Published in final edited form as:
Stem Cells. 2008 October ; 26(10): 2496–2505. doi:10.1634/stemcells.2008-0356.

MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries

Merav Bar[1,*], **Stacia K. Wyman**[2,*], **Brian R. Fritz**[2], **Junlin Qi**[3,4], **Kavita S. Garg**[2,6], **Rachael K. Parkin**[2], **Evan M. Kroh**[2], **Ausra Bendoraite**[2], **Patrick S. Mitchell**[2], **Angelique Nelson**[4,5], **Walter L. Ruzzo**[6,7], **Carol Ware**[4,5], **Jerald P. Radich**[1], **Robert Gentleman**[8], **Hannele Ruohola-Baker**[3,4], and **Muneesh Tewari**[1,2]

[1]Division of Clinical Research, Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA 98109, USA

[2]Division of Human Biology, Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA 98109, USA

[3]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

[4]University of Washington Institute for Stem Cell and Regenerative Medicine, Seattle, WA 98195, USA

[5]Department of Comparative Medicine, University of Washington, Seattle, WA 98195, USA

[6]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195-2350, USA

[7]Department of Genome Sciences, University of Washington Seattle, Washington 98195-5065, USA

[8]Public Health Sciences Division, Fred Hutchinson Cancer Research Center (FHCRC), Seattle, WA 98109, USA

Abstract

We used massively parallel pyrosequencing to discover and characterize microRNAs (miRNAs) expressed in human embryonic stem cells (hESC). Sequencing of small RNA cDNA libraries derived

Corresponding Author: Muneesh Tewari, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Mailstop D4-100, Seattle, WA 98109, Tel: (206) 667-5165, Fax: (206) 667-4723, mtewari@fhcrc.org.
*These authors contributed equally.

Author Contributions:
Merav Bar: Conception and design, Collection and/or assembly of data, Data analysis and interpretation, Manuscript writing
Stacia K.Wyman: Conception and design, Collection and/or assembly of data, Data analysis and interpretation, Manuscript writing
Brian R.Fritz.: Collection and/or assembly of data, Data analysis and interpretation, Manuscript writing
Junlin Qi: Provision of study material or patients, Collection and/or assembly of data
Kavita S.Garg: Collection and/or assembly of data, Data analysis and interpretation
Rachael K. Parkin: Collection and/or assembly of data
Evan M. Kroh: Collection and/or assembly of data
Ausra Bendoraite: Collection and/or assembly of data
Patrick S. Mitchell: Collection and/or assembly of data
Angelique Nelson: Provision of study material or patients
Walter L. Ruzzo: Conception and design
Carol Ware: Conception and design, Provision of study material or patients
Jerald Radich: Financial support, Data analysis and interpretation
Robert Gentleman: Conception and design, Data analysis and interpretation
Hannele Ruohola-Baker: Provision of study material or patients
Muneesh Tewari: Conception and design, Financial support, Data analysis and interpretation, Manuscript writing.

from undifferentiated hESC and from isogenic differentiating cultures yielded a total of 425,505 high-quality sequence reads. A custom data analysis pipeline delineated expression profiles for 191 previously annotated miRNAs, 13 novel miRNAs and 56 candidate miRNAs. Further characterization of a subset of the novel miRNAs in Dicer-knockdown hESC demonstrated Dicer-dependent expression, providing additional validation of our results. A set of 14 miRNAs (9 known and 5 novel) were noted to be expressed in undifferentiated hESC and then strongly down-regulated with differentiation. Functional annotation analysis of predicted targets of these miRNAs and comparison to a null model using non-hESC-expressed miRNAs identified statistically enriched functional categories, including chromatin remodeling and lineage-specific differentiation annotations. Finally, integration of our data with genome-wide chromatin immunoprecipitation data on OCT4, SOX2 and NANOG binding sites implicates these transcription factors in the regulation of nine of the novel/candidate miRNAs identified here. Comparison of our results to those of recent deep sequencing studies in mouse ESC and human ESC show that most of the novel/candidate miRNAs found here were not identified in the other studies. The data indicate that hESC express a larger complement of miRNAs than previously appreciated, and provide a resource for further studies of miRNA regulation of hESC physiology.

Keywords

microRNA; embryonic stem cells; deep sequencing; pyrosequencing

Introduction

The establishment of human embryonic stem cell lines has provided exciting new opportunities for the development of cell-based therapies to restore and maintain human health (reviewed in 1). The full realization of the therapeutic potential of hESC will, however, require an understanding of the molecular regulatory networks that control properties such as self-renewal and differentiation. MicroRNAs are small (typically ~22 nts in length) non-coding RNAs that play critical roles in molecular regulatory networks by post-transcriptional regulation of specific messenger RNA (mRNA) targets via direct base-pairing interactions 2. Genetic inactivation of the molecular machinery essential for proper maturation of miRNAs has been shown to cause aberrant stem cell self-renewal and/or differentiation in *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*[3–6], indicating that post-transcriptional regulation by miRNAs plays an important role in the networks that regulate stem cell activities.

A handful of studies have sought to characterize the expression of miRNAs in murine or human embryonic stem cells by using hybridization-based methods to evaluate a subset of the known miRNAs7–11, by sequencing clones from small RNA cDNA libraries 12, [13] and, more recently, by high throughput sequencing of murine ESC (mESC) [14] or hESC 15. In the study described here, we used massively parallel pyrosequencing 16 of small RNA cDNA libraries to characterize the ensemble of both known and novel miRNAs expressed in undifferentiated human embryonic stem cells, as well as in isogenic spontaneously differentiating cell populations. We performed a functional ontology analysis of the subset of known and novel miRNAs most likely to be relevant to hESC pluripotency, and integrated our results with genome-wide data on OCT4, SOX2 and NANOG promoter occupancy to identify novel miRNAs likely to be regulated by these transcription factors. The majority of the new hESC-expressed miRNAs discovered in our study were not identified in recent mESC and hESC miRNA sequencing studies, suggesting that our data complement the earlier studies and that the full repertoire of miRNAs expressed in hESC is larger than previously appreciated. We envision that our report will serve as a resource for future studies aimed at understanding and ultimately modulating hESC regulatory networks.

Materials and Methods Summary

H1 hESC were cultured under feeder-free conditions and Dicer-knockdown hESC were generated using a short hairpin-expressing lentivirus. RNA was extracted using the mirVana™ miRNA isolation kit (Ambion) and quantitative reverse transcription-PCR (qRT-PCR) was performed using TaqMan miRNA assays purchased from Applied Biosystems. Small RNA libraries were generated from 100 μg total RNA as previously described (17, http://web.wi.mit.edu/bartel/pub/protocols/miRNACloningUpdate0705.pdf). Massively parallel pyrosequencing was carried out at 454 Life Sciences, Inc.. Processing and annotation of sequences based on identity to known transcribed RNAs or as novel miRNAs was performed using a custom bioinformatics pipeline described in more detail in Supplementary Methods. Targets for known miRNAs were obtained by querying the TargetScan database and those for novel miRNAs by querying TargetScan Custom [18, 19]. Functional annotation analysis of predicted miRNA targets was done using Gene Ontology as described in detail in Supplementary Methods.

Additional details of Materials and Methods are supplied in Supplementary Methods.

Results

Overview of small RNA cDNA library sequencing

We chose to characterize the canonical H1 line as a representative, well-studied human embryonic stem cell line. Small RNAs were isolated from undifferentiated H1 human embryonic stem cells (designated Undiff-hESC) that were cultured under feeder-free conditions to avoid contamination with potential murine miRNAs from feeder cells. We reasoned that miRNAs with ESC-specific functions might be expected to exhibit changes in expression concurrent with loss of pluripotency and commitment to differentiation. For comparison, therefore, we also characterized miRNAs expressed in isogenic, spontaneously differentiating populations of H1 cells (designated Diff-hESC) that had been triggered to differentiate by culture for 10–14 days in the absence of basic fibroblast growth factor (bFGF) and conditioned medium (additional detail is provided in Supplemental Methods). Analysis of expression of pluripotency and differentiation markers in Undiff-hESC and Diff-hESC RNA samples using qRT-PCR demonstrated (i) the expression of pluripotency markers in Undiff-hESC and (ii) the loss of pluripotency marker expression and induction of markers corresponding to endoderm, mesoderm and ectoderm lineages in the Diff-hESC cell population (Supplemental Figure 1).

Small RNA cDNA libraries were generated by ligating 5’ and 3’ linkers to 18–24 nt size-fractionated RNA, followed by reverse transcription and PCR amplification (Figure 1A). Massively parallel pyrosequencing using the 454 Life Sciences’ platform generated 281,543 and 143,962 high-quality sequence reads corresponding to 18,227 and 16,096 nonredundant sequences derived from Undiff-hESC and Diff-hESC cultures, respectively. To analyze the sequence data, we constructed a custom computational pipeline. The initial operations of the pipeline included identifying sequence matches to databases of previously annotated RNAs (e.g., known miRNAs, other noncoding RNAs, messenger RNAs) and to repetitive sequence elements (Figure 1B; additional details are provided in Supplemental Methods). In the next sections, we discuss in detail the sequences matching to known miRNAs and the identification of novel miRNAs.

MicroRNA profiling: previously annotated miRNAs

Matches to known miRNAs in miRBase (release 9.0) [20, 21] represented 62.6% of the Undiff-hESC reads and 64.4% of the Diff-hESC reads. A total of 191 known miRNAs were identified

(Supplemental Table 1), with a greater total number of miRNAs expressed in Diff-hESC than in Undiff-hESC (Figure 2).

The cloning frequency of individual miRNAs, expressed as a percentage of total reads obtained from a given sample, can be used to compare relative expression of miRNAs between samples [22–24], keeping in mind that there are limitations related to the fact that replicate sequencing datasets are not available for each of the samples. Differential expression of miRNAs between Undiff-hESC and Diff-hESC cultures, assessed using this approach, is shown in Figure 3. We were particularly interested in miRNAs that are expressed in Undiff-hESC and diminish in expression with differentiation, as these miRNAs might be expected to participate in ESC-specific functions. The ten most over-expressed miRNAs in Undiff-hESC were hsa-miR-302b, -miR-302c, -miR-302d, -miR-92b, -miR-20b, -miR-519d, -miR-302a, -miR-324-3p, -miR-187, and -miR-18b (Figure 3A; all miRNA names in this paper refer to *H. sapiens* miRNAs unless otherwise specified; for brevity, we have hereafter omitted the hsa- prefix). All of these miRNAs had ≥2.5-fold over-expression and Fisher's exact test *P*-values of <0.05. Fisher's exact test has been shown to be useful for assessing statistical significance of digital gene expression data of the kind obtained in our study [25]. In addition, some miRNAs were over-expressed in Undiff-hESC but could not be assigned a fold over-expression value because there were no corresponding reads in Diff-hESC. Eight such miRNAs were represented by three or more reads each in Undiff-hESC: miR-518b, miR-520g, miR-524*, miR-363*, miR-154, miR-184, miR-518c, and miR-512-3p (Figure 3B).

As an independent confirmation of over-expression in Undiff-hESC, we sought to use commercially available TaqMan® qRT-PCR assays to measure the expression of the 18 miRNAs mentioned above in RNA from Undiff-hESC and from Diff-hESC. Effective qRT-PCR assays were available for 14 of the 18 miRNAs and confirmed over-expression by ≥2.5-fold for 13 of these 14 miRNAs (Figure 4A).

To determine whether these observations extend beyond the H1 cell line, we used qRT-PCR to examine expression of these 13 miRNAs in cultures of Undiff-hESC and Diff-hESC corresponding to the BG01 hESC line. We chose the BG01 line because it was derived at a different institution than the H1 cell line and has phenotypic characteristics that distinguish it from the H1 line [26]. All 13 miRNAs assayed were significantly over-expressed (≥2.5 fold) in BG01 Undiff-hESCs relative to BG01 Diff-hESC cultures. Furthermore, the pattern of fold over-expression values seen across the set of miRNAs closely paralleled that observed in H1 hESC (Figure 4B).

As additional validation of our deep sequencing-based miRNA differential expression results, we used qRT-PCR to assay the expression of five representative miRNAs found to be over-expressed in H1 Diff-hESC relative to H1 Undiff-hESC: miR-23a, miR-27b, miR-125a, miR-152, and miR-324-5p (Supplemental Table 1). For all five of these miRNAs, qRT-PCR results confirmed over-expression by at least 5-fold in Diff-hESC relative to Undiff-hESC expression patterns (data not shown).

Although the sequencing approach described here characterized a wide range of miRNAs expressed in hESC, we also noted an instance in which a miRNA expected to be expressed was not detected. MicroRNA miR-367, which is generated from the same primary transcript as miR-302a-d, was strikingly absent in our dataset. Using TaqMan qRT-PCR, however, we found strong expression of miR-367 in RNA from Undiff-hESC (data not shown). Landgraf et al. also observed this phenomenon (i.e., the absence of certain abundant miRNAs in sequencing data) in their high-throughput miRNA sequencing study, which they attributed to the influence of sequence-specific miRNA secondary structure on linker ligation efficiency [23].

Analysis of sequencing data to identify novel hairpin-derived small RNAs

Having profiled the complement of known miRNAs, we then identified sequences that represent novel star forms of known miRNAs (Figure 1B). MicroRNAs are generated from ~80 nt stem-loop precursor RNA transcripts that are processed by Dicer to generate a ~22 nt mature double-stranded RNA. One of the strands of the mature duplex is preferentially loaded into the miRNA-induced silencing complex, whereas the other strand, designated the miRNA "star form," is thought to be degraded [27]. We discovered a total of 52 novel star forms of known miRNAs (Supplemental Table 2 and Supplemental Methods). As expected, the relative abundance of most of the novel star forms was lower than that of their corresponding miRNAs.

The data analysis pipeline (Figure 1B) identified matches to other classes of known noncoding RNAs (Supplemental Table 3), to annotated protein-coding messenger RNAs in the RefSeq database (less than 1% of reads in each dataset), and to repetitive sequences (i.e., matches to Repbase or any sequences mapping to 20 or more loci in the genome). Sequences remaining after applying these filters were aligned to the human genome sequence (NCBI Build 36.1) [28]. Sequences were required to match the human genome sequence perfectly to be carried further for additional analysis. The only exception to this was that sequences having an additional 1–3 non-templated nucleotides present at the 3' terminus were trimmed of the non-templated bases before additional analysis. Other investigators have observed this phenomenon of non-templated addition of nucleotides to miRNAs at the 3' terminus 29, 30 and have adopted a similar approach in their analysis 29. As additional filtering steps, sequences that intersected the RepeatMasker track in the UCSC Genome Browser and sequences outside the 20–24 nucleotide length range were removed from further analysis.

The data processing steps described thus far yielded 3,115 and 1,994 unique sequences corresponding to novel small RNAs in the Undiff-hESC and Diff-hESC datasets, respectively. These unique sequences corresponded to 5,595 and 3,921 genomic loci that could potentially generate these small RNAs.

A fundamental criterion for defining miRNAs is their biogenesis from a predicted fold-back hairpin precursor transcript that contains the mature miRNA sequence within one arm of the hairpin 31. Therefore, the data analysis pipeline next screened genomic loci corresponding to unique sequences for the presence of a predicted hairpin secondary structure. We used several criteria for designating a sequence as "folding into a hairpin" including free energy minimization, shape probability (as determined by the RNAshapes program 32), and the Randfold-computed 33 *P*-value of predicted secondary structures. We also required that the pairing characteristics of the hairpin be such that the novel sequence is wholly contained within one arm of a putative hairpin precursor sequence and that the degree of base-pairing be consistent with that observed in precursors corresponding to known miRNAs in miRBase (a detailed description of hairpin folding criteria is provided in Supplemental Methods). The thresholds chosen to define novel hairpins were sufficiently stringent such that only 86% of the known miRNAs in miRBase release 9.0 would satisfy the hairpin folding criteria. By this analysis, 531 and 364 of the novel small RNAs from Undiff-hESC and Diff-hESC datasets, respectively, were found to be potentially derived from a precursor hairpin structure.

These sequences were then sorted with respect to chromosomal coordinates into groups sharing 5' ends. From each of these groups we chose a "canonical" sequence representing the group of sequences arising from that genomic locus. The canonical sequences were chosen based on common 5' terminus, abundance, and sequence length (see Supplemental Methods for details). This process further refined our sequences into a combined list of 285 unique sequences (potentially derived from 315 genomic loci) that we designated as "novel hairpin-derived small RNAs."

Identifying novel and candidate miRNAs

In order to find novel miRNAs, we screened the set of novel hairpin-derived small RNAs using criteria similar to those used in recent miRNA discovery studies [30, 34]: (1) pairing characteristics of the hairpin (an absolute requirement, as screened for in the previous section), (2) the required presence of multiple reads sharing the same 5' terminus, (3) evolutionary conservation, as reflected by an apparent conserved hairpin with identical seed region in another species (with greater weight given to non-primate conservation), (4) absence of annotation indicating non-miRNA biogenesis (an absolute requirement, as screened out in earlier steps of the pipeline), (5) shared seed region with a known animal miRNA and (6) presence of corresponding miRNA star form read(s). As in analyses by Ruby et al.[30, 34], we considered the finding of both a miRNA and a corresponding miRNA star form as compelling evidence for biogenesis from a hairpin precursor.

Thirteen of our sequences sufficiently met these criteria to be designated novel miRNAs (Table 1). Six of these sequences met five of the above criteria and six met four of the criteria. One sequence (U755.1-4/D10092.1) met only three criteria but was included based on the abundance of reads with consistent 5' ends (19 reads) and differential expression between Undiff-hESC and Diff-hESC (Table 1 and Supplemental Table 4; this sequence was also subsequently experimentally validated to exhibit Dicer-dependent expression, as described in the next section). Of note, three of the novel miRNAs had a seed region shared with previously annotated animal miRNAs (Table 1; more detail is provided in Supplemental Table 4). In addition, mapping the novel miRNA sequences to the reference human genome sequence revealed that 11 of the 13 novel miRNAs are present in introns of other genes (and encoded on the same strand as the respective host genes), much like many previously annotated miRNAs (Table 1; Supplemental Table 4). The sequence of miRNA star forms corresponding to novel miRNAs is provided in Supplemental Table 4 as well as in the context of the predicted precursor structure in Supplemental Figure 2.

The remaining sequences comprised 268 RNAs (corresponding to 291 genomic loci) that (i) had length 20–24 nt, (ii) met folding criteria and (iii) had an absence of annotation indicating non-miRNA biogenesis, but that did not meet sufficient additional criteria to be confidently annotated as novel miRNAs. We sought to select from this list the most promising sequences to designate as "candidate miRNAs" that might be confirmed in the future as *bona fide* miRNAs as additional evidence accumulates.

We required candidate miRNAs to have at least three reads, or to have two reads and additional supporting evidence of either a homologous conserved hairpin in at least one other vertebrate species or of a seed region shared with a known animal miRNA. In addition, a handful of sequences represented only by singleton reads were included in the candidate miRNA list because there was an abundance of evidence supporting their annotation as miRNA candidates: a given singleton sequence either had homologous hairpins conserved in multiple non-primate vertebrates *and* had a primate homologous hairpin; *or* it had a single conserved non-primate homologous hairpin, a primate homologous hairpin, *and* it had a shared seed sequence with a known animal miRNA. Taken together, this allowed us to refine a final list of 56 candidate miRNAs (originating from 68 potential genomic loci) (see Supplemental Table 4).

It is important to note that although all our analysis was initially performed using miRBase release 9.0, with the availability of miRBase release 10.0 we compared all our hairpin-derived small RNA sequences to release 10.0 and reclassified those that corresponded to newly deposited known miRNA sequences. This affected only 13 of the hairpin-derived small RNAs, which are listed at the bottom of Supplemental Table 1. The more recent availability of miRBase release 11.0 reclassified three novel and two candidate miRNAs as known. These

miRNAs are highlighted in blue or orange and annotated in the last column of Supplemental Table 4.

Experimental validation of novel miRNAs

In order to provide additional support for the assertion that the novel miRNAs identified in this study are *bona fide* miRNAs, we sought to use Custom TaqMan® Small RNA Assays to examine expression of the novel miRNAs in hESC transduced with either an shRNA directed against Dicer or transduced with a vector control lentivirus. qRT-PCR for Dicer mRNA confirmed substantial knockdown (84% decrease) relative to vector control transduced cells (Figure 5A). We were able to obtain robust Taqman qRT-PCR assays for three of the novel miRNAs. All three of these miRNAs showed significantly diminished expression in the Dicer knockdown hESC relative to vector control (Figure 5B). SnoRNAs, which are not expected to be processed by Dicer, served as negative controls and were not reduced in expression by Dicer knockdown, whereas three known miRNAs (serving as positive controls) showed diminished expression in Dicer knockdown hESC as expected (Figure 5B). The results are strong evidence that these novel miRNAs are products of Dicer-dependent maturation and suggest that the same is likely to hold true for other novel miRNAs identified here.

Potential regulation of novel and candidate miRNAs by the pluripotency-associated transcription factors OCT4, SOX2 and NANOG

Given the profound influence of the pluripotency-associated transcription factors OCT4, SOX2 and NANOG on gene expression in hESC [35], we hypothesized that some of the novel and candidate miRNAs discovered in this study may be regulated by these transcription factors. To investigate this hypothesis further, we turned to results of published chromatin immunoprecipitation-microarray (ChIP-chip) experiments that had identified genome-wide binding sites for these factors in hESC [36]. Boyer at al. subjected each of these three transcription factors to chromatin immunoprecipitation followed by analysis of bound DNA on microarrays containing 60-mer DNA oligonucleotide probes covering the region from −8 kb to +2 kb relative to the transcription start sites for 17,917 annotated human genes.

In order to determine whether OCT4, SOX2 or NANOG binding sites identified in the ChIP-chip experiments correspond to genomic regulatory regions for the miRNAs discovered in our study, we first sought to define transcriptional start sites (TSSs) corresponding to our novel and candidate miRNAs. For sequences that were intronic to a well-annotated RefSeq gene (and encoded from the same strand), co-transcription with the host gene was presumed and the TSS was taken to be the annotated TSS of the host gene. For the remaining sequences, we used AceView gene models (which rely heavily on EST data) and Eponine TSS predictions, whenever available, from the UCSC Genome Browser tracks to define a TSS. We were able to identify (i) a RefSeq-based TSS for 10 of the 13 novel and 36 of the 56 candidate miRNAs discovered in this study and (ii) an Aceview and/or Eponine-based TSS for 0 of the 3 remaining novel miRNAs and for 2 of the remaining 21 candidate miRNAs. We then intersected these TSSs with the genome-wide OCT4, SOX2 and NANOG bound sites defined by the ChIP-chip data of Boyer et al., requiring that a ChIP-defined binding site be located between −8 kb and +2kb of the TSS of our novel and candidate miRNAs. We found that 2 of the 10 novel miRNAs for which TSSs could be defined and 6 of the 38 candidate miRNAs for which TSSs could be defined had evidence for OCT4, SOX2 and/or NANOG binding at their genomic loci. The data is annotated in Supplemental Table 4 and described in more detail in Supplemental Table 5. Collectively, it appears that 8 out of the 48 novel and candidate miRNAs for which TSSs could be defined are associated with occupancy of and potential regulation by these pluripotency-associated transcription factors.

Independent EST-based evidence for expression of novel and candidate miRNAs in hESC and related multipotent cells

Many known miRNAs are encoded in introns of other genes and are co-transcribed with their host genes. In these cases, it is possible to use expression of the host gene as a surrogate to infer expression of the intronic miRNA. We used this approach to obtain independent evidence for transcription of the novel and candidate miRNAs in hESC or embryonal carcinoma/teratocarcinoma cell lines, by first identifying those miRNAs that were intronic to host transcripts defined by ESTs, and then asking whether any of the host transcript ESTs were derived from hESC or human embryonal carcinoma/teratocarcinoma cell lines. This approach provided independent validation of transcription in hESC or embryonal carcinoma/teratocarcinoma cell lines for seven of the 13 novel miRNAs, and for 21 candidate miRNA loci of the 68 loci corresponding to candidate miRNAs identified in this study. These results are annotated in Supplemental Table 4.

Comparison with results from murine ESC deep sequencing

Calabrese et al. recently characterized miRNAs in murine ESC (mESC) by deep sequencing, defining 46 novel and 52 candidate miRNAs, many of which are genomic repeat associated [14]. Considering non-repeat-associated miRNAs, their study reported 22 novel and 21 candidate miRNAs. We compared our list of novel and candidate miRNAs identified in hESC to the corresponding list derived in the murine ESC study and we did not find any sequences in common. Given that miRNAs can be poorly conserved in overall sequence, we next compared the seed region of novel and candidate miRNAs identified in our hESC data to the novel and candidate miRNAs identified in mESC. There we found a single seed region match between one of our candidates (D12354.1) and one of the mESC novel miRNAs (mmu-miR-466j). It is notable that even when known miRNAs are considered, dramatic differences between miRNA expression in mESC and hESC exist.

Comparison of novel and candidate miRNA data with results from a recent hESC small RNA sequencing study

While the current manuscript was under preparation, Morin et al. [15] reported the discovery of 83 novel miRNAs (corresponding to 104 genomic loci) in RNA from undifferentiated hESC and embryoid bodies derived from the same. We compared our novel and candidate miRNAs to the ones discovered by Morin et al. and found, notably, minimal overlap. Of the 83 novel miRNAs reported by Morin et al., only 22 were present at all in our raw sequencing data. Of these 22 sequences, only three met our criteria for annotation as novel or candidate miRNAs (highlighted in blue in Supplementary Table 4). The other 19 were ruled out by our classification scheme for reasons such as being repeat-associated, or matching other previously annotated features like tRNAs or RefSeq genes. Conversely, Morin et al. discovered only 3 of our 13 novel and only 1 of our 56 candidate miRNAs. One explanation for the differing results is that we studied spontaneously differentiated cultures of hESC as compared to their study of embryoid bodies; however, when results only from the Undiff-hESC cultures are compared (which represent more similar cell populations), the overlap is still minimal. Alternative explanations include (i) technical differences in library construction and sequencing platforms used, and (ii) that the diversity of novel miRNAs/small RNAs in hESC is greater than anticipated, such that neither our study nor that of Morin et al. have reached saturation.

Undifferentiated hESC-associated miRNAs and their predicted targets

MiRNAs that are expressed in Undiff-hESC but that diminish in expression with the induction of differentiation are of particular interest because they may participate in functions related to the pluripotent state. We defined a set of known and novel miRNAs falling into this category by selecting all miRNAs exhibiting a 4-fold or greater over-expression in Undiff-hESC (in

cases were the miRNA was identified in both Undiff-hESC and Diff-hESC), or being represented by at least 10 reads in the Undiff-hESC cells (in cases where the miRNA was not detected at all in the Diff-hESC population). This set of "Undiff-hESC-associated miRNAs" comprised 14 miRNAs, of which five were novel and nine were previously annotated miRNAs (Figure 6).

We undertook a functional annotation analysis to gain insight into the processes that might be regulated by these miRNAs. We began by obtaining computationally predicted targets of the known and novel miRNAs in this group using TargetScan and TargetScan Custom, respectively. The TargetScan algorithm uses seed region matches between miRNAs and their potential targets, as well as phylogenetic conservation of those matches to identify predicted targets of miRNAs. MiRNAs with the same seed region (e.g., miR-20b and miR-519d) are therefore considered as one because they have identical TargetScan target predictions.

We used the Gene Ontology [37] to obtain functional descriptions of the predicted miRNA targets for each of the Undiff-hESC miRNAs, focusing our analysis on 12,821 genes found to be expressed in Undiff-hESC based on microarray analysis ("Undiff-hESC-expressed genes", with details of microarray analysis provided in Experimental Procedures and Supplemental Methods). The Gene Ontology (GO) project uses a controlled vocabulary to describe gene products in a variety of organisms. After intersecting the predicted targets of each miRNA with the group of 12,821 hESC-expressed genes, we initially identified GO Biological Process (BP) categories that were significantly enriched (data not shown).

Although many of these categories encompassed biological functions relevant to ESC and early development, miRNAs in general may target genes involved in early developmental processes. Therefore, in order to identify enriched categories that are more specific to the Undiff-hESC-associated miRNAs, we took a different approach based on comparison to a null distribution generated by analyzing the GO functional annotations of the 295 known miRNAs (representing 266 unique seed regions) that were *not* detected as being expressed in Undiff-hESC in our sequencing dataset. We then compared the statistical significance of over-represented GO categories associated with targets for each Undiff-hESC-associated miRNA to this null distribution. From this comparison, we selected only those categories that, for the targets of a given Undiff-hESC-associated miRNA, returned a NullP-value<0.01 (see Supplemental Methods for details of the NullP-value calculation). The full results of this analysis are given in Supplemental Table 6, in which categories of particular interest are highlighted in yellow. Of particular interest is the over-representation of myeloid/erythroid differentiation and chromatin remodeling genes in predicted targets of the novel miRNA U739.1-6, as well as over-representation of BMP signaling pathway and cell differentiation categories in predicted targets of miRNA U755.1-4/D10092.1.

To enable convenient access to predicted targets of all new miRNAs reported in this study, we used the TargetScan Custom algorithm to predict targets for all of the novel miRNAs and candidate miRNAs reported here. A complete list of these predicted targets is available in Supplemental Tables 7 and 8.

Discussion

The work reported here was motivated by the hypothesis that the entire repertoire of miRNAs expressed in hESC had not yet been elucidated, and by genetic evidence indicating that miRNAs play a critical role in embryonic stem cell function. The massively parallel sequencing approach allowed us to be comprehensive (i.e., identifying not only known but also novel miRNAs), and the "digital" nature of the data permitted a semi-quantitative estimation of the relative expression level for many miRNAs. Using the deep sequencing approach, we identified

13 novel and 56 candidate miRNAs, as well as 191 previously annotated miRNAs. We suggest that some of the novel miRNAs identified here may be hESC-specific, by virtue of not having been identified in recent high-throughput sequencing-based surveys of miRNA expression across various differentiated tissues [23, 24, 38].

Although the overall trend was for an increase in expression of most miRNAs in the cell population following loss of pluripotency and differentiation, a subset of five novel and nine known miRNAs (designated as Undiff-hESC-associated) clearly showed the opposite expression pattern and may represent a miRNA signature of pluripotency in hESC cultures. This group of miRNAs also represents a critical set for future functional studies because they may regulate pluripotency or other hESC-specific functions. Although genetic perturbation experiments will ultimately be required to unravel the functions of these miRNAs, the number of potential perturbations and specific phenotypes that could be tested is vast, especially when multiple miRNAs are considered. Our results from Gene Ontology analysis of predicted targets may help in this regard by suggesting hypotheses of function for specific Undiff-hESC-associated miRNAs to guide further investigation.

It is particularly notable that the novel and candidate miRNAs discovered here were not found in recent small RNA sequencing datasets from mESC and hESC. The difference from mESC may be explained, at least in part, by the known phenotypic differences between hESC and mESC [35], as reflected in differences even in the expression of *known* miRNAs between hESC and mESC (e.g., expression of miR-302 family miRNAs) [14]. The lack of overlap of novel miRNAs discovered here with those in a recently published hESC dataset 15, however, suggests that even despite obtaining sequences on the scale of massively parallel sequencing studies, we cannot yet exclude the possibility that the entire space of miRNAs expressed in hESC is still not fully elucidated. That said, it is important to note that many of the novel miRNAs in all the sequencing studies are expressed at low levels and their functional roles have not yet been characterized. Further studies will be needed to understand the biological significance of the complement of novel hESC-expressed miRNAs identified in this and other studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Harlan Robins, Stephen Tapscott, Beverly Torok-Storb and John Byon for critical review of the manuscript. We would also like to thank anonymous reviewers for helpful comments and suggestions. We thank Iain Russell, David Petrillo and Matthew Rockwell for facilitating access to Custom TaqMan® Small RNA Assays, and acknowledge the assistance of the Genomics and Scientific Imaging Shared Resources at the Fred Hutchinson Cancer Research Center.

Acknowledgment of research funding support:

This work was supported by Oncology Training Grant 5 T32 CA009515-21/22 and Career Development in Pediatric and Medical Oncology award NCI / 5 K12 CA076930 to M. Bar, Chromosome Metabolism Training Grant 5 T32 CA09657-16 to S. Wyman, Interdisciplinary Training in Cancer Research Grant CA80416 to K. Garg, NIEHS P30ES07033 to W. L. Ruzzo, P41 HG004059-01 to R. Gentleman, NIGMS / P20 GM069983-01 and NIGMS / P01 GM081619-01 to C. Ware, the Tietze Award and grants from NIH and MOD to H. Ruohola-Baker, and a Pilot Award from the NIH/NCI Cancer Center Support Grant 5 P30 CA015704 and FHCRC New Development funds to M. Tewari. We acknowledge financial support for sequencing from Roche Diagnostics.

References

1. Mimeault M, Hauke R, Batra SK. Stem cells: a revolution in therapeutics-recent advances in stem cell biology and their therapeutic applications in regenerative medicine and cancer therapies. Clin Pharmacol Ther 2007 Sep;82(3):252–264. [PubMed: 17671448]
2. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 2004 Jul;5(7):522–531. [PubMed: 15211354]
3. Reinhart BJ, Slack FJ, Basson M, et al. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 2000 Feb 24;403(6772):901–906. [PubMed: 10706289]
4. Hatfield SD, Shcherbata HR, Fischer KA, Nakahara K, Carthew RW, Ruohola-Baker H. Stem cell division is regulated by the microRNA pathway. Nature 2005 Jun 16;435(7044):974–978. [PubMed: 15944714]
5. Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ. Characterization of Dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A 2005 Aug 23;102(34):12135–12140. [PubMed: 16099834]
6. Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet 2007 Mar;39(3):380–385. [PubMed: 17259983]
7. Strauss WM, Chen C, Lee CT, Ridzon D. Nonrestrictive developmental regulation of microRNA gene expression. Mamm Genome 2006 Aug;17(8):833–840. [PubMed: 16897339]
8. Chen C, Ridzon D, Lee CT, Blake J, Sun Y, Strauss WM. Defining embryonic stem cell identity using differentiation-related microRNAs and their potential targets. Mamm Genome 2007 May;18(5):316–327. [PubMed: 17610011]
9. Josephson R, Ording CJ, Liu Y, et al. Qualification of Embryonal Carcinoma 2102Ep As a Reference for Human Embryonic Stem Cell Research. Stem Cells 2007 Feb;25(2):437–446. [PubMed: 17284651]
10. Lakshmipathy U, Love B, Goff LA, et al. MicroRNA expression pattern of undifferentiated and differentiated human embryonic stem cells. Stem Cells Dev 2007 Dec;16(6):1003–1016. [PubMed: 18004940]
11. Wu H, Xu J, Pang ZP, et al. Integrative genomic and functional analyses reveal neuronal subtype differentiation bias in human embryonic stem cell lines. Proc Natl Acad Sci U S A 2007 Aug 21;104(34):13821–13826. [PubMed: 17693548]
12. Houbaviy HB, Murray MF, Sharp PA. Embryonic stem cell-specific MicroRNAs. Dev Cell 2003 Aug;5(2):351–358. [PubMed: 12919684]
13. Suh MR, Lee Y, Kim JY, et al. Human embryonic stem cells express a unique set of microRNAs. Dev Biol 2004 Jun 15;270(2):488–498. [PubMed: 15183728]
14. Calabrese JM, Seila AC, Yeo GW, Sharp PA. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. Proc Natl Acad Sci U S A 2007 Nov 13;104(46):18097–18102. [PubMed: 17989215]
15. Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 2008 Feb 19;
16. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 Sep 15;437(7057):376–380. [PubMed: 16056220]
17. Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 2001 Oct 26;294(5543):858–862. [PubMed: 11679671]
18. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005 Jan 14;120(1):15–20. [PubMed: 15652477]
19. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007 Jul 6;27(1):91–105. [PubMed: 17612493]
20. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006 Jan 1;34(Database issue):D140–D144. [PubMed: 16381832]

21. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res 2004 Jan 1;32(Database issue):D109–D111. [PubMed: 14681370]
22. Aravin A, Tuschl T. Identification and characterization of small RNAs involved in RNA silencing. FEBS Lett 2005 Oct 31;579(26):5830–5840. [PubMed: 16153643]
23. Landgraf P, Rusu M, Sheridan R, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 2007 Jun 29;129(7):1401–1414. [PubMed: 17604727]
24. Cummins JM, He Y, Leary RJ, et al. The colorectal microRNAome. Proc Natl Acad Sci U S A 2006 Mar 7;103(10):3687–3692. [PubMed: 16505370]
25. Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res 1997 Oct;7 (10):986–995. [PubMed: 9331369]
26. Ware CB, Nelson AM, Blau CA. A comparison of NIH-approved human ESC lines. Stem Cells 2006 Dec;24(12):2677–2684. [PubMed: 16916927]
27. Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell 2003 Oct 17;115(2):199–208. [PubMed: 14567917]
28. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001 Feb 15;409(6822):860–921. [PubMed: 11237011]
29. Berezikov E, Cuppen E, Plasterk RH. Approaches to microRNA discovery. Nat Genet 2006 Jun; 38:S2–S7. [PubMed: 16736019]
30. Ruby JG, Jan C, Player C, et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 2006 Dec 15;127(6):1193–1207. [PubMed: 17174894]
31. Ambros V, Bartel B, Bartel DP, et al. A uniform system for microRNA annotation. Rna 2003 Mar;9 (3):277–279. [PubMed: 12592000]
32. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 2006 Feb 15;22(4):500–503. [PubMed: 16357029]
33. Bonnet E, Wuyts J, Rouze P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics 2004 Nov 22;20(17):2911–2917. [PubMed: 15217813]
34. Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Res 2007 Dec;17(12):1850–1864. [PubMed: 17989254]
35. Boiani M, Scholer HR. Regulatory networks in embryo-derived pluripotent stem cells. Nat Rev Mol Cell Biol 2005 Nov;6(11):872–884. [PubMed: 16227977]
36. Boyer LA, Lee TI, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 2005 Sep 23;122(6):947–956. [PubMed: 16153702]
37. Ashburner M, Ball CA, Blake JA, et al. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000 May;25(1):25–29. [PubMed: 10802651]
38. Berezikov E, Thuemmler F, van Laake LW. Diversity of microRNAs in human and chimpanzee brain. Nat Genet 2006 Dec;38(12):1375–1377. [PubMed: 17072315]

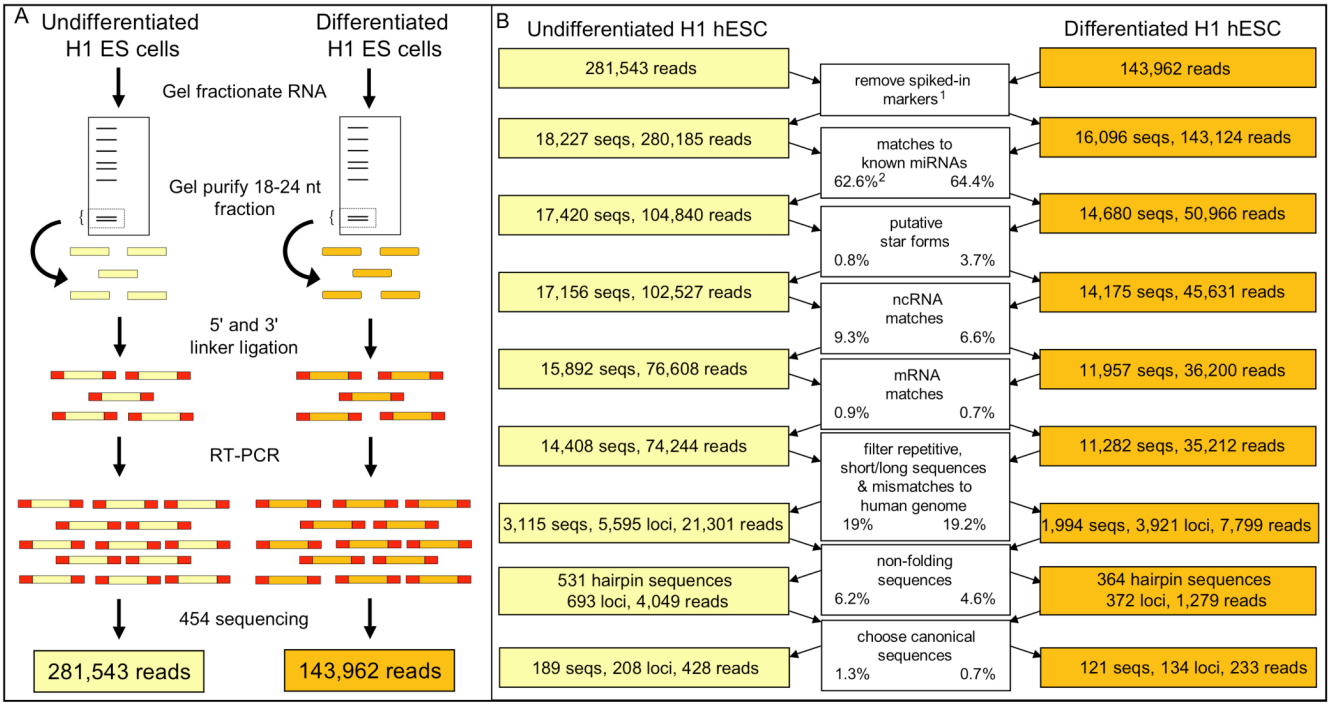


Figure 1. Small RNA library generation and data analysis pipeline

(A) The small RNA library generation and sequencing scheme is shown. Small RNAs were isolated from undifferentiated H1 hESC and isogenic spontaneously differentiating cultures. Following 3' and 5' linker ligation, RT-PCR was performed to generate two independent cDNA libraries of small RNAs that were then used as templates for massively parallel pyrosequencing (454 sequencing). (B) The flow chart describes the data analysis pipeline. "Seqs" represent nonredundant sequences derived after collapsing multiple reads of identical sequence. The columns flanking the middle column indicate the number of sequences and reads remaining at each step of the data analysis. At the end of the pipeline, 189 sequences in the Undiff-hESC dataset and 121 in the Diff-hESC dataset met our criteria for canonical hairpin-derived sequences. Loci numbers are higher because some canonical sequences map to more than one locus in the genome. [1]The initial step of the data analysis was removal of sequences corresponding to 18 nt and 24 nt RNA markers that had been spiked into the total RNA prior to gel electrophoresis. [2]Percentages of total reads from Undiff-hESC and Diff-hESC datasets that were classified into the designated categories and filtered out at each step are listed in the middle boxes.

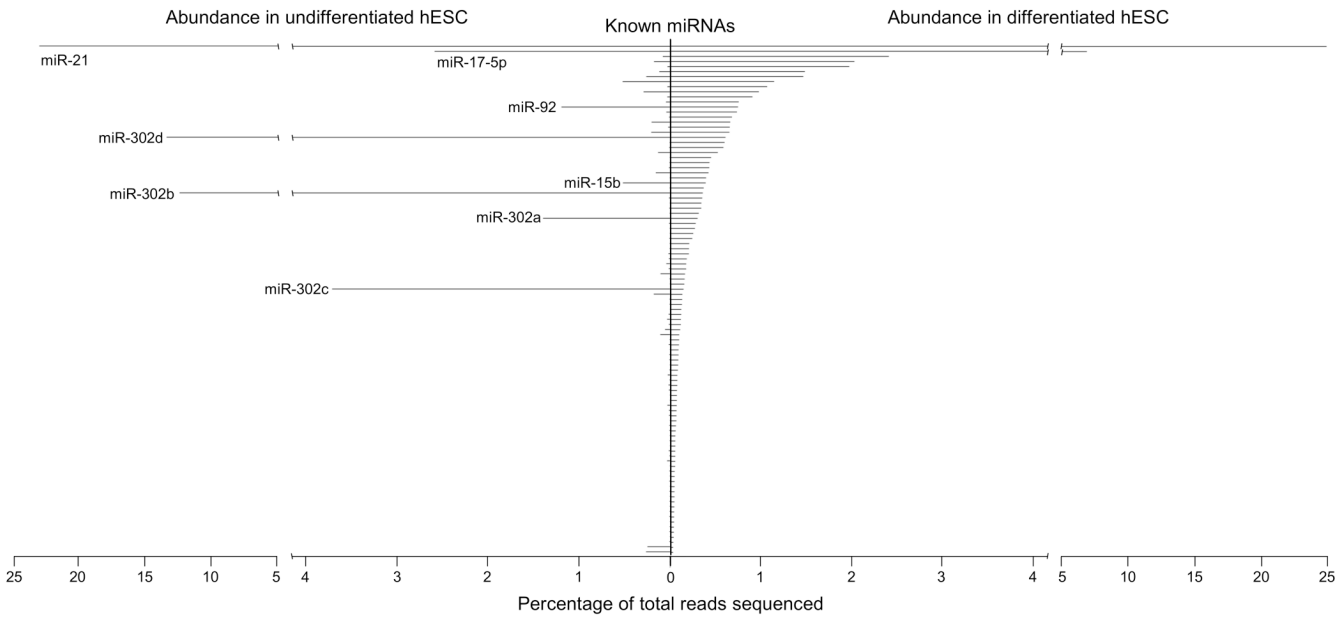
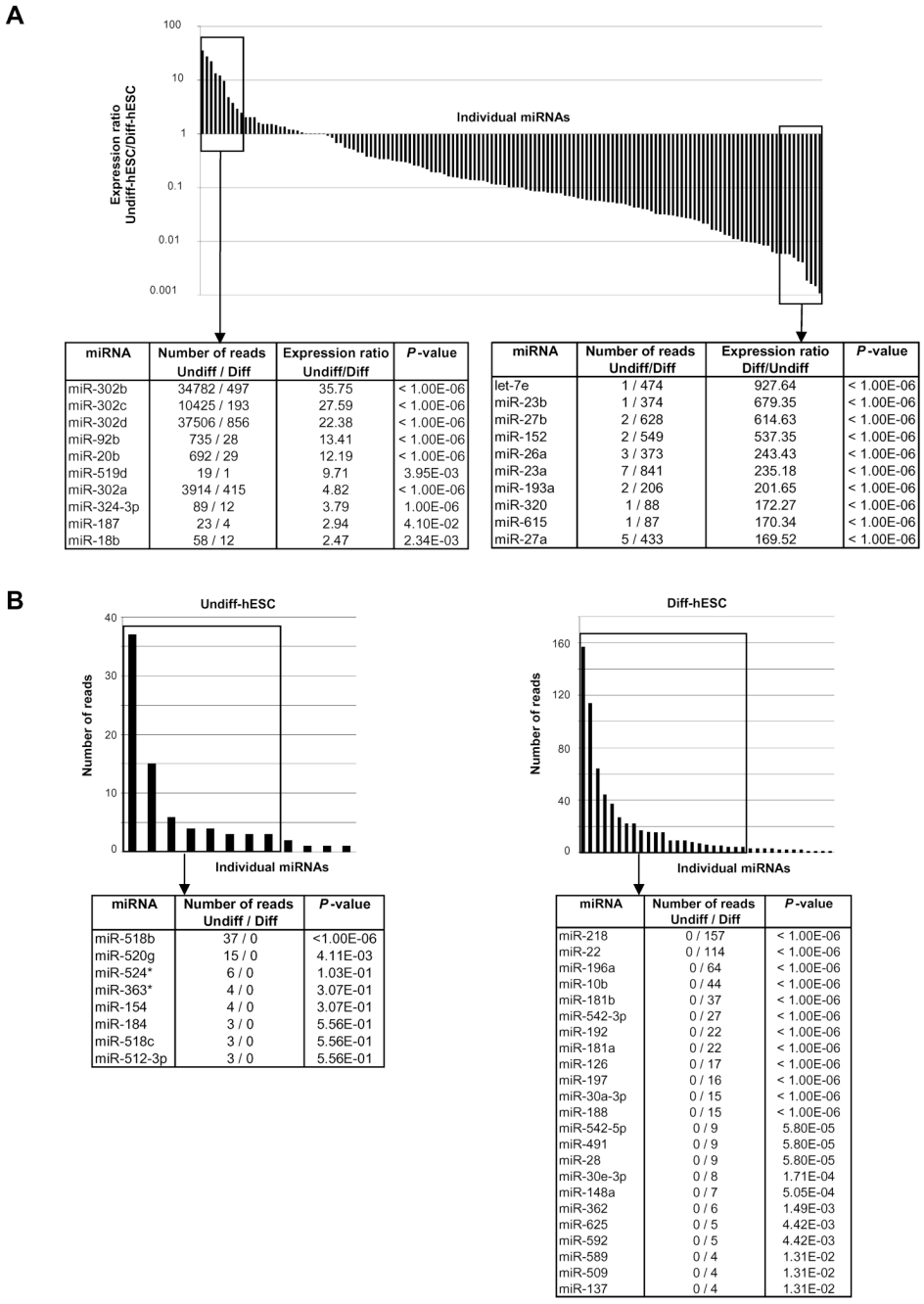


Figure 2. Global view of known miRNAs detected in hESC sequencing datasets

The percentage of total reads for a given miRNA in Undiff-hESC or Diff-hESC reflects its relative abundance in each cell population. The 100 most abundant miRNAs are shown arranged in order of decreasing abundance in Diff-hESC (as the lower abundance miRNAs would not be visible on the graph with the same scale). Selected miRNAs present at high abundance in Undiff-hESC are identified by name. A full list of all the known miRNAs identified in this study and their relative abundance in each dataset is available in Supplemental Table 1.



miRNA	Number of reads Undiff / Diff	Expression ratio Undiff/Diff	*P*-value
miR-302b	34782 / 497	35.75	< 1.00E-06
miR-302c	10425 / 193	27.59	< 1.00E-06
miR-302d	37506 / 856	22.38	< 1.00E-06
miR-92b	735 / 28	13.41	< 1.00E-06
miR-20b	692 / 29	12.19	< 1.00E-06
miR-519d	19 / 1	9.71	3.95E-03
miR-302a	3914 / 415	4.82	< 1.00E-06
miR-324-3p	89 / 12	3.79	1.00E-06
miR-187	23 / 4	2.94	4.10E-02
miR-18b	58 / 12	2.47	2.34E-03

miRNA	Number of reads Undiff/Diff	Expression ratio Diff/Undiff	*P*-value
let-7e	1 / 474	927.64	< 1.00E-06
miR-23b	1 / 374	679.35	< 1.00E-06
miR-27b	2 / 628	614.63	< 1.00E-06
miR-152	2 / 549	537.35	< 1.00E-06
miR-26a	3 / 373	243.43	< 1.00E-06
miR-23a	7 / 841	235.18	< 1.00E-06
miR-193a	2 / 206	201.65	< 1.00E-06
miR-320	1 / 88	172.27	< 1.00E-06
miR-615	1 / 87	170.34	< 1.00E-06
miR-27a	5 / 433	169.52	< 1.00E-06

miRNA	Number of reads Undiff / Diff	*P*-value
miR-518b	37 / 0	<1.00E-06
miR-520g	15 / 0	4.11E-03
miR-524*	6 / 0	1.03E-01
miR-363*	4 / 0	3.07E-01
miR-154	4 / 0	3.07E-01
miR-184	3 / 0	5.56E-01
miR-518c	3 / 0	5.56E-01
miR-512-3p	3 / 0	5.56E-01

miRNA	Number of reads Undiff / Diff	*P*-value
miR-218	0 / 157	< 1.00E-06
miR-22	0 / 114	< 1.00E-06
miR-196a	0 / 64	< 1.00E-06
miR-10b	0 / 44	< 1.00E-06
miR-181b	0 / 37	< 1.00E-06
miR-542-3p	0 / 27	< 1.00E-06
miR-192	0 / 22	< 1.00E-06
miR-181a	0 / 22	< 1.00E-06
miR-126	0 / 17	< 1.00E-06
miR-197	0 / 16	< 1.00E-06
miR-30a-3p	0 / 15	< 1.00E-06
miR-188	0 / 15	< 1.00E-06
miR-542-5p	0 / 9	5.80E-05
miR-491	0 / 9	5.80E-05
miR-28	0 / 9	5.80E-05
miR-30e-3p	0 / 8	1.71E-04
miR-148a	0 / 7	5.05E-04
miR-362	0 / 6	1.49E-03
miR-625	0 / 5	4.42E-03
miR-592	0 / 5	4.42E-03
miR-589	0 / 4	1.31E-02
miR-509	0 / 4	1.31E-02
miR-137	0 / 4	1.31E-02

Figure 3. Differential expression of known miRNAs between Undiff-hESC and Diff-hESC

(A) Expression ratios (percent of total reads in Undiff-hESC divided by percent of total reads in Diff-hESC) are shown for all known miRNAs that were detected in both Undiff-hESC and Diff-hESC datasets. Specific data pertaining to the 10 most differentially expressed miRNAs at both ends of the spectrum are displayed in the inset. (B) The absolute number of reads obtained for miRNAs that were solely detected in either Undiff-hESC or in Diff-hESC is shown.

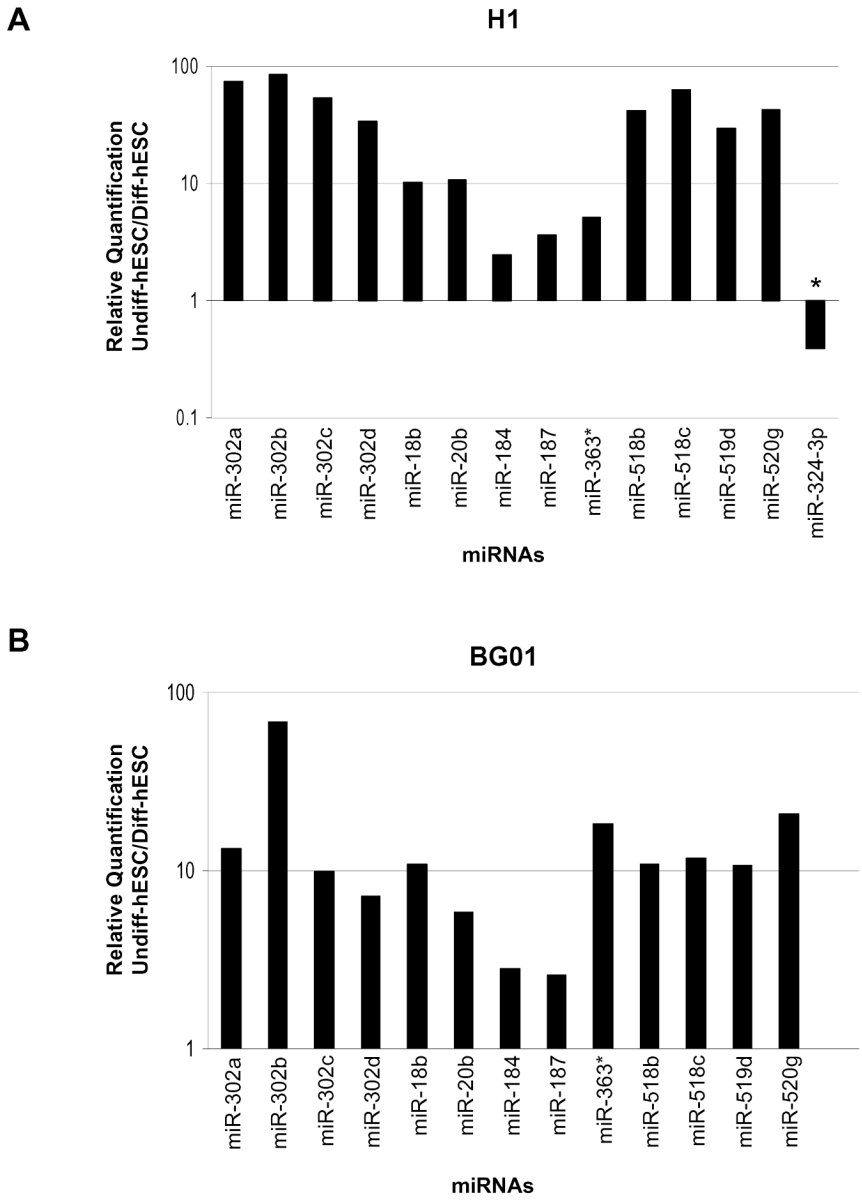


Figure 4. qRT-PCR (TaqMan) assays of selected miRNAs found by deep sequencing to be over-expressed in H1 Undiff-hESC relative to H1 Diff-hESC

Values on the y-axis (Relative Quantification) represent the relative expression of a given miRNA in Undiff-hESC relative to Diff-hESC as measured by qRT-PCR. (A) Results of qRT-PCR in H1 hESC were consistent with those from deep sequencing in 13 out of 14 cases. The one miRNA for which over-expression in H1 Undiff-hESC was not confirmed is indicated by an asterisk. (B) The 13 miRNAs confirmed to be over-expressed in H1 Undiff-hESC showed the same pattern of over-expression in BG01 hESC.

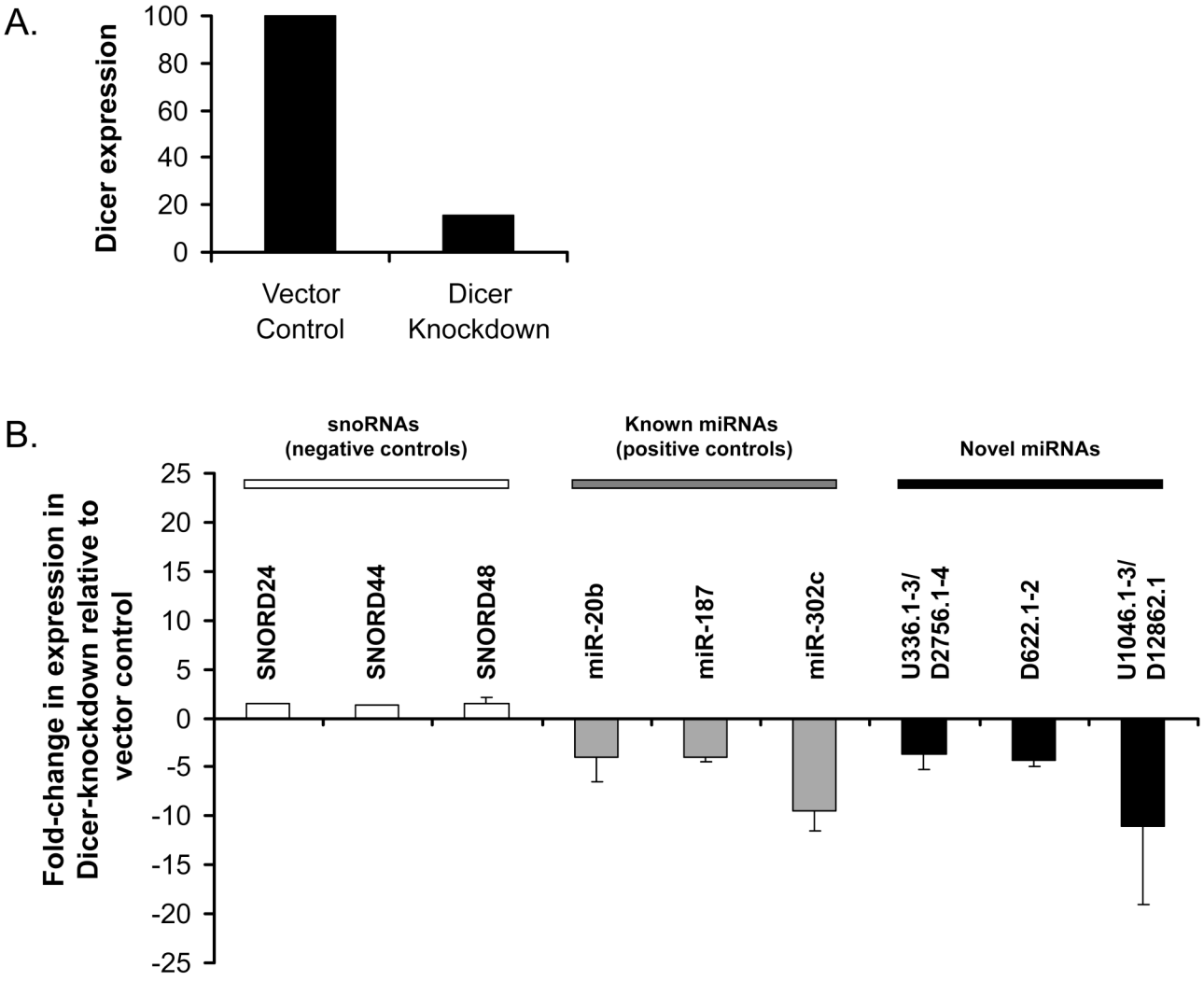
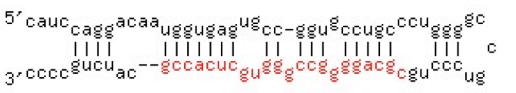
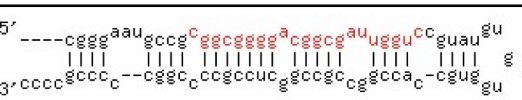
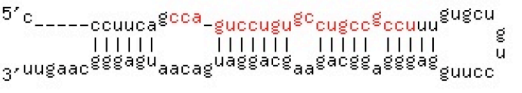
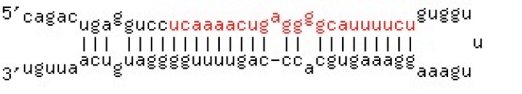
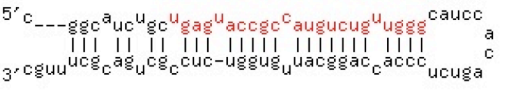


Figure 5. Dicer-dependent expression of novel miRNAs

(A) Dicer mRNA expression (measured by qRT-PCR) in vector control hESC vs. Dicer knockdown hESC is shown. The Relative Quantification method (RQ) was used and Dicer expression in vector control hESC is arbitrarily set to 100. As shown, Dicer transcript levels are reduced by 84% in the Dicer knockdown hESC compared to vector control hESC. (B) Custom TaqMan® Small RNA Assays were used to measure the expression of the three novel miRNAs indicated in both Dicer-knockdown and vector control H1 hESC. As negative controls, three snoRNAs (which are not expected to undergo Dicer processing) were measured using TaqMan qRT-PCR assays of similar design. The degree of expression of each small RNA (snoRNA or miRNA) in the Dicer-knockdown cells was compared to that in the vector control cells, and expressed as a fold-change relative to vector control. The expression of all three novel miRNAs was diminished significantly in Dicer-knockdown cells, whereas the snoRNAs did not show such a reduction and appeared in fact to show modestly elevated expression under Dicer-knockdown conditions. Three known miRNAs, serving as positive controls, showed the expected decrease in expression in Dicer knockdown hESC relative to vector control hESC.

A

Novel miRNA	Putative secondary structure	Number of reads Undiff/Diff	*P*-value
U739.1-6		26/0	0.000030
U336.1-3/D2756.1-4		57/7	0.000050
U755.1-4/D10092.1		18/1	0.006383
U1046.1-3/D12862.1		11/1	0.070936
U1123.1/D10070.1		11/1	0.070936

B

miRNA	Number of reads Undiff/Diff	*P*-value
miR-302d	37506/856	<1.0E-06
miR-302b	34782/497	<1.0E-06
miR-302c	10425/193	<1.0E-06
miR-302a	3914/415	<1.0E-06
miR-92b	735/28	<1.0E-06
miR-20b	692/29	<1.0E-06
miR-518b	37/0	<1.0E-06
miR-519d	19/1	0.003953
miR-520g	15/0	0.004111

Figure 6. Undiff-hESC-associated miRNAs

The set of five novel (A) and nine known (B) miRNAs that were at least 4-fold over-expressed in Undiff-hESC, or represented by at least 10 reads in the Undiff-hESC (in cases where the miRNA was not detected in Diff-hESC), is shown. *P*-values are calculated using Fisher's exact test. For each novel miRNA, the putative hairpin secondary structure generated by RNAshapes is shown. The red letters in the hairpin indicate the mature miRNA sequence obtained in this study. Putative secondary structures for all 13 of the novel miRNAs discovered in this study are given in Supplemental Table 4.

SeqID	Length	Coordinates	Strand	Sequence	Reads	Reads in star form	Shared seed region with animal miRNA?	Homologous hairpin		Intronic?
								primate	other vertebrate	
U336.1-3/D2756.1-4	21	chr11:61339257-61339277	m	CGGCGGGGACGGCGAUUGGUC	64		X	X		X
U739.1-6	22	chr19:1767168-1767189	m	CGCAGGGGCCGGGUGCUCACCG	26	2				X
D622.1-2	22	chr17:19188419-19188440	m	UUUCCGGCUCGCGUGGGUGUGU	19	5		X	X	X
U755.1-4/D10092.1	21	chr16:84332776-84332796	m	CCAGUCCUGUGCCUGCCGCCU	19					X
U1123.1/D10070.1	23	chrX:113904011-113904033	+	UGAGUACCGCCAUGUCUGUUGGG	12	1		X	X	X
U1046.1-3/D12862.1	22	chr19:58867044-58867065	+	UCAAAACUGAGGGGCAUUUUCU	12			X		
U3863.1-2	22	chrX:113792320-113792341	+	UACCCAGAGCAUGCAGUGUGAA	4			X	X	X
D10711.1/U3615.1	22	chr12:55874603-55874624	+	CCUCACACCUGCCUCGCCCCCC	4		X			X
U2732.1	22	chr6:166842842-166842863	m	UCUGCCCCCUCCGCUGCUGCCA	3		X		X	X
U10426.1-2	21	chr22:18616668-18616688	m	UGCAGGACCAAGAUGAGCCCU	3			X	X	X
U2971.1	22	chr20:62043308-62043329	m	CCCUGUGCCCGGCCCACUUCUG	3	1		X		X
[1]D15980.1/U18158.1	20	chr10:21825511-21825530	m	CCCCAGGGCGACGCGGCGGG	2	3		X	X	

[1]This sequence was chosen as the dominant sequence from the 5'-3' pair because it had multiple homologous hairpins in other vertebrates, and therefore a conserved seed region, where its star form (though more abundant by 1 read) did not.