 2006b; Margulies et al. 2005). The recently released Illumina sequencing platform provides two to three orders of magnitude of greater depth from a single run when compared to the output of current competing technologies (Berezikov et al. 2006b).

1.5 Perspectives and research goals

Three next-generation sequencing technologies have been released in the past few years (Bennett 2004; Margulies et al. 2005; Shendure et al. 2005). Each technology generates sequence reads in varying quantities with different lengths and error profiles. The work outlined in the accompanying manuscript describes a focused view of the miRNA profiling capabilities of the Illumina sequencing method. Many new algorithms and tools are required to facilitate analysis of the large data sets produced by these technologies. A major goal of this project was to produce a set of bioinformatics tools, including a database and associated software, which would allow a skilled researcher to gain biologically relevant and statistically meaningful insights from this type of data. This methodology was first used to profile the miRNA changes involved in differentiation of human embryonic stem cells. A secondary goal of this project was to identify genes potentially regulated by these miRNAs, in hopes of identifying genes of central importance to the pluripotency or self-renewal of these cells. The following chapter discusses progress towards these goals and should provide a foundation for future analyses using this or similar sequencing approaches.

1.6 References

Aravin, A. and Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett* **579:** 5830-40.

Bartel, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116:** 281-97.

Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* **5:** 433-438.

Berezikov, E., Cuppen, E., and Plasterk, R. H. 2006a. Approaches to microRNA discovery. *Nat Genet* **38 Suppl:** S2-7.

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. 2006b. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38:** 1375-7.

Cai, X., Hagedorn, C. H., and Cullen, B. R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna* **10:** 1957-66.

Castoldi, M., Benes, V., Hentze, M. W., and Muckenthaler, M. U. 2007. miChip: A microarray platform for expression profiling of microRNAs based on locked nucleic acid (LNA) oligonucleotide capture probes. *Methods* **43:** 146-152.

Cummins, J. M., He, Y., Leary, R. J., Pagliarini, R., A., D. L.,Jr, Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A. E., Labourier, E., et al. 2006. The colorectal microRNAome. *Proc Natl Acad Sci U S A* **103:** 3687-92.

Davison, T. S., Johnson, C. D., and Andruss, B. F. 2006. Analyzing micro-RNA expression using microarrays. *Methods Enzymol* **411:** 14-34.

Doench, J. G., Petersen, C. P., and Sharp, P. A. 2003. siRNAs can function as miRNAs. *Genes Dev.* **17:** 438-442.

Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., Law, T. F., Grant, S. R., Dangl, J. L., et al. 2007. High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. *PLoS ONE* **2:** e219.

Griffiths-Jones, S. 2006. miRBase: the microRNA sequence database. *Methods Mol Biol* **342:** 129-38.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34:** D140-4.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91-105.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. 2004. Human MicroRNA targets. *PLoS Biol* **2:** e363.

Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., and Carrington, J. C. 2007. Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. *PLoS Biol* **5:** e57.

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. 2007. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315:** 1137-40.

Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., et al. 2005. Combinatorial microRNA target predictions. *Nat Genet* **37:** 495-500.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., et al. 2007. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129:** 1401-14.

Lewis, B. P., Burge, C. B., and Bartel, D. P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J. J., Hammond, S. M., Joshua-Tor, L., and Hannon, G. J. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305:** 1437-1441.

Lu, C., Meyers, B. C., and Green, P. J. 2007. Construction of small RNA cDNA libraries for deep sequencing. *Methods* **43:** 110-117.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95-8.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-80.

Meister, G. and Tuschl, T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431:** 343-349.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S. S., Skogen, J. W., and Meyers, B. C. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res* **34:** D731-5.

O'Toole, A. S., Miller, S., Haines, N., Zink, M. C., and Serra, M. J. 2006. Comprehensive thermodynamic analysis of 3' double-nucleotide overhangs neighboring Watson-Crick terminal base pairs. *Nucleic Acids Res* **34:** 3338-44.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., et al. 2005. Identification of microRNAs of the herpesvirus family. *Nat Methods* **2:** 269-76.

Porkka, K. P., Pfeiffer, M. J., Waltering, K. K., Vessella, R. L., Tammela, T. L., and Visakorpi, T. 2007. MicroRNA expression profiling in prostate cancer. *Cancer Res* **67:** 6130-5.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. 2006. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* **20:** 3407-25.

Ruby, J. G., Jan, C. H., and Bartel, D. P. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448:** 83-86.

Sethupathy, P., Megraw, M., and Hatzigeorgiou, A. G. 2006. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods* **3:** 881-886.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728-1732.

Tolia, N. J. and Joshua-Tor, L. 2007. Slicer and the Argonautes. *Nat Chem Biol.* **1**: 36-43.

Vester, B. and Wengel, J. 2004. LNA (locked nucleic acid): high-affinity targeting of complementary RNA and DNA. *Biochemistry* **43:** 13233-41.

Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J. K., and Sun, Q. 2007. Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). *Genome Biol* **8:** R96.

Zhao, J. J., Hua, Y. J., Sun, D. G., Meng, X. X., Xiao, H. S., and Ma, X. 2006. Genome-wide microRNA profiling in human fetal nervous tissues by oligonucleotide microarray. *Childs Nerv Syst* **22:** 1419-25.

2 Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells

2.1 List of Authors

Ryan D. Morin, Michael D. O'Connor, Malachi Griffith, Florian Kuchenbauer, Allen Delaney, Anna-liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, Connie J. Eaves and Marco A. Marra.

2.2 Introduction

MicroRNAs (miRNAs) are short RNA molecules, 19-25nt in length, that derive from stable fold-back sub-structures of larger transcripts (Cai et al. 2004). In most cases, primary miRNA transcripts (pri-miRNAs) are cleaved by a complex of Drosha and its cofactor DGCR-8, producing one or more pre-miRNA hairpins, each with a 2-nt 3' overhang (Lund et al. 2004). Nuclear export of these precursors is mediated by Exportin 5 (Lund et al. 2004) after which, they are released as short double-stranded RNA duplexes following a second cleavage by Dicer (Lund et al. 2004). Based on the thermodynamic stability of each end of this duplex, one of the strands is thought to be preferentially incorporated into the RISC protein complex, producing a biologically active miRNA and an inactive miRNA* (O'Toole et al. 2006).

Once assembled within RISC, miRNAs can elicit down-regulation of target genes by blocking their translation, inducing Ago-2 mediated degradation or potentially inducing deadenylation (Aravin et al. 2005; Giraldez et al. 2006). In

*A version of this chapter has been submitted to Genome Research to be considered for publication as a research article animals, miRNA-target interactions occur through semi-complementary base pairing, usually within the 3'-UTR of the target transcript. The relaxed complementarity between miRNAs and their target sites poses many problems in computational target prediction; hence, this is an active field of study with numerous emerging approaches that rely on differing methodologies and assumptions. Common target prediction algorithms assess complementarity of miRNAs to their targets (John et al. 2004), many enforcing strong pairing within the seed region of the miRNA (positions 2-8) (Grimson et al. 2007; Lewis et al. 2005). As they were discovered only relatively recently, the study of miRNAs is a young and rapidly changing field in which undiscovered genes remain and the sequence modifications and activity of mature miRNAs are not well understood.

Before the effect of miRNAs on gene regulation can be globally studied, a robust method for profiling the expression level of each miRNA in a sample is required. The current commercially available high-throughput methodologies rely on primers or probes designed to detect each of the current reference miRNA sequences residing in miRBase, which acts as the central repository for known miRNAs (Griffiths-Jones 2006). These systems, which are often available in array form, allow concurrent profiling of many miRNAs (Zhao et al. 2006). These approaches feature good reproducibility and facilitate clustering of samples by similar miRNA expression profiles (Davison et al. 2006; Porkka et al. 2007). However, probe-based methodologies are generally restricted to the detection and profiling of only the known miRNA sequences previously identified by sequencing or homology searches.

Sequencing-based applications for identifying and profiling miRNAs have been hindered by laborious cloning techniques and the expense of capillary DNA sequencing (Cummins et al. 2006; Pfeffer et al. 2005). Nevertheless, direct small RNA sequencing has several advantages over hybridization-based methodologies. Discovery of novel miRNAs need not rely on querying candidate regions of the genome, but rather can be achieved by direct observation and validation of the folding potential of flanking genomic sequence (Berezikov et al. 2006a; Cummins et al. 2006). Direct sequencing also offers the potential to detect variation in mature miRNA length, as well as enzymatic modification of miRNAs such as RNA editing (Kawahara et al. 2007) and 3' nucleotide additions (Aravin et al. 2005; Landgraf et al. 2007). In contrast to capillary sequencing, recently available 'next-generation' sequencing technologies offer inexpensive increases in throughput, thereby providing a more complete view of the miRNA transcriptome. With the added depth of sequencing now possible, an opportunity to identify low-abundance miRNAs or those exhibiting modest expression differences between samples exists, including those that may not be detected by hybridization-based methods. Next-generation miRNA profiling has already been realized in a few organisms (Fahlgren et al. 2007; Kasschau et al. 2007; Rajagopalan et al. 2006; Yao et al. 2007) using the Massively Parallel Signature Sequencing (MPSS) methodology (Nakano et al. 2006) and more recently the Roche/454 platform (Margulies et al. 2005). The recently released Illumina sequencing platform provides approximately two orders of magnitude of greater depth than current competing technologies from a single run (Berezikov et al.

2006b), hence providing a more cost-effective option when compared to other sequencing strategies (http://www.solexa.com).

We sought to use this new technology to extensively profile and identify changes in miRNA expression that occur within a previously characterized model system. Pluripotent human embryonic stem cells (hESCs) can be cultured under non-adherent conditions that induce them to differentiate into cells belonging to all three germ layers and form cell aggregates termed embryoid bodies (EBs) (Itskovitz-Eldor et al. 2000; Bhattacharya et al. 2004). Samples of undifferentiated hESCs and differentiated cells from EBs were chosen for miRNA profiling, first because the pluripotency of ESCs is known to require the presence of miRNAs (Bernstein et al. 2003; Song et al. 2006; Wang et al. 2007) and second because specific changes in miRNA expression are thought to accompany differentiation (Chen et al. 2007). The hESC messenger RNA transcriptome has been extensively studied by ourselves and others (Abeyta et al. 2004; Bhattacharya et al. 2005; Bhattacharya et al. 2004; Boyer et al. 2005; Hirst et al. 2007; Sato et al. 2003) but to date, very little is known about the specific miRNAs that may play roles in their pluripotency or differentiation (Suh et al. 2004).

2.3 Results

2.3.1 Sequencing and annotation of small RNAs

Sequencing of small RNA libraries yielded 6,147,718 and 6,014,187 37 nucleotides (nt) unfiltered sequence reads from hESCs and EBs, respectively. After removal of reads containing ambiguous base calls, 5,261,520 (hESC) and 5,192,421 (EB) unique sequences remained with counts varying between 1 and 38,390. After mapping to the human genome, a total of 766,199 (hESC) and 724,091 (EB) genome-mapped/trimmed small RNA sequences were represented by 4,351,479 and 3,886,865 reads. These sequences were annotated as one of the known classes of small RNA genes or degradation fragments of larger non-coding RNAs (Methods). Sequences derived from 362 distinct miRNA genes were identified. Based on their median expression level, miRNAs were the most abundant class of small RNAs on average, but spanned the entire range of expression, with sequence counts from singletons up to ~120 thousand (Figure 2.1, panel a). The piRNAs, previously thought to exist only in germline cells, were also found at relatively low levels. A total of 9,012 and 4,606 sequence reads corresponded to piRNAs in hESCs and EBs, respectively. The sampling depth provided by the Illumina technology, spanning approximately 5 orders of magnitude, appears to extend the dynamic range of miRNA expression within a cell by two additional orders of magnitude (Berezikov et al. 2006a).

2.3.2 Variability in microRNA processing

In both libraries, miRNAs frequently exhibited variation from their 'reference' sequences, producing multiple mature variants that we hereafter refer to as isomiRs. In many cases the miRNA* sequence and its isomiRs were also observed in our libraries. The existence of isomiRs (Figure 2.2 shows an example) is commonly reported in miRNA cloning studies but generally dismissed with either the sequence matching the miRBase record or the most frequently observed isomiR chosen as a reference sequence (Landgraf et al. 2007; Ruby et al. 2006). It appears that much of the isomiR variability can be explained by variability in either Dicer or Drosha cleavage positions within the pri-miRNA hairpin. Notably, our data show that choosing a different isomiR sequence for measuring miRNA abundance can affect our ability to detect differential miRNA abundance. Based on our analysis, the read count for the most abundant isomiR, rather than the miRBase reference sequence, provides the most robust approach for comparing miRNA expression between libraries (see Supplementary Methods in Appendix B). In 140 cases, this most abundant sequence did not correspond exactly to the current miRBase reference sequence. Most commonly, this was due to length variation at the 3' end of the miRNAs, though isomiRs with variation at either the 5' end or both ends were also observed. This suggests either that the relative abundance of isomiRs may vary across tissues or that the original submission of this miRNA to miRBase was incorrect. The latter appears more likely when we consider that following the most recent update to miRBase, more of the most abundant sequences in our libraries then corresponded to the updated miRBase reference sequences.

2.3.3 Enzymatic modification of microRNAs

Although some reads were detected that may represent the result of pre-miRNA editing by adenosine deaminases (observed as A to G transitions) or cytidine deaminases (producing C to U transitions), examples of these were infrequent and few were significantly above the background level of other apparent sequencing errors. The more prevalent type of modification noted amongst the miRNAs were single-nucleotide 3' extensions, which occurred in multiple isomiRs from nearly all observed miRNA genes (316 genes). These modifications produce an isomiR that matches the genome at every position except the terminal nucleotide. The nucleotide most commonly added was Adenine (1130 distinct examples), followed by Uridine (1008 examples), Cytosine (733 examples) then Guanine (508 examples).

The prevalence of modifications differed amongst the miRNAs, with some miRNAs having more reads representing modified than unmodified forms. End modifications were not limited to the most common isomiR, nor did they show significant differences between the two libraries, suggesting that they may not bear direct significance in hESCs, but rather may reflect a general cellular process of miRNA modification (see Figure 2.2 for an example). For some of the miRNAs showing the highest ratios of modified/unmodified isomiRs, the same modification was previously observed across multiple human and mouse tissues (Landgraf et al. 2007). The study by Landgraf and colleagues employed a different sequencing approach, suggesting that these variations are not inherent to the sequencing strategy used. For example, hsa-miR-326 was found with a terminal Adenine addition in 66% of the corresponding reads in both of our libraries and multiple members of the family hsa-miR-30 showed Uridine additions in more than half of their reads. Both of these specific modifications were noted in the Landgraf study, where the same modifications were also observed in the orthologous mouse miRNAs. The significance of the apparent evolutionary conservation of these modifications is addressed below.

2.3.4 MicroRNAs differentially expressed between hESCs and EBs

We tested for differential expression between the hESC and EB libraries using all sequences, regardless of their annotation; hence, each isomiR was independently tested for differential expression. The total number of differentially expressed sequences corresponding to the miRBase reference sequence (120 total) was less than the total number of miRNAs identified as differentially expressed using the most abundant isomiR for each pre-miRNA arm (190 total from 165 miRNA genes) (Table 2.1 and Supplementary Table 1, Appendix A). All 120 of the miRNA genes identified as differentially expressed, based only on the sequences corresponding perfectly to those in miRBase, were amongst the 165 identified by the most the abundant sequence in our data. Employing the count of the most abundant sequence for the calculation of differential expression also allows identification of the supposed miR* sequences that are differentially expressed between hESCs and EBs but are currently absent from miRBase (25 total). This reinforces the concept that the most abundant, rather than the reference miRNA sequence, is the most useful for identifying differentially expressed miRNAs.

2.3.5 Novel microRNA genes

We sought to identify novel miRNA genes amongst the un-classified sequences in our libraries. After annotation, 40,699 of the small RNA sequences in the hESC library and 40,598 in the EB library (ignoring singletons) remained unclassified because they derived from un-annotated regions of the human genome. As they were of generally low expression and often similar in sequence to known genomic repeats, we assumed many of these were either randomly derived fragments of various RNAs or regulatory siRNAs (Berezikov et al. 2006b; Watanabe et al. 2006). To identify candidate novel miRNAs amongst these, we employed both in-house and publicly available algorithms (Methods). Once good candidates were identified, we compared both their mature sequences and the predicted pre-miRNA sequences to those of all currently known miRNAs to aid in classifying them into families. The total set of novel miRNA candidates, included in Supplementary Table 2 (Appendix A), comprises 129 unique miRNA sequences from up to 170 distinct genes. Compared to other miRNAs in our data, these miRNAs exhibited modest expression (mean count=58), perhaps indicating that most of them may not perform significant functions in these cell types. However, 31 exhibited significant differential expression between the two libraries and are thus potentially biologically important in the cells in which they were present (Table 2.2).

2.3.6 Targets of differentially expressed microRNAs

Strong base pairing between the seed region of a miRNA and the UTR of its target mRNA is important for its activity; hence many target prediction algorithms enforce strong seed complementarity and evolutionary conservation in the complementary region of potential targets (Grimson et al. 2007). As such, the repertoire of predicted targets of miRNAs with identical seeds often overlaps considerably. We sought to determine whether miRNAs sharing identical seeds demonstrated coexpression. By considering positions 2-8 of all non-singleton isomiRs of known miRNAs from each library, we identified 1009 distinct seeds. Many of these represented the non-canonical seeds of isomiRs arising from variation at their 5' end. 114 of these seeds were significantly over-represented in the hESC library while 106 are over-represented in the EB library (Fisher Exact Test, alpha=0.05/1009) (Table 2.3 and Supplementary Table 3 in Appendix A).

As expected, we found that many of the seeds with the largest changes in relative levels between the two libraries corresponded to the differentially expressed miRNAs in Table 2.1. Notably, in some cases, we observed pairs of miRNAs that shared a common seed yet exhibited inverse expression changes. A clear example of this from our data is the inverse abundances of hsa-miR-302a (hESC-enriched) and hsa-miR-327 (EB-enriched), which share the seed sequence AAGUGCU. This behavior may mask the effect of differential expression of some miRNAs in each library. Hence, focusing on the net change in seed levels, rather than distinct miRNAs, may be important in this context. Some miRNAs appear in this table more than once, suggesting that more than one isomiR from a single miRNA gene can contribute to net changes in miRNAs with a given seed during differentiation, and these isomiRs could potentially regulate different sets of transcripts.

The cooperative action of multiple miRNAs can be multiplicative and, in some cases, synergistic (Grimson et al. 2007); hence, transcripts with more predicted target sites for co-expressed miRNAs should be most drastically affected by those miRNAs. In an attempt to highlight potentially significant targets of differentially expressed miRNAs, we identified genes with predicted target sites for multiple hESC-enriched miRNAs or EB-enriched miRNAs using TargetScan. Measures were taken to compensate for UTR length, miRNAs with identical target sites, and a general preponderance of target sites in some genes (Methods). A total of 591 likely cooperative targets of EB-enriched and 461 targets of hESC-enriched miRNAs were identified by this approach. As these genes are likely under redundant post-transcriptional regulation by multiple miRNAs, they could comprise genes of central importance to the maintenance of these cells. Surprisingly, these two gene lists showed a significant overlap of 64 genes (p<0.0001, permutation test), suggesting that some of the miRNA-regulated genes in hESCs may also be regulated in EBs by a different set of miRNAs.

The genes that have been highlighted herein as likely targets of the differentially expressed miRNAs would be expected to be significant to hESC biology. This was supported by an examination of the Gene Ontology 'biological process' classifications that are significantly over-represented amongst these genes. This analysis revealed that many of the genes in both groups have been previously associated with differentiating stem cells and included those involved in differentiation, development and regulation of transcription (Skottman et al.

2005). Interestingly, some of these GO terms were enriched only in the predicted targets of hESC-enriched or EB-enriched miRNAs. For example, genes involved in programmed cell death were enriched amongst the predicted targets of hESC-enriched miRNAs while those involved in cell proliferation were enriched amongst the predicted targets of EB-enriched miRNAs (Figure 2.3).

2.4 Discussion

These data highlight the potential of a new massively parallel sequencing strategy to profile miRNA expression in hESCs and EBs and the changes that occur during differentiation. Between the two libraries, we identified 190 differentially expressed known and 31 novel miRNAs corresponding to 165 and 33 distinct miRNA genes. This is a striking advance in contrast to a recent array-based study of murine ESCs, which revealed a much smaller number of miRNAs differentially expressed between ESCs and EBs (Chen et al. 2007). The latter study identified only 23 candidate miRNAs as either ESC-specific or down-regulated during differentiation, as well as 10 miRNAs that appear to be enriched in EBs. It is encouraging that, of the 27 listed mouse miRNAs with identifiable human orthologs, 14 exhibited expression patterns consistent with the hESC results presented here, whilst many of the others exhibited insignificant changes in expression. Further, of the additional differentially expressed miRNAs reported here (Table 2.1), many were originally discovered in human or murine ESCs (Houbaviy et al. 2003; Suh et al. 2004).

Some of these previous studies generalized these miRNAs as hESC-specific in their expression. However, compared to EBs, our method did not detect miRNAs that were expressed exclusively in hESCs; this is likely a consequence of the increased sampling depth of the method used here. On the other hand, we did find 16 miRNA sequences in EBs that were not observed in hESCs (Supplementary Table 1, Appendix A). This is consistent with the postulate that the number of expressed miRNAs increases during differentiation (Strauss et al. 2006) and further supports the importance of miRNAs during ESC differentiation (Wang et al. 2007). However, because EBs likely represent a more diverse population of cells, the total numbers of expressed miRNAs in any given cell type may not necessarily increase. Thus, the miRNAs highlighted by this method include those previously implicated in pluripotency and differentiation, but the added sampling depth demonstrates that these miRNAs are not uniquely expressed by undifferentiated hESCs. This result may suggest that the processes involved in differentiation that are under miRNA-mediated regulation are dictated by the repertoire and relative miRNA levels, rather than the number of distinct miRNAs being expressed.

With few exceptions (Berezikov et al. 2006b), recent large-scale cloning efforts have provided minimal yields of new miRNA genes. This is may be due to the dominance of the highly expressed miRNAs in small RNA libraries as well as difficulties associated with library normalization (Landgraf et al. 2007; Shivdasani 2006). Here, we present 170 candidates for novel miRNA genes that have passed multiple levels of annotation criteria. This gene list was obtained by combining two separate miRNA classification tools relying on different assumptions and input data. MiPred relies on a Random Forest algorithm to classify miRNAs based on structural properties (Jiang et al. 2007). Our custom classifier, which uses a support vector machine (SVM), assigns a classification score based on a separate set of structural descriptors including data specific to this sequencing method. As a result, this classifier has provided the best classification accuracy of any machine learning-based miRNA classification method (see Methods).

At least 31 of the novel miRNAs identified in this study show evidence of regulated expression during hESC differentiation (Table 2.2), suggesting that they may also have roles in maintaining the pluripotent status of hESCs or their ability to self-renew. In support of this, five of these miRNA sequences were found to share a common seed sequence with at least one of the differentially expressed known miRNAs (Table 2.1). Many of the remaining novel miRNAs with overall low abundance and insignificant expression differences may later prove unimportant in the context of hESC biology. It is encouraging, however, that many of the known miRNAs resided in the same range of expression as these novel miRNAs (Figure 2.1, panel a), suggesting that many miRNAs may exhibit low-level expression in hESCs detectable only by deep sequencing. Like many of the known miRNAs present in these libraries, these low-abundance novel miRNAs may be observed in higher quantities amongst other tissues where they are functionally important.

In addition to novel miRNA genes, this study has aided in identifying diverse population of variants of known miRNAs, collectively termed isomiRs. The functional implications of the widespread 3' modifications and RNA editing are unclear in the context of hESCs as neither process appeared to vary significantly between the two libraries. It is important to note that the diverse 3' nucleotide additions observed here reiterate previous observations of this nature (Landgraf et al. 2007). By comparing some of the most prevalent 3' additions in our libraries to those from diverse human and mouse tissues (Landgraf et al. 2007), it is evident that the nature of the added nucleotide is nonrandom and evolutionarily conserved.

Some of the remaining non-canonical isomiRs, which appear to derive from variability in Dicer and Drosha cleavage positions, were quite abundant. By including all isomiRs and grouping those that share an identical seed sequence, we revealed 222 seeds that were represented in significantly different quantities these two libraries (Table 2.3 and Supplementary Table 3, Appendix A). As the non-canonical isomiRs share most of their sequence with the most highly expressed isomiRs, it is possible that these variants share a common set of targets with the canonical miRNAs. Those isomiRs resulting from variation at the 5' end may be of particular interest as they bear a different seed sequence than the reference miRNA, thus indicative of their potential to target different transcripts. However, whether any of these non-canonical variants associate within RISC remains to be experimentally determined. If they are found to associate within RISC, the presence of isomiRs may have implications in future annotation of miRNAs and the development of new target prediction algorithms. A recent update to miRBase resulted in better agreement between the most abundant sequences in our libraries, but some of the most prevalent isomiRs in our libraries still do not perfectly correspond to the miRBase entry. This suggests that the most abundant sequences have yet to be reliably determined for all miRNAs. If this is the case, further large-scale efforts such as this should result in the truly most common isomiR replacing the erroneous sequences residing in miRBase.

There was a large overlap of predicted target genes for the most prevalent miRNAs enriched in either hESCs or EBs. Since a large proportion of these genes were predicted by TargetScan to be targets of numerous miRNAs, genes were weighted based on their total number of predicted target sites. By taking only the genes with weights above a certain threshold, we derived two relatively small groups of genes that are strong candidates for cooperative targeting by hESC-enriched or EB-enriched miRNAs. This enabled resolution of key gene classes in each group (Figure 2.3), but it does not assert that these are true in vivo targets, nor does it provide a comprehensive list of the targets of these miRNAs. Further directed validation and gene expression profiling will be necessary to elucidate the true in-vivo targets. These lists do, however, provide the foundation for an alternate view of gene regulation in hESCs. Previously, focus has been placed on changes to mRNA levels during differentiation. By shifting focus to the genes targeted by miRNAs in this context, it is plausible that genes previously unlinked to differentiation and pluripotency may be discovered.

Adaptation of a common adapter ligation method to a novel sequencing platform has allowed us to generate a view of the small RNA component of differentiating hESCs at an unprecedented depth. Following a combination of novel miRNA annotation and discovery techniques, we have revealed the largest list to date of miRNA sequences. This list includes a diverse population of miRNA variants, termed isomiRs, with variation at both the 5' and 3' ends. We also present numerous novel human miRNAs, some of which may be important in the control of hESC pluripotency or differentiation. Notably, many of the miRNAs exhibiting the most significant differences between hESCs and EBs appear to regulate genes involved in transcriptional regulation, differentiation and development.

2.5 Methods

2.5.1 Small RNA Library Preparation

Embryonic stem cell samples were derived from lines maintained at the Terry Fox Laboratory. The stem cell line known as H9 was chosen because it is well characterized and was used in the development of differentiation protocols at this facility. Undifferentiated H9 hESCs were cultured on Matrigel (BD Biosciences, San Jose, CA) coated dishes in maintenance medium consisting of DMEM/F12 containing 20% Knockout Serum Replacer (Invitrogen, Carlsbad, CA), 0.1 mM β-mercaptoethanol, 0.1 mM non-essential amino acids, 1 mM glutamine and 4 ng/ml FGF2 (R&D Systems) and conditioned by mitotically inactivated mouse embryonic fibroblasts (Xue et al. 2005). For EB differentiation, hESCs were harvested via 0.05% trypsin (Invitrogen) supplemented with 0.5 mM $CaCl_2$ and the resultant cell aggregates cultured in non-adherent dishes (BD Biosciences) for up to 30 days in maintenance medium lacking FGF2 (medium changes performed as necessary) (Itskovitz-Eldor et al. 2000; Dvash et al. 2004). At appropriate time points RNA was extracted into Trizol and aliquots of total RNA from Day 0 (hESC) and Day 15 (EB) were subjected to miRNA library construction as follows.

For each library, 10 µg of DNase I (DNA-*free*™ kit; Ambion, Austin TX) treated total RNA was size fractionated on a 15% Tris-Borate-EDTA (TBE) urea polyacrylamide gel and a 15-30 base pair fraction excised. RNA was eluted from the polyacrylamide gel slice in 600 µl of 0.3 M NaCl overnight at 4°C. The resulting gel slurry was passed through a Spin-X filter column (Corning) and precipitated in two 300 µl aliquots by the addition of 800 µl of EtOH and 3 µl of Mussel Glycogen (5 mg/ml; Invitrogen). After washing with 75% EtOH, the pellets were allowed to air dry at 25°C and pooled in DEPC water. The 5' RNA adapter (5'-GUU CAG AGU UCU ACA GUC CGA CGA UC-3') was ligated to the RNA pool with T4 RNA ligase (Ambion) in the presence of RNAse Out (Invitrogen) overnight at 25°C. The ligation reaction was stopped by the addition of 2X formamide loading dye. The ligated RNA was size fractionated on a 15% TBE urea polyacrylamide gel and a 40-60 base pair fraction excised. RNA was eluted from the polyacrylamide gel slice in 600 µl of 0.3 M NaCl overnight at 4°C. The RNA was eluted from the gel and precipitated as described above followed by re-suspension in DEPC-treated water.

The 3' RNA adapter (5'- pUC GUA UGC CGU CUU CUG CUU GidT -3'; p = phosphate; idT = inverted deoxythymidine) was subsequently ligated to the precipitated RNA with T4 RNA ligase (Ambion) in the presence of RNAse Out (Invitrogen) overnight at 25°C. The ligation reaction was stopped by the addition of 10 ul of 2X formamide loading dye. Ligated RNA was size fractionated on a 10% TBE urea polyacrylamide gel and the 60-100 base pair fraction excised. The RNA was eluted from the polyacrylamide gel and precipitated from the gel as described above and re-suspended in 5.0 μl of DEPC water. The RNA was converted to single stranded cDNA using Superscript II reverse transcriptase (Invitrogen) and Illumina's small RNA RT-Primer (5'-CAA GCA GAA GAC GGC ATA CGA-3') following the manufacture's instructions. The resulting cDNA was PCR-amplified with Hotstart Phusion DNA Polymerase (NEB) in 15 cycles using Illumina's small RNA primer set (5' -CAA GCA GAA GAC GGC ATA CGA- 3'; 5'-AAT GAT ACG GCG ACC ACC GA-3').

PCR products were purified on a 12% TBE urea polyacrylamide gel and eluted into elution buffer (5:1, LoTE: 7.5 M Ammonium Acetate) overnight at 4°C. The resulting gel slurry was passed through a Spin-X filter (Corning) and precipitated by the addition of 1100 μl of EtOH, 133 μl 7.5 M Ammonium Acetate and 3 μl of Mussel Glycogen (20 mg/ml; Invitrogen). After washing with 75% EtOH, the pellet was allowed to air dry at 25°C and dissolved in EB buffer (Qiagen) by incubation at 4°C for 10 min. The purified PCR products were quantified on the Agilent DNA 1000 chip (Agilent) and diluted to 10 nM for sequencing on the Illumina 1G analyzer.

2.5.2 Small RNA Genome Mapping and Quantification

No reads aligned to the genome after position 28, so we trimmed all reads at 30-nt to reduce the number of unique sequences while still retaining at least 2-nt of linker sequence for the longest alignments. We counted the occurrences of each unique sequence read and used only the unique sequences for further analysis. We aligned each sequence to the human reference genome (NCBI build 36.1) using megablast (version 2.2.11). We filtered these alignments and retained only those that included the first nucleotide of the read and were devoid of insertions, deletions and mismatches. For every read, the longest alignment was determined, and this subsequence, as well as the positions for every alignment of this length was stored in a database (to a maximum of 100 alignments). The sequence following the aligned region was checked for presence of the 3' linker sequence. The counts for all reads containing the same aligned sub-sequence were summed to provide a metric of the total frequency of that small RNA molecule in the original RNA sample. The presence and length of intervening sequences between the alignment and the linker was recorded to enable identification and separate quantification of small RNAs with 3' additions.

2.5.3 Differential Expression Detection

All unique small RNA sequences were compared between the two libraries (hESC and EB) for differential expression using the Bayesian method developed by Audic & Claverie (Audic and Claverie 1997). This approach was developed for analysis of digital gene expression profiles and accounts for the sampling variability of sequences with low counts by modeling the Poisson distribution. Sequences were deemed significantly differentially expressed if the P-value given by this method was < 0.001 and there was at least a 1.5-fold change in sequence counts between the two libraries. Unless stated otherwise, the most frequently observed isomiR was used as the diagnostic sequence for comparison of miRNA expression between libraries. Multiple testing corrections were unnecessary as comparison between two libraries is considered a single test by this method (Stekel et al. 2000).

2.5.4 Small RNA annotation

Each small RNA sequence was annotated if its full sequence contained one or more nucleotides of recognizable 3' linker sequence, had at most 25 perfect full-length alignments to the human genome, and was detected at least twice. Genomic positions of each sequence were compared to genome annotations obtained from the UCSC Genome Browser download page (http://hgdownload.cse.ucsc.edu/downloads.html). The data used for annotation included the EnsEMBL genes, RepeatMasker and sno/miRNA tables. The positions of all miRNA genes were also downloaded from miRBase and used for this positional annotation (http://microrna.sanger.ac.uk/sequences/ftp.shtml). Sequences overlapping annotations from these tables were classified into one of the following classes, listed in the priority used: miRNA, tRNA, rRNA, scaRNA, CD-Box, scRNA, snoRNA, snRNA, srpRNA, genomic repeat or known transcript (exonic or intronic). Transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), which are involved in the cellular translation machinery, were the most abundant classes of sequences after miRNAs, yet these are generally considered relatively uninteresting in the context of small RNAs. The other aforementioned classes of small RNAs are involved in either subcellular targeting (srpRNA), splicing or other chemical modifications required for the maturation of ncRNAs and mRNAs (CD-Box/snRNAs/snoRNAs/scaRNA). All mapped sequences were also directly searched against the currently known human Piwi-interacting RNAs (piRNAs) (Aravin et al. 2007) found currently in piRNAbank (Lakshmi and Agrawal 2007). PiRNAs were previously thought to exist only in mammalian germline cells and function in the silencing of genomic retrotransposons. As piRNAs are considerably longer than miRNAs, we used the first 18-nt of the piRNAs to search for perfect matches amongst our sequences.

Sequences that did not overlap any of the aforementioned annotations (in any one of their genomic positions), or had more than 25 perfect alignments to the genome were automatically classified as 'unknown'. The unknown sequences with 25 or fewer perfect alignments to the genome were used, along with those corresponding to introns and genomic repeats, for novel miRNA prediction. Sequences identified as miRNAs were named based on the specific miRNA gene they overlapped as well as the arm (either 5' or 3') with respect to the pre-miRNA. Where miRNA naming was ambiguous due to identical sequences shared by related miRNA genes, sequences were named arbitrarily by one of their possible parent miRNA genes. For those miRNA genes producing identical mature sequences (e.g., hsa-miR-9-1), the trailing number was dropped from the name. MiRBase (release 10.0) reference sequences were used for the comparison of the common isomiRs to the reference miRNA sequences.

2.5.5 Novel microRNA detection

Candidate miRNA gene loci were identified by finding distinct small RNA sequences lacking annotations that shared partially overlapping genomic positions on the same strand (termed 'hotspots'). These hotspots are suggestive of the presence of two or more isomiRs and were observed for nearly all the known miRNAs in our libraries. 300-nt of genomic sequence flanking each seed was extracted, reverse-complemented where appropriate, and folded using RNALfold (Hofacker 2003), which identifies locally stable sub-structures within a query RNA sequence. Following the known structural constraints of miRNAs (Ambros et al. 2003), the largest un-branched fold back sub-structure was identified from each structure and redundant structures were then removed. Structures were also removed if the seed region (where the original small RNA sequences derived) spanned the loop. The remaining structures and their sequences comprised a set of candidate miRNA genes with expressed sequence and more than one candidate isomiR. The putative pre-miRNA sequences were then enriched for likely real pre-miRNA hairpins using two machine-learning approaches. The previously published method, termed MiPred, relies on a Random Forest (RF) algorithm (Jiang et al. 2007) and uses a combination of structural and thermodynamic parameters. The second approach was devised specifically for its application in this study and employs a support vector machine (SVM) classifier and mostly uses parameters not used by the RF method.

The SVM functionality was provided in the e1071 package for R. The parameters used by this classifier include the relative proportion of each nucleotide and the number of each type of base pair in the optimal pre-miRNA folded structure. As well, some parameters describing the folded structure are common to other SVM approaches including MFEI, AMFE (Zhang et al. 2006), Normalized Pairing Propensity (Ng Kwang Loong et al. 2007) and loop length. The parameters that are unique to this type of sequencing data are descriptors of the seed itself and its positioning within the pre-miRNA structure. Specifically, these include seed length, positioning with respect to the loop, as well as the ratio of paired to unpaired nucleotides within the seed. This classifier was trained using the folded flanking sequence of the miRNAs present in the two libraries for positive examples and the folded flanking sequences of small RNA sequences classified as either tRNA, rRNA snRNA or snoRNA for negative examples. After training, the classifier showed a sensitivity of 0.973 and specificity of 0.988, which represents the upper level of discrimination of pre-miRNAs reported for any machine learning method to date.

The intersection of the positive predictions of the SVM and Random Forest methods (98 total) was used as an initial set of reliable novel miRNA predictions. This was supplemented with miRNAs showing either significant differential expression between the two libraries, significant sequence similarity to known miRNAs, or overlap with an EvoFold prediction for an evolutionarily conserved hairpin (Pedersen et al. 2006).

2.5.6 Cooperative miRNA target prediction

TargetScan (release 4.0) predicted targets for known miRNAs were downloaded from the download page (http://www.targetscan.org/). Only miRNAs with counts of at least 100 in either hESC or EB were included in target analyses. Genes with target sites for at least two co-expressed miRNAs (either hESC-enriched or EB-enriched) were identified as potential cooperative targets. To compensate for potentially biasing some genes with many predicted target sites (due to various issues), these genes were given a lower rank than those with few predicted target sites. The rank score of a gene was calculated by dividing the number of predicted target sites for co-expressed miRNAs by the total number of target sites for that gene. In this context, the co-expressed miRNAs were those that were over-represented in the particular library and had a count of at least 100. We used a cutoff of 0.15 (rank) to produce the two sets of high-ranked candidate cooperative targets of hESC-enriched and EB-enriched miRNAs (Figure 2.3).

2.5.7 Gene Ontology analysis

GoStat (Beissbarth et al. 2004) was employed for identification of significantly enriched 'biological process' Gene Ontology (GO) (Ashburner et al. 2000) terms in the two lists of likely targets of hESC-enriched and EB-enriched miRNAs (P<0.01). Custom software was used to extract Genes and GO terms from the GoStat output. Genes and GO terms were then clustered using hierarchical clustering in R. Heat maps were created in R using the heatmap function (default parameters).

Table 2.1

Top 20 miRNAs differentially expressed between the hESC and EB libraries

miRNA gene (hsa)	Hairpin arm	Most abundant sequence	hESC count	EB count	P-value	fold change
mir-199a	3-p	ACAGTAGTCTGCACATTGGTTA	1110	13163	0	11.86
mir-372	3-p	AAAGTGCTGCGACATTTGAGCGT	1388	13653	0	9.84
mir-122a	5-p	TGGAGTGTGACAATGGTGTTTG	436	2565	0	5.88
mir-152	5-p	TCAGTGCATGACAGAACTTGG	622	3028	0	4.87
mir-10a	5-p	TACCCTGTAGATCCGAATTTGT	948	3887	0	4.10
let-7a	5-p	TGAGGTAGTAGGTTGTATAGTT	11902	2951	0	4.03
mir-302a	5-p	TAAACGTGGATGTACTTGCTTT	36800	9917	0	3.71
mir-222	3-p	AGCTACATCTGGCTACTGGGTCTC	4719	1331	0	3.55
mir-340	5-p	TTATAAAGCAATGAGACTGATT	2247	7198	0	3.20
mir-363	3-p	AATTGCACGGTATCCATCTGTA	5775	17912	0	3.10
mir-21	5-p	TAGCTTATCAGACTGATGTTGAC	39818	21003	0	1.90
mir-26a	5-p	TTCAAGTAATCCAGGATAGGCT	4892	8530	0	1.74
mir-26b	5-p	TTCAAGTAATTCAGGATAGGTT	1003	2957	1.39E-278	2.95
mir-130a	3-p	CAGTGCAATGTTAAAAGGGCAT	2334	4798	2.20E-265	2.06
mir-744	5-p	TGCGGGGCTAGGGCTAACAGCA	4166	1516	9.13E-259	2.75
mir-302b	3-p	TAAGTGCTTCCATGTTTTAGTAG	15169	8855	1.39E-213	1.71
mir-30d	5-p	TGTAAACATCCCCGACTGGAAGCT	2798	4988	3.29E-205	1.78
mir-146b	5-p	TGAGAACTGAATTCCATAGGCTGT	703	2075	2.27E-196	2.95
mir-25	3-p	CATTGCACTTGTCTCGGTCTGA	24268	15875	1.52E-189	1.53
mir-373	3-p	GAAGTGCTTCGATTTTGGGGTGT	60	788	5.75E-184	13.13
mir-423	5-p	TGAGGGGCAGAGAGCGAGACTTT	9844	5538	4.50E-162	1.78
mir-320	3-p	AAAAGCTGGGTTGAGAGGGCGA	5967	2978	1.33E-148	2.00
mir-1	3-p	TGGAATGTAAAGAAGTATGTAT	7421	4051	2.39E-137	1.83
mir-371	5-p	ACTCAAACTGTGGGGGCACTT	1649	3020	6.89E-133	1.83
mir-302d	3-p	TAAGTGCTTCCATGTTTGAGTGT	8599	5047	3.95E-119	1.70

Table 2.2

Putative novel miRNAs exhibiting significant differential expression

Chromo-some	Length (nt)	most common sequence	hESC count	EB count	Fold Change	Putative miRNA family	Seed sequence
chr2	19	AATGGATTTTTGGAGCAGG	374	245	1.53	-	AUGGAUU
chr11	17	TAAGTGCTTCCATGCTT	84	131	1.56	mir-503	AAGUGCU
chrX	22	TTCATTCGGCTGTCCAGATGTA	1269	774	1.64	-	UCAUUCG
chr16	19	GCCTGTCTGAGCGTCGCTT	194	114	1.70	-	CCUGUCU
chr1	21	TATTCATTTATCCCCAGCCTACA	126	67	1.88	mir-664	AUUCAUU
chr19	17	TCCCTGTTCGGGCGCCA	670	1302	1.94	mir-761	CCCUGUU
chr1	19	TGGATTTTTGGATCAGGGA	40	85	2.13	-	GGAUUUU
chr1	17	AACAGCTAAGGACTGCA	72	31	2.32	-	ACAGCUA
chr6	19	GTGTAAGCAGGGTCGTTTT	639	273	2.34	mir-565	UGUAAGC
chrX	22	TACGTAGATATATATGTATTTT	20	54	2.70	mir-620	ACGUAGA
chr5	18	GTCCCTGTTCAGGCGCCA	24	66	2.75	-	UCCCUGU
chr15	26	GATGATGATGGCAGCAAATTCTGAAA	81	27	3.00	mir-376	AUGAUGA
chr22	19	ACGTTGGCTCTGGTGGTG	35	11	3.18	mir-650	CGUUGGC
chr8	18	ATCCCCAGCACCTCCACC	59	18	3.28	mir-650	UCCCCAG
chr10	22	TGCTGGATCAGTGGTTCGAGTC	16	57	3.56	mir-502	GCUGGAU
chr15	23	CCTCAGGGCTGTAGAACAGGGCT	163	45	3.62	-	CUCAGGG
chr4	22	CTGGACTGAGCCGTGCTACTGG	44	161	3.66	mir-661	UGGACUG
chr10	22	ACTCGGCGTGGCGTCGGTCGTG	228	50	4.56	-	CUCGGCG
chr2	24	TTGCAGCTGCCTGGGAGTGACTTC	63	13	4.85	-	UGCAGCU
chr4	17	TCCGAGTCACGGCACCA	309	55	5.62	mir-345	CCGAGTC
chr1	17	GTGGGGGAGAGGCTGTC	103	17	6.06		UGGGGGG
chr1	22	ATGGGTGAATTTGTAGAAGGAT	7	45	6.43	mir1072	UGGGUGA
chr7	19	AAACTGTAATTACTTTTGG	21	3	7.00	mir-548a	AACUGUA
chrX	18	GCATGGGTGGTTCAGTGG	407	52	7.83	mir-146a	CAUGGGU
chr10	24	AGCCTGGAAGCTGGAGCCTGCAGT	25	3	8.33	mir-566	GCCUGGA
chrX	22	AGGAGGAATTGGTGCTGGTCTT	25	2	12.50	*mir-766**	GGAGGAA
chr22	17	AAAAAGACACCCCCCAC	189	21	9.00	mir396b	AAAAGAC
chr14	22	ACCCGTCCCGTTCGTCCCCGGA	4	53	13.25	mir-638	CCCGUCC
chr11	19	ATGGATAAGGCTTTGGCTT	31	2	15.50	-	UGGAUAA
chr15	18	CGGGCGTGGTGGTGGGGG	35	2	17.50	mir-566	GGGCGUG
chr11	24	GTGGGAAGGAATTACAAGACAGTT	72	2	36.00	-	UGGGAAG

Table 2.3

Top 25 statistically over-represented seeds of miRNA sequences found in hESCs or EBs (Fisher's Exact Test with Bonferroni correction)

Seed	hESC count	EB count	Fold Change	Corrected P-value	miRNAs with seed
AGUAGUC	86	1478	17x	0	**mir-199a**
CAGUAGU	1801	22288	12.4x	0	mir-199a
ACAGUAG	241	2749	11.4x	0	**mir-199a**
UUAAACG	3452	398	8.7x	0	**mir-302a-5p**
CUUAAAC	16762	2332	7.2x	0	mir-302a-5p
GGAGUGU	846	5216	6.2x	0	mir-122
AGUGCUG	2484	12641	5.1x	0	mir-372,mir-512
ACCCUGU	1503	5689	3.8x	0	mir-10
AAACGUG	55712	16461	3.4x	0	**mir-302a-5p,**mir-424
UAUAAAG	3002	9702	3.2x	0	mir-340
GAGAACU	1547	4976	3.2x	0	mir-146
UUGCACG	1845	5561	3.0x	0	mir-363
GAGGUAG	29280	10258	2.9x	0	let-7
CAGUGCA	3596	8559	2.4x	0	mir-148,mir-152,**mir-130**
AAAGCUG	16448	7613	2.2x	0	mir-320
AGUGCAA	4463	9594	2.1x	0	mir-454,mir-301,mir-130
UCAAGUA	6898	14089	2.0x	0	mir-26
GCUACAU	47591	24280	2.0x	0	mir-221,mir-222
GUAAACA	8789	15479	1.8x	0	mir-30
GAGGGGC	17586	10518	1.7x	0	mir-423/mir-885
AGCUUAU	65376	39424	1.7x	0	mir-21,mir-590
CUGGACU	17819	25378	1.4x	0	mir-378
GCAGCAU	153193	187218	1.2x	0	mir-103,mir-107,mir-885
GCGGGGC	5220	1956	2.7x	8.76E-305	mir-744
CUCAAAC	4306	8173	1.9x	1.62E-296	**mir-92a-5p,**mir-371

miRNAs indicated in bold are cases in which their non-canonical isomiR contains this seed

Figure 2.1 a

Distribution of sequence counts for the different classes of small RNAs.

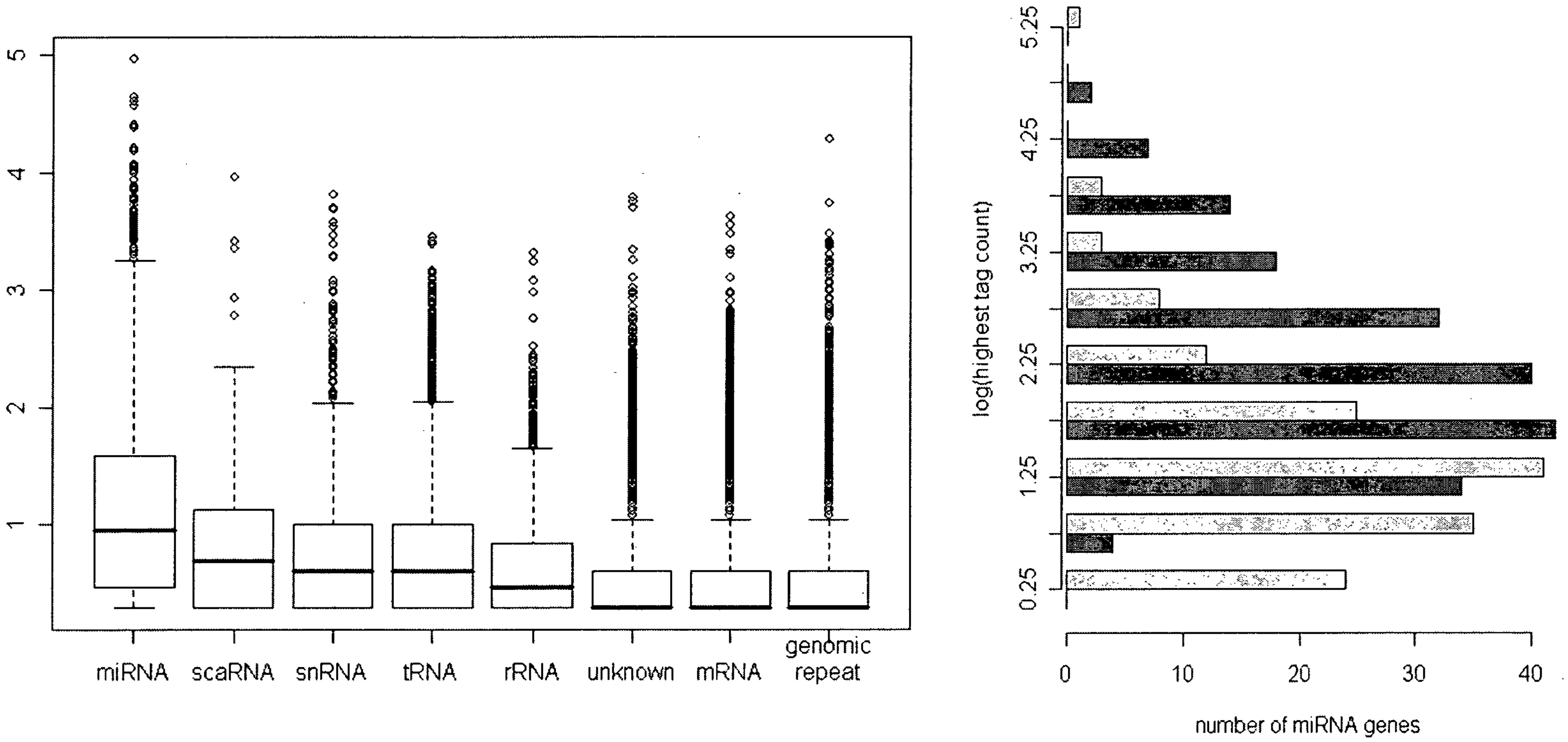
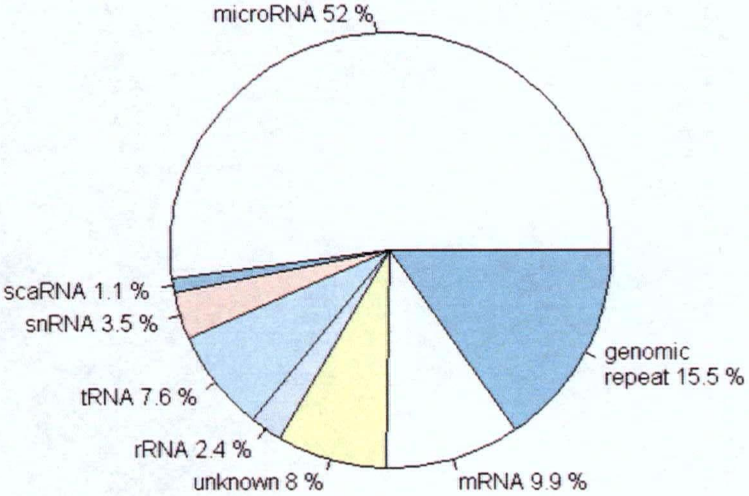


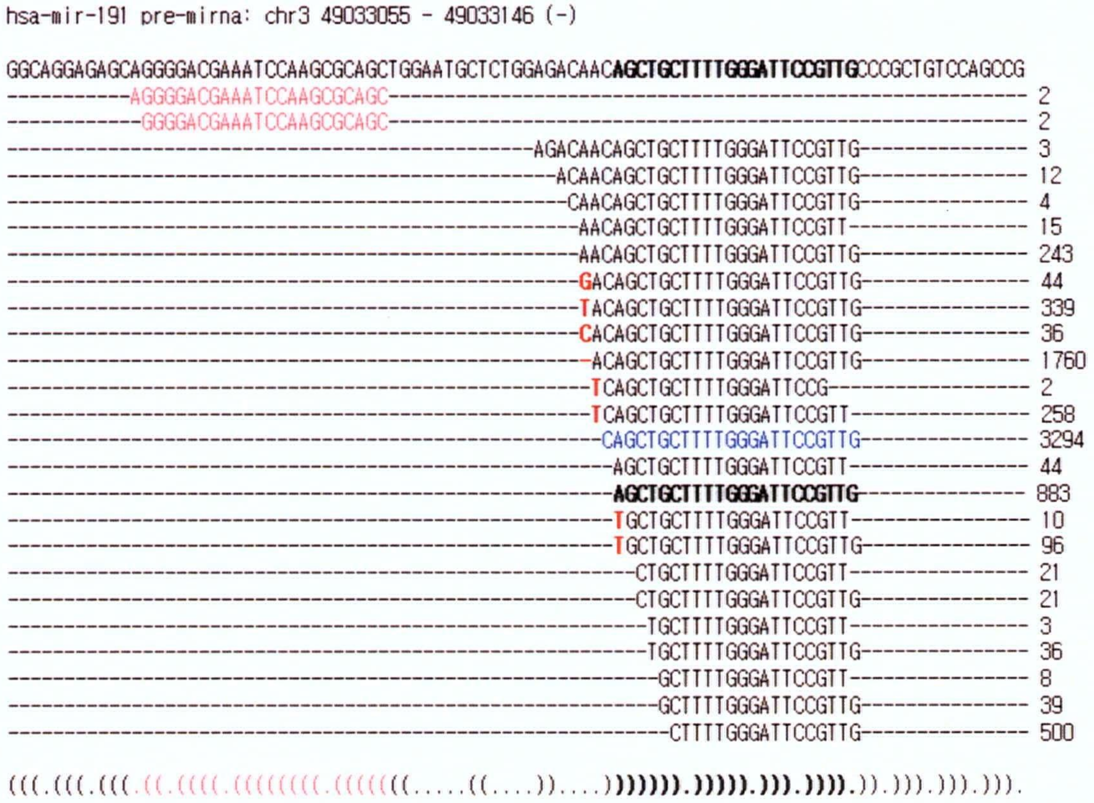
Figure 2.1 b



a) The box plot (left) shows the relative expression levels of sequences in each of the 8 major classes of small RNAs ($\log_{10}$ transformation) in the hESC small RNA library. MiRNAs were the most highly expressed class (mean sequence count=253, median sequence count=9). The most abundant miRNA in this library was mir-103, which had 91,398 instances of the most common isomiR in this library and nearly 120,000 in the matched library (EB). The highest log-transformed count between the two libraries (right) for all miRNAs identified as differentially expressed (dark grey) is roughly normal (mean=2.32, median=2.17), representing a count of 1743 and 151 respectively). The miRNAs detected in at least one of the libraries but were not significantly differentially expressed are shown in light grey for comparison. There is a slight enrichment of miRNAs with lower absolute expression in this group (mean=1.35,median=1.24), suggesting miRNAs with higher absolute expression levels are more likely to be identified as differentially expressed. b) Total counts for the 8 classes in the box plot are summarized. They are represented as a fraction of the total sequences that had at least one perfect alignment to the human reference genome (1,631,559 total).

Figure 2.2

The repertoire of isomiRs and 3' modifications of hsa-mir-191.



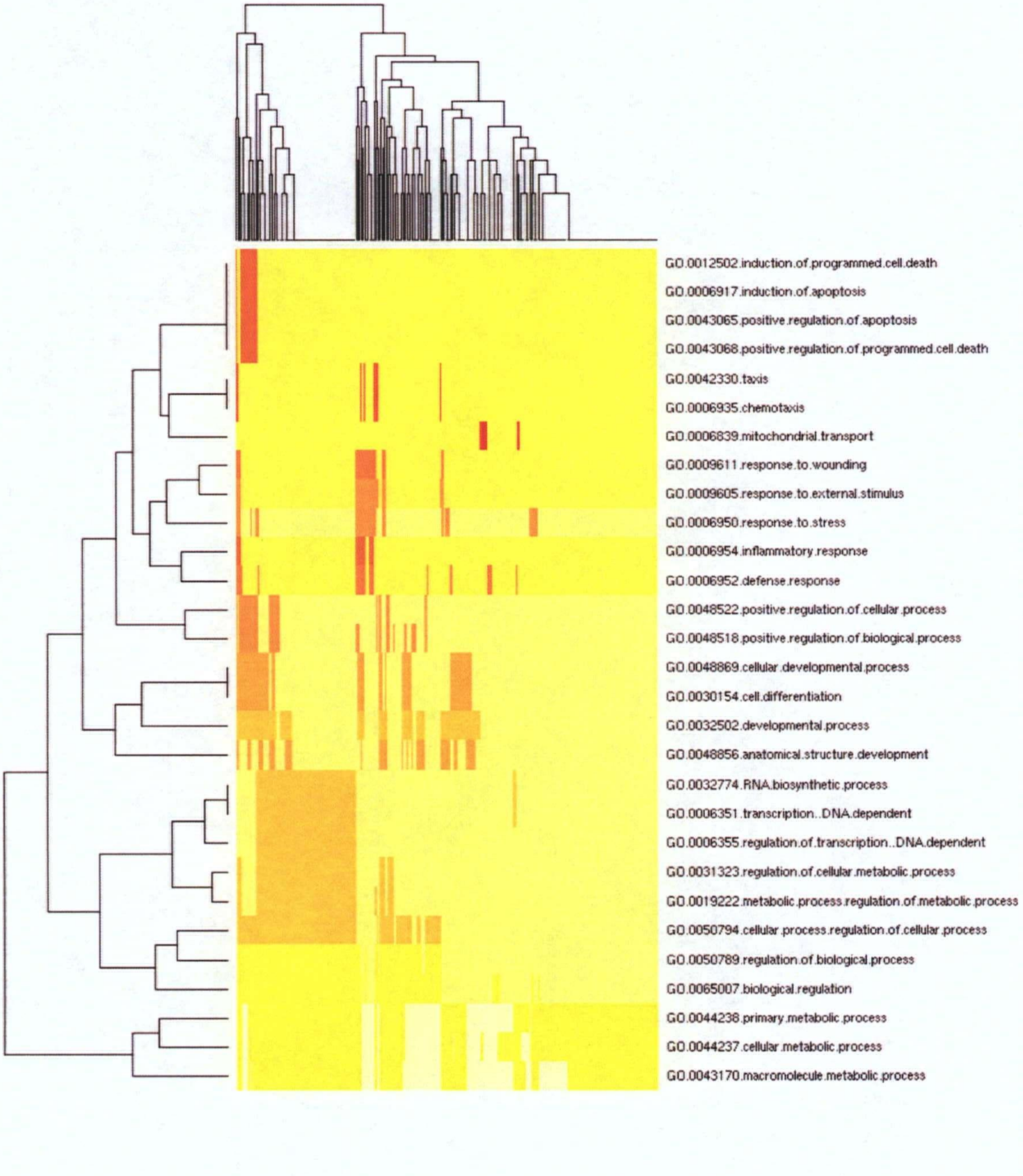
```
hsa-mir-191 pre-mirna: chr3 49033055 - 49033146 (-)

GGCAGGAGAGCAGGGGACGAAATCCAAGCGCAGCTGGAATGCTCTGGAGACAACAGCTGCTTTTGGGATTCCGTTGCCCGCTGTCCAGCCG
-----------AGGGGACGAAATCCAAGCGCAGC--------------------------------------------------------- 2
------------GGGGACGAAATCCAAGCGCAGC--------------------------------------------------------- 2
-----------------------------------------------AGACAACAGCTGCTTTTGGGATTCCGTTG--------------- 3
-------------------------------------------------ACAACAGCTGCTTTTGGGATTCCGTTG--------------- 12
--------------------------------------------------CAACAGCTGCTTTTGGGATTCCGTTG--------------- 4
---------------------------------------------------AACAGCTGCTTTTGGGATTCCGTT---------------- 15
---------------------------------------------------AACAGCTGCTTTTGGGATTCCGTTG--------------- 243
---------------------------------------------------GACAGCTGCTTTTGGGATTCCGTTG--------------- 44
---------------------------------------------------TACAGCTGCTTTTGGGATTCCGTTG--------------- 339
---------------------------------------------------CACAGCTGCTTTTGGGATTCCGTTG--------------- 36
----------------------------------------------------ACAGCTGCTTTTGGGATTCCGTTG--------------- 1760
----------------------------------------------------TCAGCTGCTTTTGGGATTCCG------------------ 2
----------------------------------------------------TCAGCTGCTTTTGGGATTCCGTT---------------- 258
-----------------------------------------------------CAGCTGCTTTTGGGATTCCGTTG--------------- 3294
------------------------------------------------------AGCTGCTTTTGGGATTCCGTT---------------- 44
------------------------------------------------------AGCTGCTTTTGGGATTCCGTTG--------------- 883
------------------------------------------------------TGCTGCTTTTGGGATTCCGTT---------------- 10
------------------------------------------------------TGCTGCTTTTGGGATTCCGTTG--------------- 96
--------------------------------------------------------CTGCTTTTGGGATTCCGTT---------------- 21
--------------------------------------------------------CTGCTTTTGGGATTCCGTTG--------------- 21
---------------------------------------------------------TGCTTTTGGGATTCCGTT---------------- 3
---------------------------------------------------------TGCTTTTGGGATTCCGTTG--------------- 36
----------------------------------------------------------GCTTTTGGGATTCCGTT---------------- 8
----------------------------------------------------------GCTTTTGGGATTCCGTTG--------------- 39
-----------------------------------------------------------CTTTTGGGATTCCGTTG--------------- 500

(((.(((.(((.((.((((.((((((((.(((((((.....((....))....))))))).))))).))).)))).)).))).))).))).
```

A diverse variety of isomiRs were observed for many of the known and novel miRNAs. Sequences representing the miR* were also commonly observed (highlighted in pink). The reference miRNA sequence from miRBase (highlighted in bold) was generally not the most frequently observed isomiR (shown in blue). Though variation at the 3' end is generally much more common than at the 5' end, target preference of these isomiRs may differ. Sequences with evidence of 3' additions of nucleotides (Red) were common, with certain miRNAs more heavily modified than others. The predicted structure of the pre-miRNA is represented at the bottom.

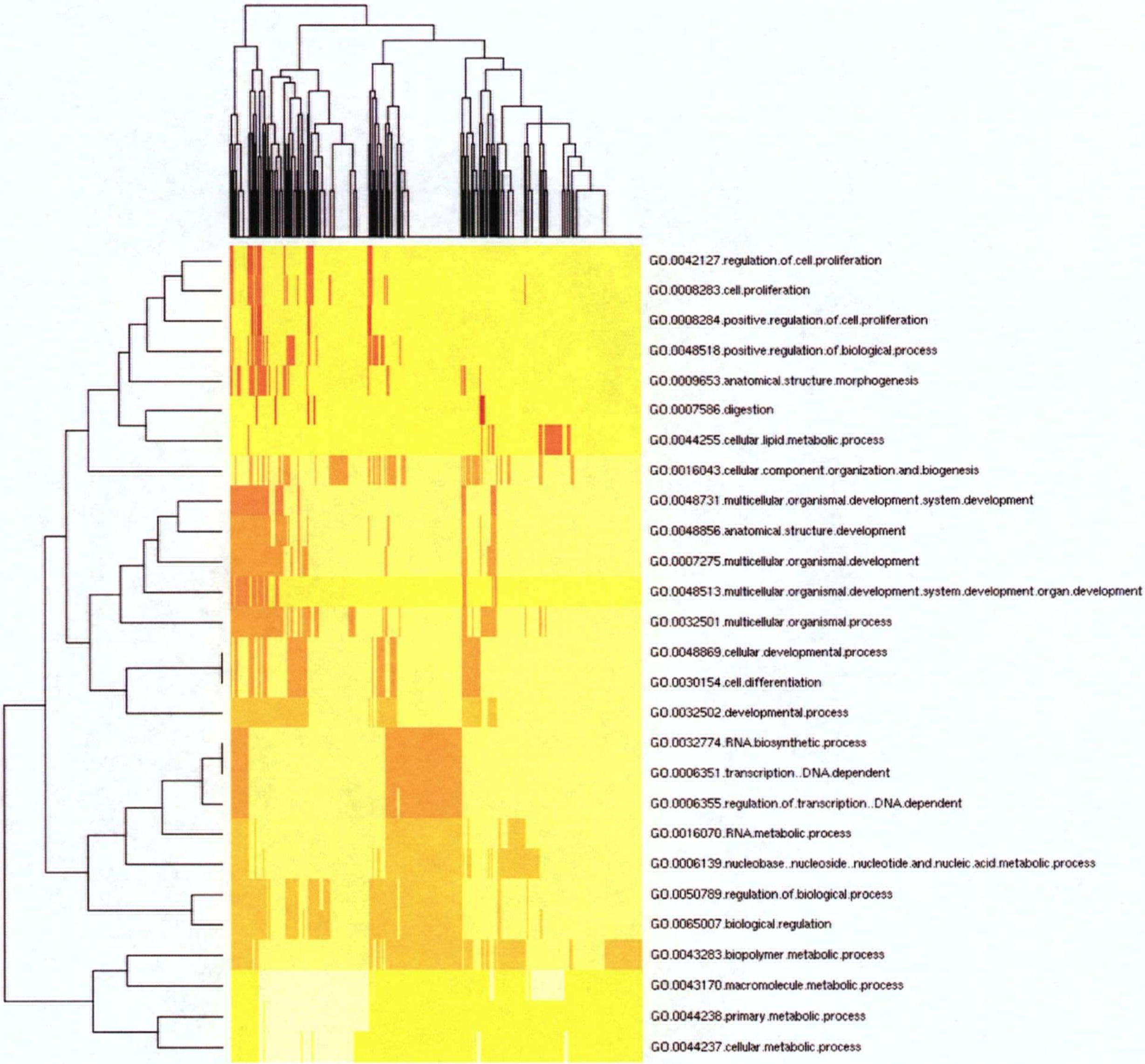
Figure 2.3

Clustering of over-represented Gene Ontology classes in predicted targets of differential microRNAs.

a)



b)



Shown are heat map representations of Gene Ontology (GO) (Ashburner et al. 2000) terms over-represented amongst predicted cooperative targets (y-axis) of (a) hESC-enriched miRNAs and (b) EB-enriched miRNAs. All genes with statistically over-represented GO annotations were included (p<0.01, x-axis) as identified by GoStat (Beissbarth et al. 2004). GO terms common to both sets of genes were those involved in transcriptional regulation, differentiation and development. Those GO terms unique to hESC-enriched miRNA targets were associated with programmed cell death, response to stress, and cell motility while those unique to EB-enriched miRNA targets describe various aspects of cell proliferation regulation as well as nucleotide and nucleic acid metabolism

2.6 References

Abeyta, M. J., Clark, A. T., Rodriguez, R. T., Bodnar, M. S., Pera, R. A., and Firpo, M. T. 2004. Unique gene expression signatures of independently-derived human embryonic stem cell lines. *Hum Mol Genet* **13:** 601-8.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Drey, S. R., Griffiths-Jones, S., Marshall, M., Matzke, M. et al. A uniform system for microRNA annotation. 2003. *RNA* **9**:277-9.

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316:** 744-747.

Aravin, A. and Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett* **579:** 5830-40.

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:** 25-9.

Audic, S. and Claverie, J. M. 1997. The significance of digital gene expression profiles. *Genome Res* **7:** 986-95.

Beissbarth, T. and Speed, T. P. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20:** 1464-1465.

Berezikov, E., Cuppen, E., and Plasterk, R. H. 2006a. Approaches to microRNA discovery. *Nat Genet* **38 Suppl:** S2-7.

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. 2006b. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38:** 1375-7.

Bernstein E., Kim S. Y., Carmell M. A., Murchison E. P., Alcorn H., Li M. Z., Mills A. A., Elledge S. J., Anderson K. V. and Hannon G. J. 2003. Dicer is essential for mouse development. Nat Genet. 35(3): 215-7.

Bhattacharya B, Cai J, Luo Y, Miura T, Mejido J, Brimble SN, Zeng X, Schulz TC, Rao MS,Puri RK. 2005. Comparison of the gene expression profile of undifferentiated human embryonic stem cell lines and differentiating embryoid bodies. *BMC Dev Biol* **5:** pre-print.

Bhattacharya, B., Miura, T., Brandenberger, R., Mejido, J., Luo, Y., Yang, A. X., Joshi, B. H., Ginis, I., Thies, R. S., Amit, M., et al. 2004. Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood* **103:** 2956-64.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947-56.

Cai, X., Hagedorn, C. H., and Cullen, B. R. 2004. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna* **10:** 1957-66.

Chen, C., Ridzon, D., Lee, C. T., Blake, J., Sun, Y., and Strauss, W. M. 2007. Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets. *Mamm Genome* .

Cummins, J. M., He, Y., Leary, R. J., Pagliarini, R., A., D. L.,Jr, Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A. E., Labourier, E., et al. 2006. The colorectal microRNAome. *Proc Natl Acad Sci U S A* **103:** 3687-92.

Davison, T. S., Johnson, C. D., and Andruss, B. F. 2006. Analyzing micro-RNA expression using microarrays. *Methods Enzymol* **411:** 14-34.

Dvash, T., Mayshar, Y., Darr, H., McElhaney, M., Barker, D., Yanuka, O., Kotkow, K. J., Rubin, L. L., Benvenisty, N., and Eiges, R. 2004. Temporal gene expression during differentiation of human embryonic stem cells and embryoid bodies. *Hum Reprod* **19:** 2875-83.

Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., Law, T. F., Grant, S. R., Dangl, J. L., et al. 2007. High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. *PLoS ONE* **2:** e219.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Van Dongen, S., Inoue, K., Enright, A. J., and Schier, A. F. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312:** 75-9.

Griffiths-Jones, S. 2006. miRBase: the microRNA sequence database. *Methods Mol Biol* **342:** 129-38.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27:** 91-105.

Hirst, M., Delaney, A., Rogers, S. A., Schnerch, A., Persaud, D. R., O'Connor M, D., Zeng, T., Moksa, M., Fichter, K., Mah, D., et al. 2007. LongSAGE profiling of nine human embryonic stem cell lines. *Genome Biol* **8:** R113.

Hofacker, I. L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res* **31:** 3429-31.

Houbaviy, H. B., Murray, M. F., and Sharp, P. A. 2003. Embryonic stem cell-specific MicroRNAs. *Dev Cell* **5:** 351-8.

Itskovitz-Eldor, J., Schuldiner M., Karsenti, D., Eden, A., Yanuka, O., Amit, M., Soreq, H. and Benvenisty, N. 2000. Differentiation of human embryonic stem cells into embryoid bodies compromising the three embryonic germ layers. Mol Med. 6(2): 88-95.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res* **35:** W339-44.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. 2004. Human MicroRNA targets. *PLoS Biol* **2:** e363.

Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., and Carrington, J. C. 2007. Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. *PLoS Biol* **5:** e57.

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. 2007. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315:** 1137-40.

Lakshmi, S. and Agrawal S. 2007. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* In press.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., et al. 2007. A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129:** 1401-14.

Lewis, B. P., Burge, C. B., and Bartel, D. P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120:** 15-20.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95-8.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-80.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S. S., Skogen, J. W., and Meyers, B. C. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Res* **34:** D731-5.

Ng Kwang Loong, S. and Mishra, S. K. 2007. Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. *Rna* **13:** 170-87.

O'Toole, A. S., Miller, S., Haines, N., Zink, M. C., and Serra, M. J. 2006. Comprehensive thermodynamic analysis of 3' double-nucleotide overhangs neighboring Watson-Crick terminal base pairs. *Nucleic Acids Res* **34:** 3338-44.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2:** e33.

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F. A., van Dyk, L. F., Ho, C. K., Shuman, S., Chien, M., et al. 2005. Identification of microRNAs of the herpesvirus family. *Nat Methods* **2:** 269-76.

Porkka, K. P., Pfeiffer, M. J., Waltering, K. K., Vessella, R. L., Tammela, T. L., and Visakorpi, T. 2007. MicroRNA expression profiling in prostate cancer. *Cancer Res* **67:** 6130-5.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. 2006. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* **20:** 3407-25.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127:** 1193-207.

Sato, N., Sanjuan, I. M., Heke, M., Uchida, M., Naef, F., and Brivanlou, A. H. 2003. Molecular signature of human embryonic stem cells and its comparison with the mouse. *Dev Biol* **260:** 404-13.

Shivdasani, R. A. 2006. MicroRNAs: regulators of gene expression and cell differentiation. *Blood* **108:** 3646-53.

Skottman, H., Mikkola, M., Lundin, K., Olsson, C., Stromberg, A. M., Tuuri, T., Otonkoski, T., Hovatta, O., and Lahesmaa, R. 2005. Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells* **23:** 1343-56.

Song, L. and Tuan, R. S. 2006. MicroRNAs and cell differentiation in mammalian development. *Birth Defects Res C Embryo Today* **78:** 140-9.

Stekel, D. J., Git, Y., and Falciani, F. The comparison of gene expression from multiple cDNA libraries. 2000. *Genome Res* **10**: 2055-61.

Strauss, W. M., Chen, C., Lee, C. T., and Ridzon, D. 2006. Nonrestrictive developmental regulation of microRNA gene expression. *Mamm Genome* **17:** 833-40.

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., Cha, K. Y., Chung, H. M., Yoon, H. S., Moon, S. Y., et al. 2004. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol* **270:** 488-98.

Wang Y., Medvid R., Melton C., Jaenisch R. and Blelloch R. 2007. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet. 39(3): 380-5.

Watanabe, T., Takeda, A., Tsukiyama, T., Mise, K., Okuno, T., Sasaki, H., Minami, N., and Imai, H. 2006. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev* **20:** 1732-43.

Xue, C., Li, F., He, T., Liu, G. P., Li, Y., and Zhang, X. 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6:** 310.

Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J. K., and Sun, Q. 2007. Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). *Genome Biol* **8:** R96.

Zhang, B., Pan, X. P., Cox, S. B., Cobb, G. P., and Anderson, T. A. 2006. Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci* **63:** .

Zhao, J. J., Hua, Y. J., Sun, D. G., Meng, X. X., Xiao, H. S., and Ma, X. 2006. Genome-wide microRNA profiling in human fetal nervous tissues by oligonucleotide microarray. *Childs Nerv Syst* **22:** 1419-25.

3 CONCLUDING REMARKS

3.1 Goals accomplished

A strong motivating factor in this project was the general lack of software for analyzing miRNA sequence data. The software that is currently available for this purpose lags far behind current technological capabilities, with recent releases still limited to analysis of data from 'dideoxy' (Sanger) sequencing methods (Tian et al. 2007). Considering the benefits of next-generation sequencing methods presented in the previous chapters, use of older sequencing approaches for miRNA profiling is likely to diminish in favour of these new methods. To catalyze this transition, the scientific community will require access to software that is capable of analyzing and summarizing data from experiments employing these new platforms. As researchers may be apprehensive to adopt new techniques, they must also be convinced of the benefits such a change will provide. To this end, many valuable bioinformatic tools for processing Illumina small RNA sequence data have been developed throughout this project. Although improvements are required to enhance the scope and robustness of this software, it has utility in interpreting the vast amount of data that is produced by this technology. Specifically, chapter two provides an example of the various types of information that can be extracted from these data using this set of tools.

Overall, the analyses included here revealed some underappreciated intricacies of miRNA biogenesis and some specific insights into gene regulation in human embryonic stem cells. These results include the observation that miRNAs demonstrate variability at the 5' and 3' ends, likely resulting from inconsistent Drosha and Dicer cleavage; undergo RNA editing by ADAR; and are subject to a plethora of 3' extensions that appear to be specific to certain miRNAs and evolutionarily conserved. The biological significance of these findings is unclear and they may later prove to be artefactual. However that most of these observations are recapitulations of other groups suggests they are at least reproducible. As well, this analysis has resulted in a list of 170 novel miRNA genes, some of which may also be important in human embryonic stem cells. Of the known miRNA genes identified here, many were already known to be important in hESCs (Suh et al. 2004). This technique also revealed the expression of miRNAs expressed at relatively low levels, whose expression has never before reported in hESCs. Some of these also demonstrated significant expression changes between these samples, thus increasing the list of miRNAs potentially important in either pluripotency or differentiation.

Key to discovering the role of these miRNAs in hESCs is the correct identification of their gene targets. Target genes, predicted by one of the most reliable target prediction algorithms (TargetScan), were ranked for their potential for cooperative regulation by co-expressed miRNAs. Although this reduced the sensitivity of target prediction considerably, it provided smaller sets of genes that are potentially important in pluripotency or differentiation. The statistically over-represented gene ontology (GO) terms (Figure 2.3) support this, as these smaller gene lists are enriched for genes involved in these processes.

3.2 Future directions

The data included here represent only a small sample of the capabilities of this new sequencing technology. The Illumina sequencing apparatus allows 8 separate libraries to be sequenced concurrently, while offering a choice between enhanced sampling depth or multiple biological replicates. Biological and technical replicates would allow more robust statistical tests to be applied, potentially revealing more subtle miRNA expression changes that were missed in this study. Samples extracted from differentiating hESCs at earlier and later time points could also reveal trends in these miRNA expression changes, which may fluctuate during the process of differentiation. Later time points may also reveal further stages of differentiation or could reveal whether the EB sample was fully differentiated (devoid of Pluripotent cells). These additional experiments would solidify our current views of the hESC miRNA transcriptome and would likely reveal the variability in miRNA levels that are unrelated to differentiation.

A complete view of the miRNA expression levels and changes associated with differentiation are of limited utility without knowledge of their gene targets. However, assays for validating miRNA/target interactions are laborious, thus few have been biologically verified to date. In fact, a publicly available database known as TarBase (Sethupathy et al. 2006) hosts the currently known interactions and it includes very few of the miRNAs highlighted in this study. Hence, an important next step would be to validate of some of the candidate targets of cooperative miRNA regulation identified here (Figure 2.3). Perhaps more importantly, it is necessary to develop large-scale efforts to identify miRNA/target interactions be developed. A few recent studies have demonstrated successful co-immunoprecipitation between components of RISC and their bound mRNAs in both human and mouse tissues (Duan et al. 2006; Beitzinger et al. 2007). Such approaches may be adapted into high-throughput methods for identifying mRNA targets of the miRNA pathway and potentially the specific sites bound by these proteins. If a full view of all mRNAs targeted by miRNAs can be produced, the field of miRNA research will be strengthened considerably.

Finally, a few of the observations in this project are of unknown functional relevance. RNA editing of pre-miRNAs can affect the mature miRNA sequence and is thought to affect a miRNAs repertoire of targets (Blow et al. 2006; Kawahara et al. 2007). However, as the relative levels of edited miRNAs found here were less than 1% of the unmodified forms, it seems unlikely that they are functional. Further work using more small RNA sequence libraries should reveal whether editing of miRNAs is a widespread phenomenon. Two other modes of miRNA variation were discussed here: 3' extension by an unknown mechanism and end variation due to variable Drosha and Dicer cleavage sites. Both can result in miRNAs of different lengths, and the 5' variants have a different seed sequence than the canonical miRNA. As they may also have different *in vivo* targets, determining whether these isomiRs associate with functional RISCs is an important next step to concluding whether they are also functional. The 3' variability of miRNAs is less likely to affect their targets, as miRNA/mRNA complementarity in this region is less important. Instead, it is possible that variation at the 3' end performs a different function entirely. A recent study has shown that the sequence at the extreme 3' end of some miRNAs can affect their subcellular localization (Hwang et al. 2007). Hence, it is possible that further study may reveal that these isomiRs may reside preferentially in different cellular locales.

3.3 References

Beitzinger, M., Peters, L., Zhu, J. Y., Kremmer, E., and Meister, G. 2007. Identification of Human microRNA Targets From Isolated Argonaute Protein Complexes. *RNA Biol* **4:** .

Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R., and Stratton, M. R. 2006. RNA editing of human microRNAs. *Genome Biol.* **7:** R27.

Duan, R. and Jin, P. 2006. Identification of messenger RNAs and microRNAs associated with fragile X mental retardation protein. *Methods Mol. Biol.* **342:** 267-276.

Hwang, H. W., Wentzel, E. A., and Mendell, J. T. 2007. A hexanucleotide element directs microRNA nuclear import. *Science* **315:** 97-100.

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A. G., and Nishikura, K. 2007. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315:** 1137-40.

Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *Rna* **12:** 192-7.

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., Cha, K. Y., Chung, H. M., Yoon, H. S., Moon, S. Y., et al. 2004. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol* **270:** 488-98.

Tian, F., Zhang, H., Zhang, X., Song, C., Xia, Y., Wu, Y., and Liu, X. 2007. miRAS: a data processing system for miRNA expression profiling study. *BMC Bioinformatics* **8:** 285.

APPENDICES
Appendix A: Supplementary Tables

Supplementary Table 1: All significantly differentially expressed miRNAs

microRNA (hsa)	arm	most common sequence	hESC count	EB count	P-value	fold change
mir-199a	3-p	ACAGTAGTCTGCACATTGGTTA	1110	13163	0	11.86
mir-372	3-p	AAAGTGCTGCGACATTTGAGCGT	1388	13653	0	9.84
mir-122a	5-p	TGGAGTGTGACAATGGTGTTTG	436	2565	0	5.88
mir-152	5-p	TCAGTGCATGACAGAACTTGG	622	3028	0	4.87
mir-10a	5-p	TACCCTGTAGATCCGAATTTGT	948	3887	0	4.10
let-7a	5-p	TGAGGTAGTAGGTTGTATAGTT	11902	2951	0	4.03
mir-302a	5-p	TAAACGTGGATGTACTTGCTTT	36800	9917	0	3.71
mir-222	3-p	AGCTACATCTGGCTACTGGGTCTC	4719	1331	0	3.55
mir-340	5-p	TTATAAAGCAATGAGACTGATT	2247	7198	0	3.20
mir-363	3-p	AATTGCACGGTATCCATCTGTA	5775	17912	0	3.10
mir-21	5-p	TAGCTTATCAGACTGATGTTGAC	39818	21003	0	1.90
mir-221	3-p	AGCTACATTGTCTGCTGGGTTTC	16275	8716	0	1.87
mir-26a	5-p	TTCAAGTAATCCAGGATAGGCT	4892	8530	0	1.74
mir-26b	5-p	TTCAAGTAATTCAGGATAGGTT	1003	2957	1.39E-278	2.95
mir-130a	3-p	CAGTGCAATGTTAAAAGGGCAT	2334	4798	2.20E-265	2.06
mir-744	5-p	TGCGGGGCTAGGGCTAACAGCA	4166	1516	9.13E-259	2.75
mir-594	5-p	ATGGATAAGGCATTGGC	1717	211	1.96E-253	8.14
mir-302b	3-p	TAAGTGCTTCCATGTTTTAGTAG	15169	8855	1.39E-213	1.71
mir-30d	5-p	TGTAAACATCCCCGACTGGAAGCT	2798	4988	3.29E-205	1.78
mir-146b	5-p	TGAGAACTGAATTCCATAGGCTGT	703	2075	2.27E-196	2.95
mir-25	3-p	CATTGCACTTGTCTCGGTCTGA	24268	15875	1.52E-189	1.53
mir-373	3-p	GAAGTGCTTCGATTTTGGGGTGT	60	788	5.75E-184	13.13
mir-423	5-p	TGAGGGGCAGAGAGCGAGACTTT	9844	5538	4.50E-162	1.78
mir-320	3-p	AAAAGCTGGGTTGAGAGGGCGA	5967	2978	1.33E-148	2.00
mir-1	3-p	TGGAATGTAAAGAAGTATGTAT	7421	4051	2.39E-137	1.83
mir-371	5-p	ACTCAAACTGTGGGGGCACTT	1649	3020	6.89E-133	1.83
mir-302d	3-p	TAAGTGCTTCCATGTTTGAGTGT	8599	5047	3.95E-119	1.70
mir-378	3-p	ACTGGACTTGGAGTCAGAAGG	4021	6149	5.05E-119	1.53
mir-30a	5-p	TGTAAACATCCTCGACTGGAAGCT	1654	2822	2.67E-105	1.71
mir-302c	3-p	TAAGTGCTTCCATGTTTCAGT	3929	1984	4.03E-95	1.98
let-7e	5-p	TGAGGTAGGAGGTTGTATAGTT	899	1794	4.88E-95	2.00
mir-92	5-p	AGGTTGGGATCGGTTGCAATGCT	788	138	2.91E-94	5.71
mir-204	5-p	TTCCCTTTGTCATCCTATGCCT	690	1505	7.42E-94	2.18
mir-421	3-p	ATCAACAGACATTAATTGGGCGC	579	1331	1.31E-90	2.30
mir-20b	5-p	CAAAGTGCTCATAGTGCAGGTAG	281	862	2.51E-86	3.07
mir-130b	3-p	CAGTGCAATGATGAAAGGGCAT	836	1626	4.23E-82	1.94
mir-146a	5-p	TGAGAACTGAATTCCAT	107	556	3.04E-78	5.20
mir-181a	5-p	AACATTCAACGCTGTCGGTGAGTTT	571	1220	8.39E-74	2.14
mir-106a	5-p	AAAAGTGCTTACAGTGCAGGTAG	236	696	4.74E-67	2.95
let-7c	5-p	TGAGGTAGTAGGTTGTATGGTT	360	30	1.94E-64	12.00
mir-184	3-p	TGGACGGAGAACTGATAAGGGT	403	44	2.15E-64	9.16
mir-708	5-p	AAGGAGCTTACAATCTAGCTGGG	397	990	3.68E-64	2.49
mir-205	5-p	TCCTTCATTCCACCGGAGTCTGT	502	0	9.77E-63	n/a
mir-127	3-p	TCGGATCCGTCTGAGCTTGGCT	81	406	9.77E-63	5.01

microRNA (hsa)	arm	most common sequence	hESC count	EB count	P-value	fold change
let-7i	5-p	TGAGGTAGTAGTTTGTGCTGTT	347	33	4.34E-59	10.52
mir-302a	3-p	TAAGTGCTTCCATGTTTTGGTGA	5239	3237	8.21E-58	1.62
mir-331	3-p	GCCCCTGGGCCTATCCTAGAA	1129	434	1.21E-53	2.60
mir-889	3-p	TTAATATCGGACAACCATTGT	0	170	2.09E-53	n/a
mir-210	3-p	CTGTGCGTGTGACAGCGGCTGA	28	247	1.54E-51	8.82
mir-302d	5-p	ACTTTAACATGGAGGCACTTGCT	258	20	8.58E-48	12.90
mir-143	3-p	TGAGATGAAGCACTGTAGCTC	314	680	7.06E-43	2.17
mir-423	3-p	AGCTCGGTCTGAGGCCCCTCAGT	1225	560	3.17E-40	2.19
mir-503	5-p	TAGCAGCGGGAACAGTTCTG	235	30	2.17E-35	7.83
mir-25	5-p	AGGCGGAGACTTGGGCAATTGCT	1002	456	1.44E-33	2.20
mir-17	3-p	ACTGCAGTGAAGGCACTTGTAG	146	377	1.27E-31	2.58
let-7b	5-p	TGAGGTAGTAGGTTGTGT	284	57	1.38E-31	4.98
mir-335	5-p	TCAAGAGCAATAACGAAAAATG	119	331	1.19E-30	2.78
mir-27b	3-p	TTCACAGTGGCTAAGTTCTGC	195	442	1.38E-30	2.27
mir-129	3-p	AAGCCCTTACCCCAAAAAGCAT	946	449	1.22E-28	2.11
mir-30e	3-p	CTTTCAGTCGGATGTTTACAG	1741	994	1.72E-28	1.75
mir-371	3-p	AAGTGCCGCCATCTTTTGAGTGT	9	114	1.31E-27	12.67
mir-19b	3-p	TGTGCAAATCCATGCAAAACTGA	492	794	7.29E-27	1.61
mir-30e	5-p	TGTAAACATCCTTGACTGGAAGCT	447	738	1.34E-26	1.65
let-7g	5-p	TGAGGTAGTAGTTTGTACAGTT	321	100	1.69E-22	3.21
mir-134	5-p	TGTGACTGGTTGACCAGAGGGG	10	97	7.80E-22	9.70
mir-181d	5-p	AACATTCATTGTTGTCGGTGGGTT	281	82	1.38E-21	3.43
mir-20a	5-p	TAAAGTGCTTATAGTGCAGGTAG	377	608	6.56E-21	1.61
mir-92b	5-p	AGGGACGGGACGCGGTGCAGTGTT	556	250	7.10E-20	2.22
mir-877	5-p	GTAGAGGAGATGGCGCAGGGG	101	10	1.50E-19	10.10
let-7f	5-p	TGAGGTAGTAGATTGTATAGTT	2004	1281	1.60E-19	1.56
mir-483	5-p	AAGACGGGAGGAAAGAAGGGAG	31	133	1.81E-19	4.29
mir-219	3-p	AGAGTTGAGTCTGGACGTCCCG	3	67	6.98E-19	22.33
mir-23b	3-p	ATCACATTGCCAGGGATTACCAC	134	286	1.57E-18	2.13
mir-93	5-p	CAAAGTGCTGTTCGTGCAGGTA	319	556	5.26E-18	1.74
mir-660	5-p	TACCCATTGCATATCGGAGTTG	46	150	2.41E-17	3.26
mir-801	5-p	TGGCACACGTAGGGCAAC	90	185	7.04E-17	2.06
let-7d	5-p	AGAGGTAGTAGGTTGCATAGTT	77	4	7.19E-17	19.25
mir-518f	3-p	GAAAGCGCTTCTCTTTAGAGGA	84	205	1.20E-16	2.44
mir-296	3-p	AGGGTTGGGTGGAGGCT	97	11	2.74E-16	8.82
mir-192	5-p	TGACCTATGAATTGACAGCCAGT	79	210	3.14E-16	2.66
mir-599	5-p	TTTGATAAGCTGACATGGGACAG	0	48	1.32E-15	n/a
mir-145	5-p	GTCCAGTTTTCCCAGGAATCCCT	57	146	4.79E-13	2.56
mir-873	5-p	GCAGGAACTTGTGAGTCTCCT	203	358	1.75E-12	1.76
mir-486	3-p	TCCTGTACTGAGCTGCCCCGAG	293	125	1.91E-12	2.34
mir-329	3-p	AACACACCTGGTTAACCTCTTT	0	35	2.74E-12	n/a
mir-20b	3-p	ACTGTAGTATGGGCACTTCCAGT	205	331	4.18E-12	1.61
mir-125a	5-p	TCCCTGAGACCCTTTAACCTGTG	189	311	5.35E-12	1.65
mir-302c	5-p	TTTAACATGGGGGTACCTGCT	522	280	9.08E-12	1.86
mir-487b	3-p	AATCGTACAGGGTCATCCACTT	0	31	5.70E-11	n/a
mir-769	5-p	TGAGACCTCTGGGTTCTGAGCT	170	298	1.78E-10	1.75
mir-28	3-p	CACTAGATTGTGAGCTCCTGGA	344	169	1.89E-10	2.04
mir-106b	3-p	CCGCACTGTGGGTACTTGCTGC	197	77	2.58E-10	2.56
mir-135b	5-p	TATGGCTTTTCATTCCTATGTGA	30	54	2.67E-10	1.80
mir-598	3-p	TACGTCATCGTTGTCATCGTCA	168	271	3.79E-10	1.61

microRNA (hsa)	arm	most common sequence	hESC count	EB count	P-value	fold change
mir-187	3-p	TCGTGTCTTGTGTTGCAGCCGG	281	130	4.07E-10	2.16
mir-27a	3-p	TTCACAGTGGCTAAGTTC	77	157	4.45E-10	2.04
mir-199b	5-p	CCCAGTGTTTAGACTATCTGTTC	1	30	4.97E-10	30.00
mir-135a	5-p	TATGGCTTTTTATTCCTATGTGA	59	132	5.44E-10	2.24
mir-323	3-p	GCACATTACACGGTCGACCTCT	0	28	5.55E-10	n/a
mir-218	5-p	TTGTGCTTGATCTAACCATGT	1	28	5.55E-10	28.00
mir-96	5-p	TTTGGCACTAGCACATTTTTGCT	44	110	6.99E-10	2.50
mir-185	5-p	TGGAGAGAAAGGCAGTTCCTG	500	311	1.11E-09	1.61
mir-151	5-p	CTAGACTGAAGCTCCTTGAGGA	1179	786	1.16E-09	1.50
mir-539	3-p	ATCATACAAGGACAATTTCTTT	0	27	1.19E-09	n/a
mir-379	5-p	TGGTAGACTATGGAACGTAGG	7	46	1.58E-09	6.57
mir-215	5-p	ATGACCTATGAATTGACAGACA	3	38	2.70E-09	12.67
mir-135a	3-p	ATATAGGGATTGGAGCCGTGGC	15	65	3.02E-09	4.33
mir-99a	5-p	AACCCGTAGATCCGATCTTGTG	4	30	4.03E-09	7.50
mir-768	5-p	GTTGGAGGATGAAAGTACGGA	69	151	4.21E-09	2.19
mir-518b	3-p	CAAAGCGCTCCCCTTTAGAGGT	215	94	4.64E-09	2.29
mir-494	3-p	TGAAACATACACGGGAAACCTCT	2	32	4.73E-09	16.00
mir-330	3-p	GCAAAGCACACGGCCTGCAGAGAG	187	86	5.09E-09	2.17
mir-410	3-p	AATATAACACAGATGGCCTGT	0	25	5.41E-09	n/a
mir-941	3-p	CACCCGGCTGTGTGCACATGTGC	114	41	1.12E-08	2.78
mir-594	3-p	GCCCCAGTGGCCTAATGGA	9	46	2.00E-08	5.11
mir-193b	3-p	AACTGGCCCTCAAAGTCCCGCT	288	146	2.74E-08	1.97
mir-367	3-p	TGGATTGTTAAGCCAATGACAGAA	45	5	2.89E-08	9.00
mir-382	5-p	GAAGTTGTTCGTGGTGGATTCG	0	24	3.59E-08	n/a
mir-335	3-p	TTTTTCATTATTGCTCCTGACC	72	137	5.41E-08	1.90
mir-29c	3-p	TAGCACCATTTGAAATCGGTTA	45	100	7.98E-08	2.22
mir-296	5-p	AGGGCCCCCCCTCAATCCTGT	60	12	9.65E-08	5.00
mir-302b	3-p	ACTTTAACATGGAAGTGCTTTCT	343	188	1.07E-07	1.82
mir-518a	3-p	AGAAGATCTCAAGCTGTGA	75	20	2.22E-07	3.75
mir-224	5-p	CAAGTCACTAGTGGTTCCGTTTAG	108	43	3.19E-07	2.51
mir-520g	3-p	ACAAAGTGCTTCCCTTTAGAGTGT	74	20	3.33E-07	3.70
mir-519a	3-p	AGAAGATCTCAGGCTGTGTC	65	16	4.78E-07	4.06
mir-22	3-p	AAGCTGCCAGTTGAAGAACTGT	359	205	5.20E-07	1.75
mir-9	3-p	TCTTTGGTTATCTAGCTGTATGA	427	256	8.66E-07	1.67
mir-221	5-p	ACCTGGCATACAATGTAGATTTCT	637	413	1.04E-06	1.54
mir-574	3-p	CACGCTCATGCACACACCCACA	65	17	1.12E-06	3.82
mir-543	3-p	AAACATTCGCGGTGCACTTCTT	2	25	1.86E-06	12.50
mir-140	3-p	ACCACAGGGTAGAACCAC	137	219	2.18E-06	1.60
mir-10b	5-p	TACCCTGTAGAACCGAATTTGT	2	23	2.43E-06	11.50
mir-455	5-p	TATGTGCCTTTGGACTACATCG	36	78	3.55E-06	2.17
mir-518c	5-p	CTCTGGAGGGAAGCACTTTCTGTT	43	8	3.69E-06	5.38
mir-197	3-p	TTCACCACCTTCTCCACCCAGC	145	66	4.75E-06	2.20
mir-760	3-p	CGGCTCTGGGTCTGTGGGG	25	2	4.90E-06	12.50
mir-516	5-p	CATCTGGAGGTAAGAAGCACTTTGT	170	92	6.14E-06	1.85
mir-486	3-p	CGGGGCAGCTCAGTACAGGATA	19	0	6.54E-06	n/a
mir-129	5-p	CTTTTTGCGGTCTGGGCTTGC	126	55	7.28E-06	2.29
mir-374b	5-p	ATATAATACAACCTGCTAAG	84	146	1.05E-05	1.74
mir-154	3-p	AATCATACACGGTTGACCTATT	0	15	1.07E-05	n/a
mir-484	5-p	TCAGGCTCAGTCCCCTCCCGAT	65	20	1.10E-05	3.25
mir-187	5-p	GCTACAACACAGGACCCGGGCG	18	0	1.23E-05	n/a

microRNA (hsa)	arm	most common sequence	hESC count	EB count	P-value	fold change
mir-211	3-p	TTCCCTTTGTCATCCTTCGCCT	28	3	1.27E-05	9.33
mir-577	5-p	GTAGATAAAATATTGGTACCTG	53	96	1.64E-05	1.81
mir-369	3-p	AATAATACATGGTTGATCTTT	2	20	1.86E-05	10.00
mir-382	3-p	AATCATTCACGGACAACACTTT	2	20	1.86E-05	10.00
mir-200a	3-p	TAACACTGTCTGGTAACGATGTT	87	136	2.28E-05	1.56
mir-485	3-p	GTCATACACGGCTCTCCTCTCT	0	14	2.28E-05	n/a
mir-766	5-p	AGGAGGAATTGGTGCTGGTCTT	25	3	2.41E-05	8.33
mir-7	3-p	AGGTAGACTGGGATTTGTTGTT	41	9	2.58E-05	4.56
mir-483	3-p	TCACTCCTCTCCTCCCGTCTT	59	101	3.59E-05	1.71
mir-520a	3-p	AAAGTGCTTCCCTTTGGACTG	363	226	3.64E-05	1.61
mir-589	5-p	TGAGAACCACGTCTGCTCTGAGC	62	23	4.20E-05	2.70
mir-519c	3-p	AGAAGATCTCAGCCTGTGAC	44	11	4.35E-05	4.00
mir-374	5-p	TTATAATACAACCTGATAAGT	59	108	4.81E-05	1.83
mir-432	5-p	TCTTGGAGTAGGTCATTGGG	0	13	4.87E-05	n/a
mir-217	5-p	TACTGCATCAGGAACTGATTGGA	1	16	5.02E-05	16.00
mir-31	5-p	AGGCAAGATGCTGGCATAGCTG	235	135	6.16E-05	1.74
mir-196b	5-p	TAGGTAGTTTCCTGTTGTTGGG	2	18	7.11E-05	9.00
mir-362	5-p	AATCCTTGGAACCTAGGTGTGAGT	14	66	8.78E-05	4.71
mir-424	3-p	CAGCAGCAATTCATGTTTTGAA	52	16	8.78E-05	3.25
mir-493	5-p	TTGTACATGGTAGGCTTTCATT	0	13	9.19E-05	n/a
mir-512	3-p	AAGTGCTGTCATAGCTGA	37	76	9.63E-05	2.05
mir-33	5-p	GTGCATTGTAGTTGCATTGCA	43	78	0.000100716	1.81
mir-182	5-p	TTTGGCAATGGTAGAACTCACACTGGT	73	32	0.000129595	2.28
mir-923	3-p	CAGGATTCCCTCAGTAA	13	1	0.000161215	13.00
mir-193b	5-p	CGGGGTTTTGAGGGCGAGATG	19	2	0.000173717	9.50
mir-874	3-p	CTGCCCTGGCCCGAGGGACCGAC	19	2	0.000173717	9.50
mir-95	3-p	TTCAACGGGTATTTATTGAGC	19	2	0.000173717	9.50
mir-542	5-p	TGTGACAGATTGATAACTGAAA	48	82	0.00020581	1.71
mir-24	3-p	TGGCTCAGTTCAGCAGGAACA	47	87	0.000216923	1.85
mir-499	5-p	TTAAGACTTGCAGTGATGTTTA	30	59	0.000231403	1.97
mir-887	3-p	GTGAACGGGCGCCATCCCGAGGCT	3	19	0.000343537	6.33
mir-155	5-p	TTAATGCTAATCGTGATAGGGGT	35	9	0.000348001	3.89
mir-185	3-p	AGGGGCTGGCTTTCCTCT	35	9	0.000348001	3.89
mir-18b	5-p	TAAGGTGCATCTAGTGCAGT	0	11	0.000382853	n/a
mir-375	3-p	TTTGTTCGTTCGGCTCGC	19	46	0.000415915	2.42
mir-203	3-p	GTGAAATGTTTAGGACCACTAG	52	85	0.000459785	1.63
mir-495	3-p	AAACAAACATGGTGCACTTCTT	0	10	0.000474427	n/a
mir-339	5-p	TGAGCGCCTCGACGACAGAGCCG	29	55	0.000616533	1.90
mir-21	3-p	CAACACCAGTCGATGGGCTGTCT	33	10	0.000631908	3.30
mir-618	3-p	AAACTCTACTTGTCCTTCTGAGT	9	27	0.000678534	3.00
mir-28	5-p	AAGGAGCTCACAGTCTATTGAGT	21	4	0.000767493	5.25
mir-520b	3-p	AAAGTGCTTCCTTTTAGAG	0	10	0.000781386	n/a
mir-641	5-p	AAAGACATAGGATAGAGTCACCT	0	10	0.000781386	n/a
mir-199a	5-p	CCCAGTGTTCAGACTACCTGTTC	1	10	0.000781386	10.00
mir-301	5-p	CAGTGCAATAGTATTGTCAAAGCAT	16	37	0.000792328	2.31
mir-504	5-p	AGACCCTGGTCTGCACTCTATCT	49	20	0.000811599	2.45
mir-372	5-p	CCTCAAATGTGGAGCACTATTC	1	12	0.000823076	12.00
mir-194	3-p	TGTAACAGCAACTCCATGTGGAA	12	31	0.000885533	2.58
mir-518a	5-p	CTGCAAAGGGAAGCCCTTTCT	2	14	0.000978556	7.00

Supplementary Table 2: All novel miRNAs reported in this study

Chromo-some	Start	End	Most abundant isomiR	hESC count	EB count	Fold Change (if significant)
chr2	177173995	177174018	AATGGATTTTTGGAGCAGG	374	245	1.53
chr11	7212576	7212600	TAAGTGCTTCCATGCTT	84	131	1.56
chrX	113855921	113855947	TTCATTCGGCTGTCCAGATGTA	1269	774	1.64
chr16	33873061	33873083	GCCTGTCTGAGCGTCGCTT	194	114	1.70
chr1	218440512	218440575	TATTCATTTATCCCCAGCCTACA	126	67	1.88
chr19	62716221	62716246	TCCCTGTTCGGGCGCCA	670	1302	1.94
chr6	131417429	131417451	TCCCTGTTCGGGCGCCA	670	1302	1.94
chr1	19096156	19096178	TGGATTTTTGGATCAGGGA	40	85	2.13
chr1	555992	556022	AACAGCTAAGGACTGCA	72	31	2.32
chr6	26645776	26645799	GTGTAAGCAGGGTCGTTTT	639	273	2.34
chrX	117404429	117404454	TACGTAGATATATATGTATTTT	20	54	2.70
chr5	41511501	41511523	GTCCCTGTTCAGGCGCCA	24	66	2.75
chr15	62841721	62841750	GATGATGATGGCAGCAAATTCTGAAA	81	27	3.00
chr22	18453595	18453657	ACGTTGGCTCTGGTGGTG	35	11	3.18
chr8	141129908	141129930	ATCCCCAGCACCTCCACC	59	18	3.28
chr10	100145015	100145040	TGCTGGATCAGTGGTTCGAGTC	16	57	3.56
chr15	50356653	50356679	CCTCAGGGCTGTAGAACAGGGCT	163	45	3.62
chr4	66825201	66825227	CTGGACTGAGCCGTGCTACTGG	44	161	3.66
chr10	105144044	105144071	ACTCGGCGTGGCGTCGGTCGTG	228	50	4.56
chr2	25405021	25405049	TTGCAGCTGCCTGGGAGTGACTTC	63	13	4.85
chr4	164234207	164234230	TCCGAGTCACGGCACCA	309	55	5.62
chr1	150066842	150066862	GTGGGGGAGAGGCTGTC	103	17	6.06
chr1	68421844	68421869	ATGGGTGAATTTGTAGAAGGAT	7	45	6.43
chr7	146706060	146706085	AAACTGTAATTACTTTTGG	21	3	7.00
chrX	21990206	21990231	GCATGGGTGGTTCAGTGG	407	52	7.83
chr10	70189095	70189123	AGCCTGGAAGCTGGAGCCTGCAGT	25	3	8.33
chr22	35428858	35428879	AAAAAGACACCCCCCAC	189	21	9.00
chr14	101096450	101096475	ACCCGTCCCGTTCGTCCCCGGA	4	53	13.25
chr11	90241994	90242016	ATGGATAAGGCTTTGGCTT	31	2	15.50
chr15	20014621	20014642	CGGGCGTGGTGGTGGGGG	35	2	17.50
chr11	93106277	93106306	GTGGGAAGGAATTACAAGACAGTT	72	2	36.00
chr1	20299	20324	TTGGGACATACTTATGCTAAA	4	1	-
chr1	21187455	21187480	AGGCATTGACTTCTCACTAGCT	2	2	-
chr1	38242469	38242487	AGAGGAACAGTTCATTC	10	2	-
chr1	86844553	86844578	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr1	94984022	94984045	ACTGGGCTTGGAGTCAGAAG	98	130	-
chr1	148091054	148091079	GCGAGATCGCGCAGGACTTT	3	0	-

Chromo-some	Start	End	Most abundant isomiR	hESC count	EB count	Fold Change (if significant)
chr1	166234525	166234550	CGGATGAGCAAAGAAAGTGGTT	7	2	-
chr1	169337500	169337525	TTAGGCCGCAGATCTGGGTGA	4	6	-
chr1	191372302	191372327	TAGTACTGTGCATATCATCTAT	184	126	-
chr1	222511371	222511397	AAAAGCTGGGTTGAGAGGGCAA	20	13	-
chr10	14518592	14518617	CAGGATGTGGTCAAGTGTTGTT	9	2	-
chr10	56037643	56037672	AAAACTGTAATTACTTTTGGAC	64	31	-
chr10	64802775	64802800	TTAGGGCCCTGGCTCCATCTCC	13	15	-
chr10	93405230	93405250	TGAACCCAGGAGGCGGA	6	9	-
chr10	104030627	104030644	AGATGGATTGTTCTGGG	2	1	-
chr10	112738724	112738749	AAAAACTGAGACTACTTTTGCA	38	26	-
chr11	3535184	3535209	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr11	8662350	8662379	ATCGAGGCTAGAGTCACGCTTGG	8	10	-
chr11	8662888	8662951	TGTGGACTTTGGTCAGTGAAC	2	1	-
chr11	61491650	61491669	GCCGAGAGTCGTCGGGGTT	3	1	-
chr11	67457297	67457322	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr11	69807738	69807763	AAAAGTACTTGCGGATTTTGCT	57	64	-
chr11	93106536	93106561	TTTGAGGCTACAGTGAGATGTG	2	2	-
chr11	93839357	93839382	AAAAGTATTTGCGGGTTTTGTC	14	14	-
chr12	9371838	9371863	TTGGGACATACTTATGCTAAA	4	1	-
chr12	43950119	43950145	AGTGAACCACAGACCTGAGATGT	3	3	-
chr12	47334539	47334568	TGGCCCTGACTGAAGACCAGCAGT	6	0	-
chr12	48914229	48914253	TGGGTGGTCTGGAGATTTGTGC	4	0	-
chr12	61368489	61368514	AAAAACTGTAATTACTTTT	4	1	-
chr12	67953235	67953252	TCATATTGCTTCTTTCT	5	2	-
chr12	78337170	78337195	AGAAGGAAATTGAATTCATTTA	4	5	-
chr12	96409820	96409846	ACTCTAGCTGCCAAAGGCGCT	2	9	-
chr12	97517768	97517795	GACTCGACTCGTGTGCGGACATTT	7	2	-
chr12	111617267	111617292	TTGGGACATACTTATGCTAAA	4	1	-
chr13	44500861	44500886	TCTGGGCAACAAAGTGAGACCT	23	33	-
chr13	53784117	53784134	TTCAAGTAATTCAGGTG	4	6	-
chr13	70373329	70373354	AAAAACTGTAATTACTTTT	4	1	-
chr13	106981564	106981588	CCTGTTGAAGTGTAATCCCCA	2	3	-
chr14	22317565	22317585	ATATAATACAACCTGCTT	45	74	-
chr14	63631551	63631577	AAAAGTAATCGCGGTTTTTGTC	27	16	-
chr14	76802325	76802346	ATCCCACCTCTGCCACCA	4	2	-
chr15	41873222	41873242	TCGTTTGCCTTTTTCTGCTT	1	2	-
chr15	51017861	51017887	TTGAGAAGGAGGCTGCTG	0	3	-
chr15	66267369	66267386	AAAAGAAATGTACAGAG	1	4	-
chr15	84114778	84114803	TAAAGAGCCCTGTGGAGACA	3	9	-

Chromo-some	Start	End	Most abundant isomiR	hESC count	EB count	Fold Change (if significant)
chr15	100318227	100318252	TTGGGACATACTTATGCTAAA	4	1	-
chr16	1952199	1952219	TCTGGAAGGAGACTGTAT	0	1	-
chr16	11307845	11307871	AAAAGTAATCGCGGTTTTTGTC	27	16	-
chr17	2598181	2598205	AGAGAAGAAGATCAGCCTGCA	4	2	-
chr17	13387635	13387661	AAAAGTAATCGCGGTTTTTGTC	27	16	-
chr17	16126095	16126119	TGGACTGCCCTGATCTGGAGA	3	2	-
chr17	76721656	76721681	ACGGTGCTGGATGTGGCCTTT	6	8	-
chr18	26132900	26132917	TAATTGCTTCCATGTTT	60	58	-
chr18	55567933	55567958	AAAAACTGTAATTACTTTT	4	1	-
chr19	23043	23068	TTGGGACATACTTATGCTAAA	4	1	-
chr19	20371125	20371152	CTGGAGATATGGAAGAGCTGTGT	76	38	-
chr19	58883558	58883583	TCTACAAAGGAAAGCGCTTTCT	12	10	-
chr19	58953309	58953334	TCTACAAAGGAAAGCGCTTTCT	12	10	-
chr2	56081400	56081425	AAATCTCTGCAGGCAAATGTG	0	4	-
chr2	62828390	62828415	TAGCAAAAACTGCAGTTACTTT	6	2	-
chr2	70333567	70333592	TCTGGGCAACAAAGTGAGACCT	23	33	-
chr2	114057048	114057073	TTGGGACATACTTATGCTAAA	4	1	-
chr2	180433815	180433840	AGTTAGGATTAGGTCGTGGAA	10	5	-
chr2	180433850	180433875	AGTTAGGATTAGGTCGTGGAA	10	5	-
chr2	189551103	189551128	AAGTGATCTAAAGGCCTACAT	0	6	-
chr2	207842293	207842318	TTGGGACATACTTATGCTAAA	4	1	-
chr2	212999243	212999270	AACTGTAATTACTTTTGCACCAAC	2	0	-
chr2	232286322	232286348	AAGTAGTTGGTTTGTATGAGATGGTT	0	1	-
chr20	2581424	2581452	TGGGAACGGGTTCCGGCAGACGCTG	4	5	-
chr20	33505224	33505250	TGGAGTCCAGGAATCTGCATTTT	2	2	-
chr20	36487253	36487280	AAACTCTGGAGCTTTCGTACATGC	7	0	-
chr20	36578662	36578687	CCAAAACTGCAGTTACTTTTGC	8	2	-
chr20	47330266	47330296	ATATATGATGACTTAGCTTTT	7	22	-
chr20	59962070	59962094	AGTGAATGATGGGTTCTGACC	2	3	-
chr22	18616665	18616690	TGCAGGACCAAGATGAGCCCT	13	14	-
chr22	20337277	20337302	GCTCTGACGAGGTTGCACTACT	7	6	-
chr22	20337312	20337340	CAGTGCAATGATATTGTCAAAGCAT	24	30	-
chr22	25281235	25281262	AAAAGTAATTGCGGTCTTTGGT	82	58	-
chr22	39818495	39818513	TCGCCTCCTCCTCTCCC	0	1	-
chr22	43975501	43975526	ACGCCCTTCCCCCCCTTCTTCA	16	3	-
chr3	70572562	70572587	AAAAACTGTAATTACTTTT	4	1	-
chr3	71673877	71673902	TCTATACAGACCCTGGCTTTTC	2	6	-
chr3	119201514	119201540	TGGAGTCCAGGAATCTGCATTTT	2	2	-
chr3	126992024	126992049	AAAAGTAATTGCGGATTTTGCC	2	3	-

Chromo-some	Start	End	Most abundant isomiR	hESC count	EB count	Fold Change (if significant)
chr3	129563700	129563720	TCCCACCGCTGCCACCC	6	5	-
chr3	150090065	150090092	GGCCAGGCCTGGTGGCTCACTTT	6	1	-
chr3	165372000	165372027	ATGGTACCCTGGCATACTGAGT	5	3	-
chr3	187985862	187985882	GAAAACTGCAGTGACATGTT	0	2	-
chr3	187987156	187987186	ACCTTCTTGTATAAGCACTGTGCTAAA	5	1	-
chr4	4136998	4137023	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr4	9166974	9166999	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr4	36104418	36104443	CGGATGAGCAAAGAAAGTGGTT	7	2	-
chr4	71081401	71081426	TTGGGACATACTTATGCTAAA	4	1	-
chr4	102470541	102470568	AGGATGAGCAAAGAAAGTAGATT	111	70	-
chr4	114247470	114247495	AACTGGATCAATTATAGGAGTG	3	2	-
chr4	148485239	148485266	AAAACTGTAATTACTTTTGTAC	21	12	-
chr4	183327488	183327513	TTTTCAACTCTAATGGGAGAGA	14	1	-
chr5	100180098	100180123	TAGCAAAAACTGCAGTTACTTT	6	2	-
chr5	109877434	109877461	AACTGTAATTACTTTTGCACCAAC	2	0	-
chr5	132791203	132791229	TGGAGTCCAGGAATCTGCATTTT	2	2	-
chr5	153706904	153706929	TGTGAGGTTGGCATTGTTGTCT	12	12	-
chr5	154045576	154045602	TTTAGAGACGGGGTCTTGCTCT	19	8	-
chr5	175727567	175727592	CTTGGCACCTAGCAAGCACTCA	20	25	-
chr6	54084910	54084927	AACAGCATTGCACTTGT	1	3	-
chr6	74286439	74286468	AAAGGAAAAGACTCATATCAA	3	0	-
chr6	132155005	132155031	AAAAGTAATCGCGGTTTTTGTC	27	16	-
chr6	132634811	132634835	TGGGAAAGGAAAAGACTC	4	2	-
chr6	133180049	133180074	AAGCCAGCCAATGAA	4	0	-
chr6	133180152	133180179	TAAGGTCCCTGAGAATGGCTAT	38	24	-
chr6	163910779	163910796	AAATACAATAAAATCTG	0	2	-
chr6	167331306	167331331	TACGCGCAGACCACAGGATGTC	4	3	-
chr7	18133379	18133404	TTGGGACATACTTATGCTAAA	4	1	-
chr7	34946939	34946964	CAAAAGTAATTGTGGATTTTGT	3	2	-
chr7	67622028	67622053	TCTGGGCAACAAAGTGAGACCT	23	33	-
chr7	91671271	91671298	TCTGGGCAACAAAGTGAGACCT	23	33	-
chr7	100156228	100156248	GTGGGGGAGAGGCTGTG	18	9	-
chr8	7043408	7043433	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr8	7983960	7983985	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr8	12479920	12479945	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr8	12538020	12538045	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr8	26962356	26962382	AAAAGTAATCGCGGTTTTTGTC	27	16	-
chr8	101105387	101105415	GGGCGACAAAGCAAGACTCTTTCTT	4	0	-
chr8	142865520	142865545	TTGGGACATACTTATGCTAAA	4	1	-

Chromo-some	Start	End	Most abundant isomiR	hESC count	EB count	Fold Change (if significant)
chr8	144193152	144193177	TTGGGACATACTTATGCTAAA	4	1	-
chr9	20214	20239	TTGGGACATACTTATGCTAAA	4	1	-
chr9	28878884	28878910	GAGACTGATGAGTTCCCGGGA	41	23	-
chr9	68292057	68292082	TTCTGGAATTCTGTGTGAGGGA	2	8	-
chr9	91868406	91868431	AAAAGTAATTGCGGATTTTGCC	2	3	-
chr9	99165696	99165721	TTGGGACATACTTATGCTAAA	4	1	-
chr9	116078285	116078309	AATGGCTTTTTGGAGCAGG	149	103	-
chrX	32569518	32569545	AAAACTGTAATTACTTTTGTAC	21	12	-
chrX	69159472	69159498	CTGTCCTAAGGTTGTTGAGTT	28	15	-
chrX	83367459	83367484	AAAAGTAATTGCGGATTTTGCC	2	3	-
chrX	93354756	93354781	AAAAGTAATTGCGGATTTTGCC	2	3	-
chrX	94204843	94204867	CAAAGGTATTTGTGGTTTTTG	2	0	-
chrX	100287314	100287339	TAGCAAAAACTGCAGTTACTTT	6	2	-
chrX	113793427	113793450	CAAGTCTTATTTGAGCACCTGTT	2	2	-
chrX	118805342	118805371	CATGACAGATTGACATGGACAATT	7	0	-
chrY	13478002	13478025	AGATGAACTTGAAAGAAGACCAT	2	1	-

Supplementary Table 3: Seed sequences differentially represented between libraries

Seed sequence	hESC count	EB count	hESC background	EB background	Fold Change	Corrected P-value
AGUAGUC	86	1478	860319	826472	17.19	0
CAGUAGU	1801	22288	860319	826472	12.38	0
ACAGUAG	241	2749	860319	826472	11.41	0
UUAAACG	3452	398	860319	826472	8.67	0
CUUAAAC	16762	2332	860319	826472	7.19	0
GGAGUGU	846	5216	860319	826472	6.17	0
AGUGCUG	2484	12641	860319	826472	5.09	0
ACCCUGU	1503	5689	860319	826472	3.79	0
AAACGUG	55712	16461	860319	826472	3.38	0
UAUAAAG	3002	9702	860319	826472	3.23	0
GAGAACU	1547	4976	860319	826472	3.22	0
UUGCACG	1845	5561	860319	826472	3.01	0
GAGGUAG	29280	10258	860319	826472	2.85	0
CAGUGCA	3596	8559	860319	826472	2.38	0
AAAGCUG	16448	7613	860319	826472	2.16	0
AGUGCAA	4463	9594	860319	826472	2.15	0
UCAAGUA	6898	14089	860319	826472	2.04	0
GCUACAU	47591	24280	860319	826472	1.96	0
GUAAACA	8789	15479	860319	826472	1.76	0
GAGGGGC	17586	10518	860319	826472	1.67	0
AGCUUAU	65376	39424	860319	826472	1.66	0
CUGGACU	17819	25378	860319	826472	1.42	0
GCAGCAU	153193	187218	860319	826472	1.22	0
GCGGGGC	5220	1956	860319	826472	2.67	8.76E-305
CUCAAAC	4306	8173	860319	826472	1.90	1.62E-296
UGGAUAA	1770	233	860319	826472	7.60	1.00E-275
GGAAUGU	9601	5384	860319	826472	1.78	4.47E-223
AAAGUGC	8585	12847	860319	826472	1.50	8.91E-221
UAAGCCA	2225	550	860319	826472	4.05	7.95E-220
GGUUGGG	1037	199	860319	826472	5.21	8.95E-127
UCAACAG	1003	2283	860319	826472	2.28	1.83E-121
UGUGCGU	129	747	860319	826472	5.79	5.50E-109
GGACGGA	616	69	860319	826472	8.93	4.40E-103
CCCCUGG	1910	773	860319	826472	2.47	3.31E-97
UCCCUUU	837	1862	860319	826472	2.22	9.74E-95
AGGAGCU	705	1649	860319	826472	2.34	5.90E-92
UAAACGU	759	165	860319	826472	4.60	1.60E-83
UCACAGU	719	1573	860319	826472	2.19	2.42E-77
CGGAUCC	115	578	860319	826472	5.03	2.41E-76
GCUCGGU	1793	817	860319	826472	2.19	8.24E-72
AUGGAUA	322	16	860319	826472	20.13	1.18E-69
UAAUAUC	0	228	860319	826472	n/a	2.34E-68
GAGAUGA	412	1036	860319	826472	2.51	8.38E-65
CACAGUG	6807	8616	860319	826472	1.27	9.60E-62
CUUGGAG	0	193	860319	826472	n/a	1.63E-57
AGUGCUU	8315	6098	860319	826472	1.36	8.08E-55

Seed sequence	hESC count	EB count	hESC background	EB background	Fold Change	Corrected P-value
CUUUAAC	860	315	860319	826472	2.73	3.22E-51
GGCGGAG	1303	602	860319	826472	2.16	3.00E-50
UUUCAGU	4065	2680	860319	826472	1.52	1.72E-49
AGCCCUU	1311	612	860319	826472	2.14	2.02E-49
AUGGCUU	350	834	860319	826472	2.38	2.40E-47
UAGAGGA	277	29	860319	826472	9.55	5.51E-47
CUGCAGU	340	815	860319	826472	2.40	8.22E-47
AACGGAA	6891	5034	860319	826472	1.37	1.63E-46
GGGACGG	1059	472	860319	826472	2.24	4.81E-44
CAAGAGC	240	643	860319	826472	2.68	6.25E-44
ACAGUAC	2763	3740	860319	826472	1.35	7.14E-40
AGCAGCG	548	175	860319	826472	3.13	1.87E-39
UUAACAU	1179	579	860319	826472	2.04	3.85E-39
UUGAUAA	0	127	860319	826472	n/a	4.54E-37
GGAUUGU	216	22	860319	826472	9.82	1.01E-36
AGUGCCG	19	194	860319	826472	10.21	2.94E-36
UGCCGCC	11	164	860319	826472	14.91	9.64E-35
GACCUAU	545	1017	860319	826472	1.87	1.72E-34
AGACGGG	56	273	860319	826472	n/a	1.73E-34
CCUUCAU	1246	660	860319	826472	1.89	1.15E-33
ACCCAUU	106	350	860319	826472	3.30	7.40E-31
AAGCUGG	810	374	860319	826472	2.17	9.93E-31
GUGACUG	16	160	860319	826472	10.00	1.61E-29
CCUGUAC	529	200	860319	826472	2.65	1.94E-29
CCCUGUA	40	216	860319	826472	5.40	3.77E-29
GAAGAUC	394	124	860319	826472	3.18	1.50E-28
ACAUUCA	3751	4644	860319	826472	1.24	6.63E-28
UGACCUA	2439	3150	860319	826472	1.29	4.08E-25
GUGCAAA	755	1198	860319	826472	1.59	8.80E-25
CAAAGCA	1176	678	860319	826472	1.73	4.92E-24
CGCACUG	490	201	860319	826472	2.44	2.68E-23
CAGGAAC	531	902	860319	826472	1.70	3.38E-23
GAGUUGA	3	93	860319	826472	31.00	3.11E-22
UCCAGUU	85	269	860319	826472	3.16	3.65E-22
AGCUCGG	700	352	860319	826472	1.99	2.77E-21
AACGUGG	221	53	860319	826472	4.17	2.77E-21
UAUAAUA	3711	4467	860319	826472	1.20	3.20E-21
CCAGUGU	0	73	860319	826472	n/a	2.44E-20
UGGCACA	93	269	860319	826472	2.89	1.30E-19
ACUAGAU	514	238	860319	826472	2.16	6.75E-19
GGUGGGG	223	61	860319	826472	3.66	1.76E-18
CAAGACU	115	14	860319	826472	8.21	2.78E-17
ACGUCAU	517	812	860319	826472	1.57	1.10E-15
GGGUUGG	184	49	860319	826472	3.76	1.88E-15
AAGACUG	96	10	860319	826472	9.60	3.14E-15
GUGCCGC	11	93	860319	826472	8.45	3.19E-15
UCAUACA	0	54	860319	826472	n/a	1.88E-14
ACACACC	6	76	860319	826472	12.67	2.07E-14
AAUACUG	3266	3839	860319	826472	1.18	2.42E-14
GGCAAGA	585	316	860319	826472	1.85	3.86E-14

Seed sequence	hESC count	EB count	hESC background	EB background	Fold Change	Corrected P-value
GUAGUCU	2	62	860319	826472	31.00	4.59E-14
CGGGGCU	203	67	860319	826472	3.03	6.61E-13
CCAGUGC	52	0	860319	826472	n/a	7.08E-13
GGCAGUG	2741	2084	860319	826472	1.32	7.16E-13
UAGACUG	2781	2119	860319	826472	1.31	8.33E-13
AAUGGAU	74	6	860319	826472	12.33	1.81E-12
CCCCAGU	44	148	860319	826472	3.36	2.06E-12
CUGUAGU	472	722	860319	826472	1.53	2.07E-12
AAAGCAC	285	120	860319	826472	2.38	2.82E-12
GAGACCU	248	440	860319	826472	1.77	3.44E-12
GGGGCAG	156	44	860319	826472	3.55	4.72E-12
CAUGGGU	49	0	860319	826472	n/a	5.41E-12
CGUGUCU	414	211	860319	826472	1.96	1.84E-11
AAGCCAG	116	27	860319	826472	4.30	1.08E-10
UAUCAGA	1019	1328	860319	826472	1.30	2.04E-10
UUUUUGC	232	95	860319	826472	2.44	3.15E-10
CACAUUA	0	40	860319	826472	n/a	4.08E-10
ACGUGGA	130	36	860319	826472	3.61	4.65E-10
UGCACGG	29	109	860319	826472	3.76	5.88E-10
UAAACAU	159	302	860319	826472	1.90	1.27E-09
GGCACAC	117	244	860319	826472	2.09	1.42E-09
UUGGAGG	419	627	860319	826472	1.50	1.45E-09
UAUAGGG	15	79	860319	826472	5.27	2.27E-09
AGCCAUG	40	0	860319	826472	n/a	2.44E-09
UGUGCUU	2	45	860319	826472	22.50	4.22E-09
UAGCUUA	124	36	860319	826472	3.44	6.27E-09
CAGCAUU	949	1226	860319	826472	1.29	7.05E-09
AUCGUAC	0	36	860319	826472	n/a	7.08E-09
CCCUGAG	721	969	860319	826472	1.34	7.20E-09
GGGCCCC	74	12	860319	826472	6.17	1.21E-08
AAGUUGU	2	43	860319	826472	21.50	1.60E-08
UCACCAC	217	95	860319	826472	2.28	4.49E-08
CUACAUC	304	155	860319	826472	1.96	4.71E-08
AUCCUUG	40	119	860319	826472	2.98	4.89E-08
AACAUUC	4	47	860319	826472	11.75	7.50E-08
UGCCCUG	43	2	860319	826472	21.50	8.01E-08
GAAUUGU	21	84	860319	826472	4.00	1.01E-07
GGCUCUG	49	4	860319	826472	12.25	1.01E-07
GUGCAAU	72	164	860319	826472	2.28	2.38E-07
GCUUAUC	520	321	860319	826472	1.62	2.98E-07
UAAUGCU	139	50	860319	826472	2.78	3.23E-07
CAGGGAU	80	18	860319	826472	4.44	3.93E-07
AUUGCAC	137973	135989	860319	826472	1.01	4.83E-07
ACUGGCC	458	278	860319	826472	1.65	1.02E-06
AAAGCGC	654	862	860319	826472	1.32	1.05E-06
GAAACAU	2	36	860319	826472	18.00	1.66E-06
AGCUGCC	442	269	860319	826472	1.64	2.67E-06
GUAAGUG	229	115	860319	826472	1.99	6.19E-06
AAGUCAC	369	217	860319	826472	1.70	6.27E-06
GACCCUG	122	45	860319	826472	2.71	9.34E-06

Seed sequence	hESC count	EB count	hESC background	EB background	Fold Change	Corrected P-value
GGAGGAA	57	10	860319	826472	5.70	9.50E-06
CACUCCU	134	236	860319	826472	1.76	1.18E-05
CUACAUU	1042	765	860319	826472	1.36	1.55E-05
AGCCAGG	27	0	860319	826472	n/a	1.72E-05
AACACUG	105	197	860319	826472	1.88	1.80E-05
AUAUAAC	0	25	860319	826472	n/a	1.81E-05
ACCCGUA	1637	1900	860319	826472	1.16	2.05E-05
ACCCGGC	148	63	860319	826472	2.35	2.48E-05
AAUGCCC	1011	744	860319	826472	1.36	3.31E-05
UUGGCAC	63	137	860319	826472	2.17	3.53E-05
UACUGCA	3	34	860319	826472	11.33	4.30E-05
AAGCUCG	112	42	860319	826472	2.67	4.46E-05
UUUAACA	41	5	860319	826472	8.20	5.62E-05
GGAGAGA	2555	2092	860319	826472	1.22	6.37E-05
AACACCA	144	63	860319	826472	2.29	6.62E-05
CUACAAC	25	0	860319	826472	n/a	6.75E-05
AUCAUAC	0	23	860319	826472	n/a	7.55E-05
UUUUCAU	203	314	860319	826472	1.55	9.19E-05
UAUUGCA	275	156	860319	826472	1.76	0.000100705
AGAACUG	9	47	860319	826472	5.22	0.000111875
ACAUUAC	0	22	860319	826472	n/a	0.000154039
CACACCU	0	22	860319	826472	n/a	0.000154039
CACGGUA	0	22	860319	826472	n/a	0.000154039
AUCAUUC	2	29	860319	826472	14.50	0.000161419
UGAACGG	11	49	860319	826472	4.45	0.00021004
GUGACAG	88	166	860319	826472	1.89	0.000210775
CUUUGGU	737	527	860319	826472	1.40	0.0002156
GAGAACC	206	108	860319	826472	1.91	0.000233133
CCUGGCA	1427	1117	860319	826472	1.28	0.000301431
UGUACAU	0	21	860319	826472	n/a	0.000314382
CACAGGG	333	204	860319	826472	1.63	0.000321823
AAGAUCU	173	86	860319	826472	2.01	0.000390211
CAAGUAA	84	157	860319	826472	1.87	0.000571587
UAAGACU	37	90	860319	826472	2.43	0.000739075
AUGUGCC	96	171	860319	826472	1.78	0.000840298
CUAGAUU	81	28	860319	826472	2.89	0.001084265
ACUGCAU	2	26	860319	826472	13.00	0.001120023
CUGUGCG	2	26	860319	826472	13.00	0.001120023
ACGCUCA	84	30	860319	826472	2.80	0.001186869
UUGUUCG	431	565	860319	826472	1.31	0.001187676
UGGAGGA	26	72	860319	826472	2.77	0.001289031
AGCAUUG	270	163	860319	826472	1.66	0.002278633
AUCUGGA	235	137	860319	826472	1.72	0.002768409
AGACUGG	57	16	860319	826472	3.56	0.002930674
GUAACAG	15	51	860319	826472	3.40	0.003219056
CGAUUGC	19	58	860319	826472	3.05	0.003246499
GGGGCUG	72	25	860319	826472	2.88	0.003559141
CUAAGCC	60	18	860319	826472	3.33	0.003716032
AAAAGCU	125	59	860319	826472	2.12	0.004052592
GGCUCAG	943	1111	860319	826472	1.18	0.004326709

Seed sequence	hESC count	EB count	hESC background	EB background	Fold Change	Corrected P-value
CUCACAC	660	805	860319	826472	1.22	0.005340937
UACCCUG	0	17	860319	826472	n/a	0.005454676
AUAAUAC	14	48	860319	826472	3.43	0.005639196
ACUGCAA	3	26	860319	826472	8.67	0.006144601
ACUCCUC	54	108	860319	826472	2.00	0.006407272
GAAUGUA	105	47	860319	826472	2.23	0.006947813
AAGCACA	80	31	860319	826472	2.58	0.009501052
UCAACGG	181	101	860319	826472	1.79	0.009508369
GAGUGUG	35	79	860319	826472	2.26	0.013463942
AUAAAGC	20	56	860319	826472	2.80	0.015650839
GACUGGG	17	0	860319	826472	n/a	0.016244698
GGGGUUU	61	21	860319	826472	2.90	0.020082178
ACAGCAG	0	15	860319	826472	n/a	0.022720944
AGCACCA	4400	3850	860319	826472	1.14	0.024924325
CUCAAAU	2	21	860319	826472	10.50	0.026958039
UCUUGAG	77	31	860319	826472	2.48	0.027743405
UUCUAAU	64	117	860319	826472	1.83	0.027744697
UAGAUAA	59	110	860319	826472	1.86	0.029032117
CUUCAUG	10	38	860319	826472	3.80	0.031394002
AGGGGCA	190	112	860319	826472	1.70	0.032600265
GUGCUUC	349	237	860319	826472	1.47	0.03499445
UGAAAUG	70	124	860319	826472	1.77	0.037971577
ACUUUAA	70	28	860319	826472	2.50	0.045295024
CAGGCUC	142	77	860319	826472	1.84	0.045618849

Appendix B: Supplementary Methods and accompanying data

Comparing miRNA expression levels using five different metrics

We assessed the use of five different metrics to summarize miRNA expression levels based on the two sequence libraries. The first three metrics are calculated from sequences with either a perfect match, or sequences with perfect 3' or 5' matches to the miRBase reference sequence. The other two were calculated using either the most common or the sum of all sequences representing a single arm of a particular miRNA.

The pairwise correlation between the hESC and EB libraries using each of the five metrics is included in Table A1 (below). The most robust discrimination between these libraries, manifested as a low R^2 value, resulted from comparison using the most frequently observed sequence from each arm of every miRNA (**R^2=0.87, Spearman Correlation**). Using only the sequences matching the miRBase reference sequence gave reasonable, but not as large, a discrimination between libraries (**R^2=0.92**). The 'sum of all tags' metric showed an even lower correlation, suggesting it may provide better discrimination between libraries. However, this metric did not perform well when used for clustering libraries of different origin in both human and mouse (not shown).

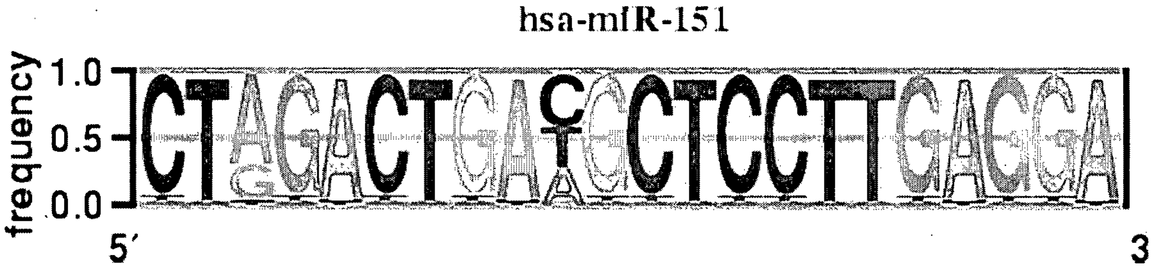
Table A1: Spearman correlation between hESC and EB libraries using each of the five metrics for miRNA expression

	hESC miRBase	hESC most common tag	hESC sum of tags	hESC 5prime miRBase match	hESC 3prime miRBase match
EB miRBase	**0.92**	0.68	0.57	0.85	0.93
EB most common tag	0.76	**0.87**	0.64	0.89	0.85
EB sum of tags	-0.16	-0.09	**0.66**	-0.16	-0.12
EB 5' miRBase match	0.77	0.70	0.29	**0.92**	0.84
EB 3' miRBase match	0.67	0.68	0.66	0.85	**0.95**

Detecting RNA editing events

The two types of RNA editing that have been characterized in mammals include Adenine and Cytosine deamination reactions. The former is observed in sequence reads as an A to G transition while the latter is observed as a C to T transition. To search for evidence of RNA editing, we aligned all unaligned reads to the human genome using a local alignment tool that allows up to two mismatches (ELAND, unpublished). Taking all the reads that aligned to miRNA loci using this tool, we produced multiple sequence alignments representing 'modified' versions of each miRNA. These multiple sequence alignments were converted to sequence logos to facilitate visualization of recurrent editing events. Figure A1 shows the identification of a known A->I editing site in has-miR-151 Blow et al. 2006) as well as a potential C->U editing site. It is also possible that the latter results from minor contamination of the mouse fibroblast feeder cells on which the hESCs were grown, considering the mismatching nucleotide corresponds to the mouse genome sequence.

Figure A1: Sequence logo representing has-miR-151 reads with imperfect genome alignments


hsa-mIR-151
frequency
5′
3′