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

Ryan D. Morin,[1] Michael D. O'Connor,[2] Malachi Griffith,[1] Florian Kuchenbauer,[2] Allen Delaney,[1] Anna-Liisa Prabhu,[1] Yongjun Zhao,[1] Helen McDonald,[1] Thomas Zeng,[1] Martin Hirst,[1] Connie J. Eaves,[2,3,4] and Marco A. Marra[1,3,4]

[1]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada; [2]Terry Fox Laboratory, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada

MicroRNAs (miRNAs) are emerging as important, albeit poorly characterized, regulators of biological processes. Key to further elucidation of their roles is the generation of more complete lists of their numbers and expression changes in different cell states. Here, we report a new method for surveying the expression of small RNAs, including microRNAs, using Illumina sequencing technology. We also present a set of methods for annotating sequences deriving from known miRNAs, identifying variability in mature miRNA sequences, and identifying sequences belonging to previously unidentified miRNA genes. Application of this approach to RNA from human embryonic stem cells obtained before and after their differentiation into embryoid bodies revealed the sequences and expression levels of 334 known plus 104 novel miRNA genes. One hundred seventy-one known and 23 novel microRNA sequences exhibited significant expression differences between these two developmental states. Owing to the increased number of sequence reads, these libraries represent the deepest miRNA sampling to date, spanning nearly six orders of magnitude of expression. The predicted targets of those miRNAs enriched in either sample shared common features. Included among the high-ranked predicted gene targets are those implicated in differentiation, cell cycle control, programmed cell death, and transcriptional regulation.

[Supplemental material is available online at www.genome.org.]

MicroRNAs (miRNAs) are short RNA molecules, 19–25 nucleotides (nt) in length, that derive from stable fold-back substructures of larger transcripts (Cai et al. 2004). In most cases, primary miRNA transcripts (pri-miRNAs) are cleaved by a complex of Drosha (currently known as RNASEN) and its cofactor DGCR8, 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 Dicer1 (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 RNA-induced silencing complex (RISC), producing a biologically active miRNA and an inactive miRNA* (O'Toole et al. 2006).

Once assembled within the RISC protein complex, miRNAs can elicit down-regulation of target genes by blocking their translation, inducing EIF2C2 (formerly AGO2) mediated degradation, or potentially inducing deadenylation (Aravin and Tuschl 2005; Giraldez et al. 2006). In animals, miRNA–target interactions occur through semicomplementary base pairing, usually within the 3′ untranslated region (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) (Lewis et al. 2005; Grimson et al. 2007). 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 (Pfeffer et al. 2005; Cummins et al. 2006). Nevertheless, direct small RNA sequencing

[3]These authors contributed equally to this work.
[4]Corresponding authors.
E-mail mmarra@bcgsc.ca; fax (604) 675-8178.
E-mail ceaves@bccrc.ca; fax (604) 877-0712.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.7179508.

18:610–621 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

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 and Tuschl 2005; Landgraf et al. 2007). In contrast with 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, 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. Next-generation miRNA profiling has already been realized in a few organisms (Rajagopalan et al. 2006; Fahlgren et al. 2007; Kasschau et al. 2007; 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 greater depth than current competing technologies (Berezikov et al. 2006b), yielding up to several million sequences from a single flow cell lane (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 nonadherent 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 and Tuan 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 (Sato et al. 2003; Abeyta et al. 2004; Bhattacharya et al. 2004,2005; Boyer et al. 2005; Hirst et al. 2007), but to date, very little is known about the specific miRNAs that may play roles in their pluripotency or differentiation (Suh et al. 2004).

Results

Sequencing and annotation of small RNAs

Sequencing of small RNA libraries yielded 6,147,718 and 6,014,187 37-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. These reads were mapped to the genome by forcing perfect alignments beginning at the first nucleotide and retaining the longest region of each read that could be aligned to the reference genome, along with all alignment positions. After mapping, a total of 766,199 (hESC) and 724,091 (EB) unique error-free trimmed small RNA sequences were represented by 4,351,479 and 3,886,865 reads. Genomic positions of all non-singleton small RNA sequences can be viewed in our genome browser (http://microrna.bcgsc.ca/cgi-bin/gbrowse/hg18). Any sequence observed more than three times was considered a reliable representation of a small RNA molecule. These sequences were annotated as one of the known classes of small RNA genes or degradation fragments of larger noncoding RNAs (see Methods). Briefly, sequences were annotated based on their overlap with publicly available genome annotations including miRNAs, rRNAs, tRNAs, other small RNAs, and genomic repeats. Sequences deriving from 334 distinct miRNA genes were identified. The miRNAs were the most abundant class of small RNAs on average, but spanned the entire range of expression, with sequence counts up to ~120,000 (Fig. 1A). The PIWI-associated RNAs (piRNAs), previously thought to exist only in mammalian germline cells, were also observed at relatively low levels. Searching the mapped sequences from the hESC and EB libraries against known piRNAs identified 460 and 378 unique sequences (including singletons) with counts of 9012 and 4606, respectively (median count = 2). These sequences correspond to 118 and 94 distinct known piRNAs (Supplemental Table 1). The sampling depth provided by one lane of the Illumina sequencing apparatus extends the dynamic range of miRNA expression within a cell, previously thought to span only ~3 orders of magnitude (Berezikov et al. 2006a).

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 (e.g., Fig. 2) 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 (Cummins et al. 2006; Ruby et al. 2006; Landgraf et al. 2007). It appears that much of the isomiR variability can be explained by variability in either Dicer1 or Drosha cleavage positions within the pre-miRNA hairpin (Fig. 2). All isomiRs of this class including their sequences and abundances are summarized in Supplemental Table 2. Notably, our data show that choosing a different isomiR sequence for measuring miRNA expression level can affect the ability to detect differential miRNA expression. 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 (Supplemental Table 3; Supplemental Methods). In 107 cases, this most abundant sequence did not correspond exactly to the current miRBase reference sequence (Supplemental Methods). 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. Following the most recent update to miRBase, far more of the most abundant sequences in our libraries corresponded with the updated miRBase reference, pointing toward the latter explanation.

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 among the

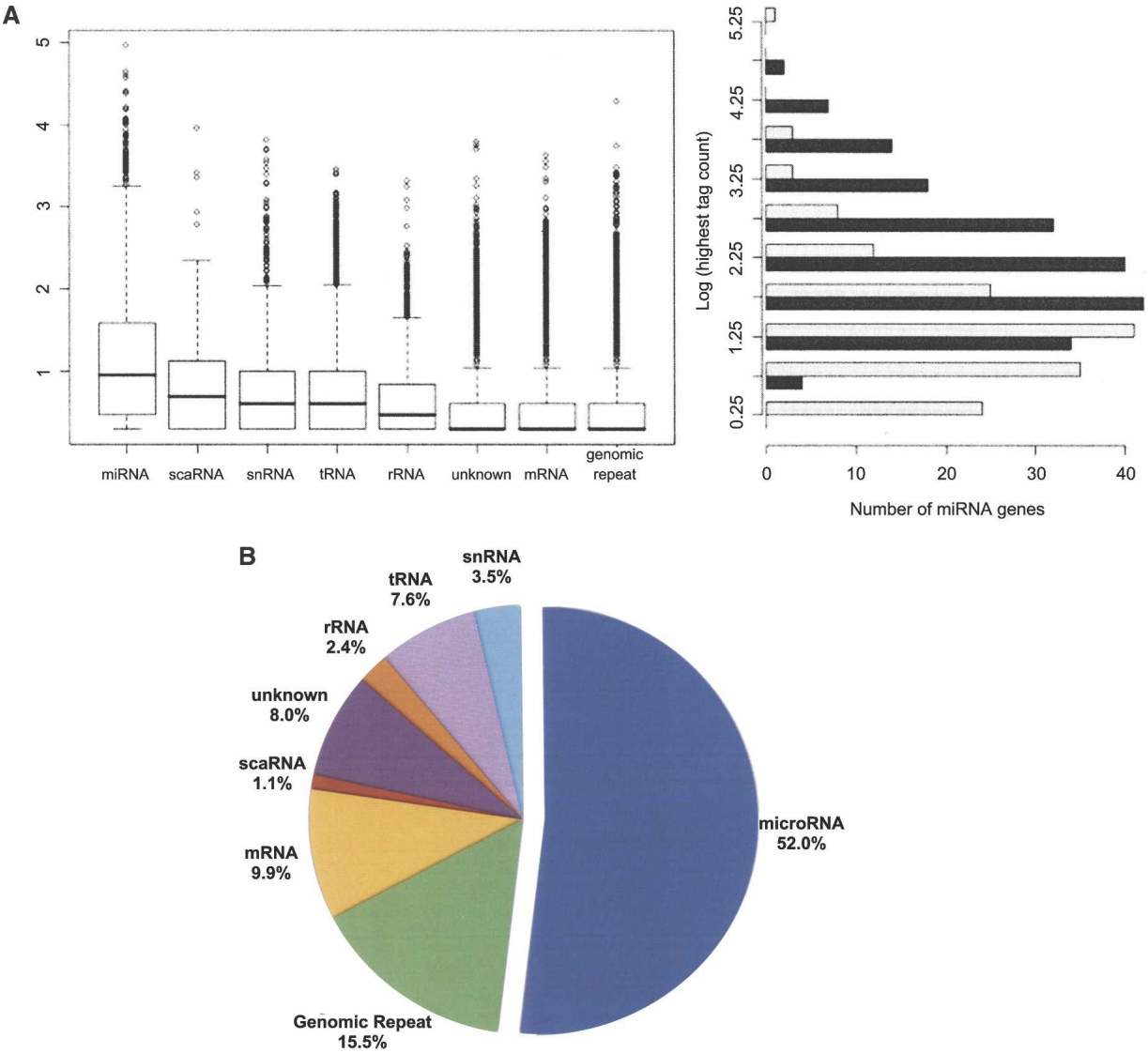

Figure 1. Distributions of sequence counts for the different classes of small RNAs. (*A*) The box plot (*left*) shows the relative expression levels of sequences in each of the eight major classes of small RNAs ($\log_{10}$ transformation) in the hESC small RNA library. miRNAs were the most highly expressed 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 (black) is roughly normal (mean = 2.32, median = 2.17), representing a tag count of 1743 and 151, respectively. The miRNAs detected in at least one of the libraries but not significantly differentially expressed are shown in light gray 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 eight 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).

miRNAs were single-nucleotide 3′ extensions, which were observed in multiple isomiRs from nearly all observed miRNA genes (316 miRNA 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 among 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. Figure 2 shows an example of 3′ modification, and Supplemental Table 4 summarizes the extensions that were most prevalent in the two libraries. 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). 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

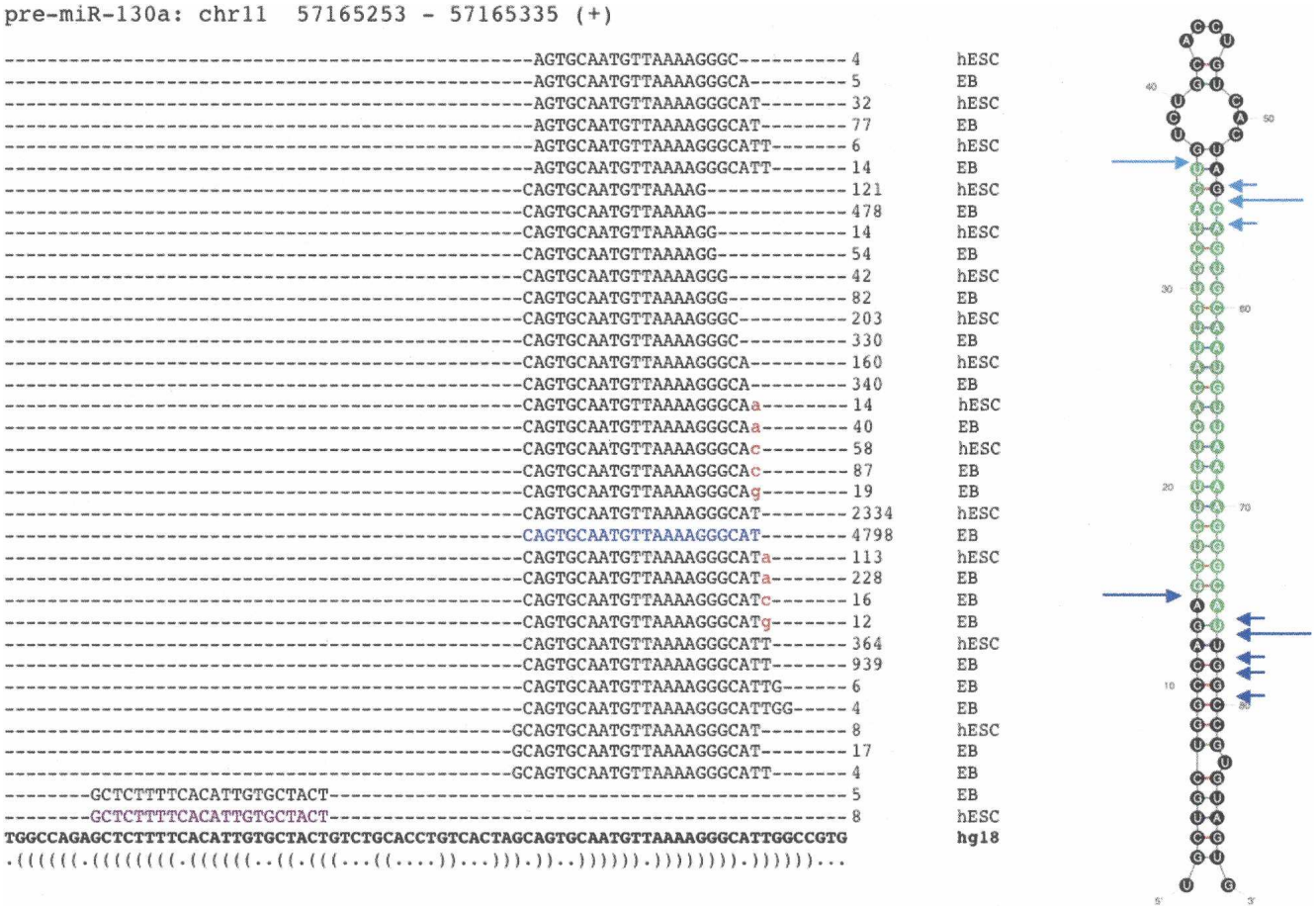

Figure 2. The repertoire of isomiRs and 3′ modifications of hsa-miR-130a. A diverse variety of isomiRs were observed for many of the known and novel miRNAs. Sequences representing the miR* were also commonly observed (highlighted in purple). The reference miRNA sequence from miRBase was, in many cases, not the most frequently observed isomiR (shown in blue). In this example, the most abundant miR* did not correspond to the current miRBase entry, although within the predicted pre-miRNA structure it derives from the correct 2-nt offset position relative to the reference miRNA (*far right*). Although variation at the 3′ end is generally much more common than at the 5′ end, target preference of the 5′ isomiRs may differ. Sequences with evidence of 3′ additions of nucleotides (red) were common, with certain miRNAs more heavily modified than others (summarized in Supplemental Table 4). The predicted structure of the pre-miRNA is represented at the *bottom* in dot-bracket notation, and as a graphic (*far right*). Experimentally inferred Drosha and Dicer1 cleavage positions are indicated in dark and light blue arrows, respectively. Large arrows represent cleavage sites for the most abundant isomiR whereas small arrows indicate those for the other isomiRs.

previous study (Landgraf et al. 2007) where the same modifications were also observed in the orthologous mouse miRNAs. Library preparation involves a ligation reaction mediated by T4 ligase in the presence of ATP (see Methods), which might explain the addition of adenosines. However, to our knowledge this cannot explain the addition of other nucleotides nor the apparent preference for 3′ extension of some miRNAs over others.

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 (119 total from 100 miRNA genes) was less than the total number of miRNAs identified as differentially expressed using the most abundant isomiR for each pre-miRNA arm (171 total from 145 miRNA genes) (Table 1; Supplemental Table 5). All 100 of the miRNA genes identified as differentially expressed, based only on the sequences corresponding perfectly with those in miRBase, were among the 145 identified by the most 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 (22 total). Taken together, this reinforces the concept that the most abundant, rather than the reference miRNA sequence, is the most useful for identifying differentially expressed miRNAs.

Novel microRNA genes

We sought to identify novel miRNA genes among the unclassified 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 unannotated 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

Table 1. Top 20 miRNAs differentially expressed between the hESC and EB libraries

microRNA ID	Pre-miRNA arm (5-p or 3-p)	Most abundant sequence (isomiR)	hESC count	EB count	*P*-value	Fold change	miRBase sequence most abundant isomiR
miR-199a	3-p	ACAGTAGTCTGCACATTGGTTA	1110	13,163	0.00	11.86	Yes
miR-372	3-p	AAAGTGCTGCGACATTTGAGCGT	1388	13,653	0.00	9.84	Yes
miR-122	5-p	TGGAGTGTGACAATGGTGTTTG	436	2565	0.00	5.88	Yes
miR-152	3-p	TCAGTGCATGACAGAACTTGG	622	3028	0.00	4.87	Yes
miR-10a	5-p	TACCCTGTAGATCCGAATTTGT	948	3887	0.00	4.10	No
let-7a	5-p	TGAGGTAGTAGGTTGTATAGTT	11,902	2951	0.00	4.03	Yes
miR-302a	5-p	TAAACGTGGATGTACTTGCTTT	36,800	9917	0.00	3.71	No
miR-222	3-p	AGCTACATCTGGCTACTGGGTCTC	4719	1331	0.00	3.55	No
miR-340	5-p	TTATAAAGCAATGAGACTGATT	2247	7198	0.00	3.20	Yes
miR-363	3-p	AATTGCACGGTATCCATCTGTA	5775	17,912	0.00	3.10	Yes
miR-21	5-p	TAGCTTATCAGACTGATGTTGAC	39,818	21,003	0.00	1.90	No
miR-221	3-p	AGCTACATTGTCTGCTGGGTTTC	16,275	8716	0.00	1.87	Yes
miR-26a	5-p	TTCAAGTAATCCAGGATAGGCT	4892	8530	0.00	1.74	Yes
miR-26b	5-p	TTCAAGTAATTCAGGATAGGTT	1003	2957	1.39×10^{-278}	2.95	No
miR-130a	3-p	CAGTGCAATGTTAAAAGGGCAT	2334	4798	2.20×10^{-265}	2.06	Yes
miR-594	5-p	ATGGATAAGGCATTGGC	1717	211	1.96×10^{-253}	8.14	No
miR-302b	3-p	TAAGTGCTTCCATGTTTTAGTAG	15,169	8855	1.39×10^{-213}	1.71	Yes
miR-744	5-p	TGCGGGGCTAGGGCTAACAGCA	4166	1516	4.17×10^{-213}	2.75	Yes
miR-30d	5-p	TGTAAACATCCCCGACTGGAAGCT	2798	4988	3.29×10^{-205}	1.78	No
miR-146b	5-p	TGAGAACTGAATTCCATAGGCTGT	703	2075	2.27×10^{-196}	2.95	No

were either randomly derived fragments of various RNAs as well as unidentified piRNAs or regulatory siRNAs (Berezikov et al. 2006b; Watanabe et al. 2006). To identify candidate novel miRNAs among these, we employed both in-house and publicly available algorithms (see Methods). First, small hot-spot regions of the genome with alignments to more than one small RNA in our library were identified, followed by folding flanking genomic sequence with RNALfold. In parallel, we classified these folded structures using a publicly available miRNA classifier (MiPred) and an in-house classifier based on a support vector machine (SVM). 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 Supplemental Table 6, comprises 83 unique miRNA sequences from up to 104 distinct genes. Compared with other miRNAs in our data, these miRNAs exhibited modest expression (mean count = 45), perhaps indicating that most of them may not perform significant functions in these cell types. However, 23 exhibited significant differential expression between the two libraries and are thus potentially biologically important in the cells in which they were present (Table 2).

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

Table 2. Putative novel miRNAs exhibiting significant differential expression

Most abundant sequence	Name	hESC count	EB count	Fold change	Seed	Notes
TTCATTCGGCTGTCCAGATGTA	miR-1298	1269	774	1.64	UCAUUCG	—
TCCCTGTTCGGGCGCCA	miR-12754b	670	1302	1.94	CCCUGUU	—
GCATGGGTGGTTCAGTGG	miR-1308	407	52	7.83	CAUGGGU	Shares seed with miR-885
AATGGATTTTTGGAGCAGG	miR-1246	374	245	1.53	AUGGAUU	—
ACTCGGCGTGGCGTCGGTCGTG	miR-1307	228	50	4.56	CUCGGCG	—
CCTCAGGGCTGTAGAACAGGGCT	miR-1266	163	45	3.62	CUCAGGG	—
TATTCATTTATCCCCAGCCTACA	miR-664	126	67	1.88	AUUCAUU	Shares seed with miR-181/miR-664
GTGGGGGAGAGGCTGTC	miR-1275	103	17	6.06	UGGGGGG	—
TAAGTGCTTCCATGCTT	miR-302e	84	131	1.56	AAGUGCU	Shares seed with miR-302 family
GATGATGATGGCAGCAAATTCTGAAA	miR-1272	81	27	3	AUGAUGA	—
TTGCAGCTGCCTGGGAGTGACTTC	miR-1301	63	13	4.85	UGCAGCU	—
CTGGACTGAGCCGTGCTACTGG	miR-1269	44	161	3.66	UGGACUG	—
TGGATTTTTGGATCAGGGA	miR-1290	40	85	2.13	GGAUUUU	—
CGGGCGTGGTGGTGGGGG	miR-1268	35	2	17.5	GGGCGUG	—
ACGTTGGCTCTGGTGGTG	miR-1306	35	11	3.18	CGUUGGC	—
ATGGATAAGGCTTTGGCTT	miR-1261	31	2	15.5	UGGAUAA	Seed is enriched in hESCs
AGCCTGGAAGCTGGAGCCTGCAGT	miR-1254	25	3	8.33	GCCUGGA	—
AGGAGGAATTGGTGCTGGTCTT	miR-766*	25	2	12.5	GGAGGAA	Shares seed with miR-923
GTCCCTGTTCAGGCGCCA	miR-1274a	24	66	2.75	UCCCUGU	—
TACGTAGATATATATGTATTTT	miR-1277	20	54	2.7	ACGUAGA	—
TGCTGGATCAGTGGTTCGAGTC	miR-1287	16	57	3.56	GCUGGAU	—
ATGGGTGAATTTGTAGAAGGAT	miR-1262	7	45	6.43	UGGGUGA	—
ACCCGTCCCGTTCGTCCCCGGA	miR-1247	4	53	13.25	CCCGUCC	—

of predicted targets of miRNAs with identical seeds often overlaps considerably. We sought to determine whether miRNAs sharing identical seeds demonstrated coexpression. We identified 1009 distinct seed sequences when considering all unique isomiRs from both libraries. Many of these represented the noncanonical seeds of isomiRs arising from variation at their 5′ end. One hundred fourteen 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 3; Supplemental Table 7).

As expected, we found that many of the seeds with the largest changes in relative levels between the two libraries corresponded with the differentially expressed miRNAs in Table 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-372 (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 coexpressed 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 (see 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 examining the significant Gene Ontology (GO) "biological process" classifications that are over-represented among 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 the biological processes were enriched only in the predicted targets of hESC-enriched or EB-enriched miRNAs. For example, genes involved in programmed cell death were enriched among the predicted targets of hESC-enriched miRNAs while those involved in cell proliferation were enriched among the predicted targets of EB-enriched miRNAs (Figure 3).

Discussion

These data highlight the potential of a new massively parallel sequencing strategy to profile miRNA expression in differentiating hESCs. Between the two libraries, we identified 171 differentially expressed known and 23 novel miRNAs corresponding with

Table 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	17×	0	**miR-199a**
CAGUAGU	1801	22,288	12.4×	0	miR-199a
ACAGUAG	241	2749	11.4×	0	**miR-199a**
UUAAACG	3452	398	8.7×	0	**miR-302a-5p**
CUUAAAC	16,762	2332	7.2×	0	miR-302a-5p
GGAGUGU	846	5216	6.2×	0	miR-122
AGUGCUG	2484	12,641	5.1×	0	miR-372, miR-512
ACCCUGU	1503	5689	3.8×	0	miR-10
AAACGUG	55,712	16,461	3.4×	0	**miR-302a-5p,** miR-424
UAUAAAG	3002	9702	3.2×	0	miR-340
GAGAACU	1547	4976	3.2×	0	miR-146
UUGCACG	1845	5561	3.0×	0	miR-363
GAGGUAG	29,280	10,258	2.9×	0	let-7a–let-7i
CAGUGCA	3596	8559	2.4×	0	miR-148, miR-152, **miR-130**
AAAGCUG	16,448	7613	2.2×	0	miR-320
AGUGCAA	4463	9594	2.1×	0	miR-454, miR-301, miR-130
UCAAGUA	6898	14,089	2.0×	0	miR-26
GCUACAU	47,591	24,280	2.0×	0	miR-221, miR-222
GUAAACA	8789	15,479	1.8×	0	miR-30
GAGGGGC	17,586	10,518	1.7×	0	miR-423/miR-885
AGCUUAU	65,376	39,424	1.7×	0	miR-21, miR-590
CUGGACU	17,819	25,378	1.4×	0	miR-378
GCAGCAU	153,193	187,218	1.2×	0	miR-103, miR-107, miR-885
GCGGGGC	5220	1956	2.7×	8.76×10^{-305}	miR-744
CUCAAAC	4306	8173	1.9×	1.62×10^{-296}	**miR-92a-5p,** miR-371

miRNAs indicated in boldface type are cases in which one of their noncanonical isomiRs bears this seed.

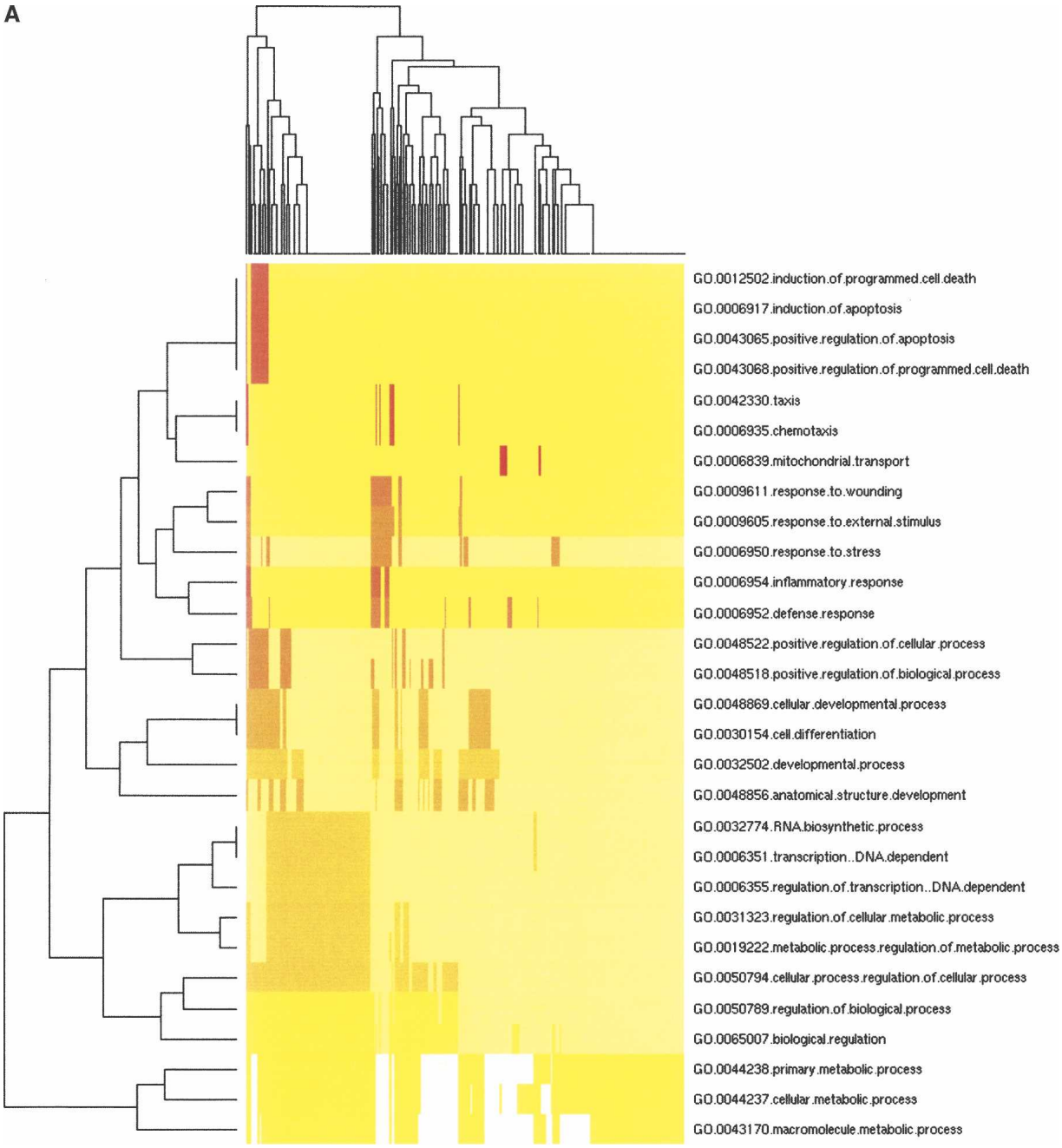

Figure 3. (Continued on next page)

168 distinct miRNA genes. This is a striking advance in contrast with 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, among the 27 listed mouse miRNAs with identifiable human orthologs, 14 exhibited expression patterns consistent with the hESC results presented here, while many of the others exhibited insignificant changes in expression. Further, of the additional differential miRNAs reported here (Table 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 with EBs, our method detected only two miRNAs that appeared exclusively in hESCs (miR-486 and miR-187); this is likely a consequence of the increased sampling depth of the method used here. On the other hand, we found 14 miRNAs in EBs (all with counts of at least 10) that were not observed in hESCs (Supplemental Table 5). 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

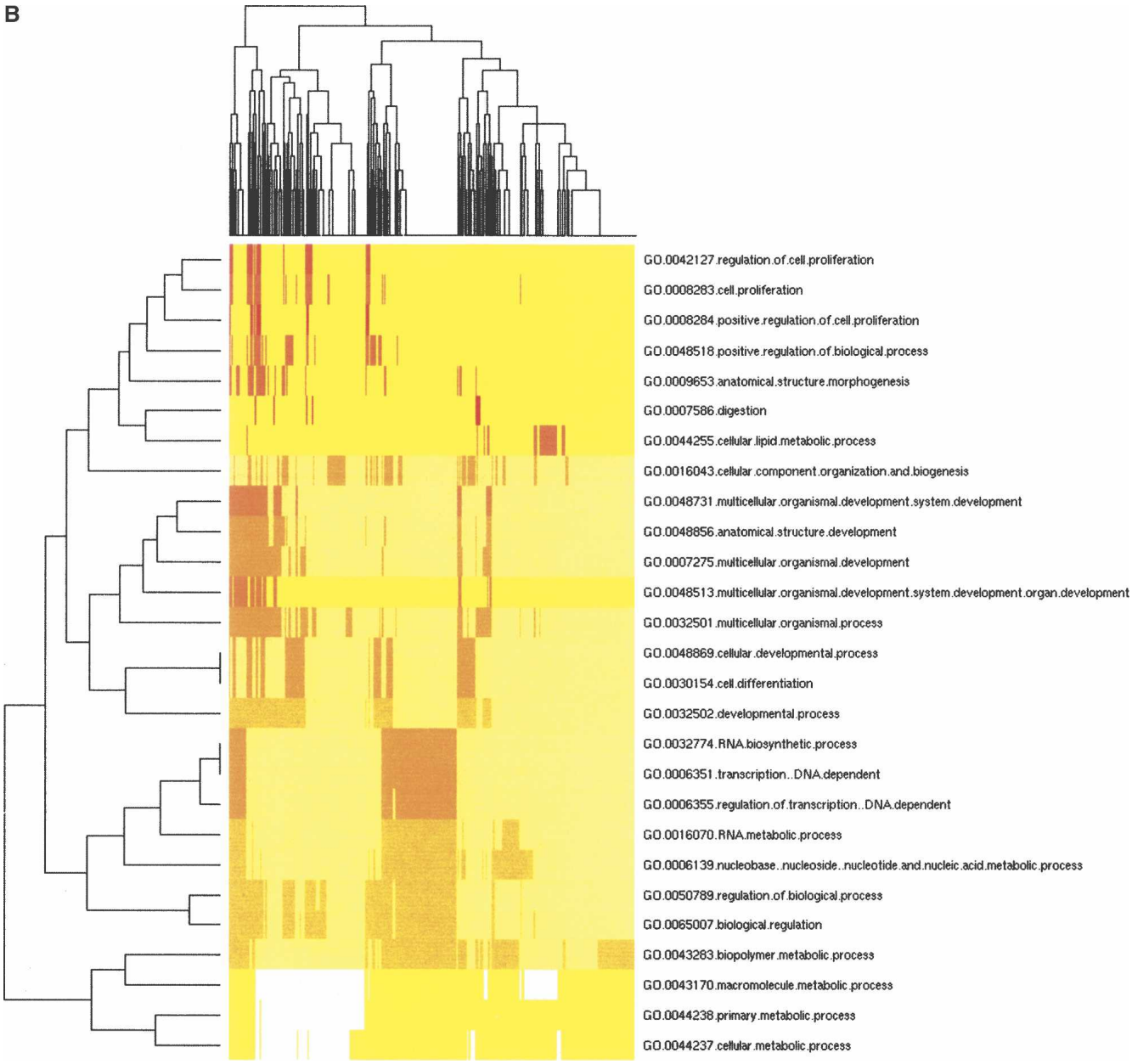

Figure 3. Clustering of over-represented Gene Ontology (GO) classes in predicted targets of differential microRNAs. Shown are heat map representations of GO (Ashburner et al. 2000) terms over-represented among predicted cooperative targets (*Y*-axis) of hESC-enriched miRNAs (*A*) and EB-enriched miRNAs (*B*). All genes with statistically over-represented GO annotations were included ($P < 0.01$, *X*-axis) as identified by GoStat (Beissbarth and Speed 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.

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 most of 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 may be due to the dominance of the highly expressed miRNAs in small RNA libraries as well as difficulties associated with library normalization (Shivdasani 2006; Landgraf et al. 2007). Here, we present 104 novel miRNA genes that have passed multiple levels of annotation criteria including expression and

folding verification. 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 (RF) algorithm to classify miRNAs based on structural properties (Jiang et al. 2007). Our custom classifier, which uses an SVM, assigns a classification score based on a separate set of structural descriptors as well as data specific to this sequencing method. As a result, this classifier has provided the best classification accuracy of any machine learning-based miRNA method (see Methods).

At least 23 of the novel miRNAs identified in this study show evidence of regulated expression during hESC differentiation (Table 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). 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 reside in the same range of expression as these novel miRNAs (Fig. 1A, black bars), 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 among other tissues where they are functionally important.

In addition to novel miRNA genes, this study has aided in identifying a diverse population of variants of known miRNAs, collectively termed isomiRs. The functional implications of the widespread 3′ modifications (Supplemental Table 4) 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 with those from diverse human and mouse tissues, it is evident that the nature of the added nucleotide is nonrandom and evolutionarily conserved.

Some of the remaining noncanonical isomiRs, which appear to derive from variability in Dicer1 and Drosha cleavage sites, were quite abundant. By including all isomiRs and grouping those that share an identical seed sequence, we revealed 222 of seeds that were represented in significantly different quantities these two libraries (Table 3; Supplemental Table 7). As the noncanonical 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 noncanonical 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 (Fig. 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 miRNAs exhibiting the most significant differences between hESCs and EBs appear to regulate genes involved in transcriptional regulation, differentiation and development.

Methods

Small RNA library preparation

Undifferentiated H9 hESCs (Hirst et al. 2007) were cultured on Matrigel (BD Biosciences) coated dishes in maintenance medium consisting of Dulbecco's Modified Eagle Medium (DMEM)/F12 containing 20% Knockout Serum Replacer (Invitrogen), 0.1 mM β-mercaptoethanol, 0.1 mM nonessential 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 were cultured in nonadherent dishes (BD Biosciences) for up to 30 d 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) treated total RNA was size fractionated on a 15% tris-borate-EDTA (TBE) urea polyacrylamide gel and a 15–30 base pair fraction was 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 ethanol and 3 µL of mussel glycogen (5 mg/mL; Invitrogen).

After washing with 75% ethanol, the pellets were allowed to air dry at 25°C and pooled in diethylpyrocarbonate (DEPC) water. The 5′ RNA adapter (5′-GUUCAGAGUUCUACAGUCCGAC GAUC-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 2× formamide loading dye. The ligated RNA was size fractionated on a 15% TBE urea polyacrylamide gel, and a 40–60 base pair fraction was 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 resuspension in DEPC-treated water.

The 3′ RNA adapter (5′-pUCGUAUGCCGUCUUCUGC UUGidT-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 µL of 2× formamide loading dye. Ligated RNA was size fractionated on a 10% TBE urea polyacrylamide gel, and the 60–100 base pair fraction was excised. The RNA was eluted from the polyacrylamide gel and precipitated from the gel as described above and resuspended 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′-CAAGCAGAAGACGGCATACGA-3′) following the manufacturer'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′-CAAGCAGAAGACG GCATACGA-3′; 5′-AATGATACGGCGACCACCGA-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 ethanol, 133 µL of 7.5 M ammonium acetate, and 3 µL of mussel glycogen (20 mg/mL; Invitrogen). After washing with 75% ethanol, 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 and diluted to 10 nM for sequencing on the Illumina 1G.

Small RNA genome mapping and quantification

Virtually no reads aligned to the genome after position 28, so we trimmed all reads at 30 nt to reduce the number of unique sequences. 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 Mega BLAST (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 subsequence 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.

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 and Claverie (1997). This approach was developed for analysis of digital gene expression profiles and accounts for the sampling variability of tags with low counts. 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.

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 with 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, CDBox, scRNA, snoRNA, snRNA, srpRNA, genomic repeat, or known transcript (exonic or intronic). All mapped sequences were also searched against the currently known human piRNAs (Aravin et al. 2007) using sequences retrieved from piRNABank (Lakshmi and Agrawal 2007). As piRNAs are considerably longer than miRNAs, we used the first 18 nt of the piRNAs to search for perfect matches among our sequences. Sequences that did not overlap any of these annotations (in any one of their genomic positions) or that 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.

Novel microRNA prediction

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"). Three hundred nucleotides of genomic sequence flanking each seed was extracted, reverse-complemented where appropriate, and folded using RNALfold (Hofacker 2003), which identifies locally stable substructures within a query RNA sequence. The largest unbranched fold-back substructure 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 an 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 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 minimum folding energy index (MFEI), adjusted minimum folding energy (AMFE) (Zhang et al. 2006), normalized pairing propensity (Ng Kwang Loong and Mishra 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 RF methods 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).

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 coexpressed miRNAs (either hESC-enriched or EB-enriched) were identified as potential cooperative targets. To compensate for potential bias, genes with numerous predicted miRNA target sites 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 target sites for coexpressed miRNAs by the total number of target sites for that gene. 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 (Fig. 3).

GO analysis

GoStat (Beissbarth and Speed 2004) was employed for identification of significantly enriched "biological process" 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 from the GoStat output. Genes and GO terms were clustered using hierarchical clustering in R. Heat maps were created in R using the heatmap function (default parameters).

Acknowledgments

We thank the staff at Illumina, Inc. for technical assistance. We also thank Rhonda Oshanek of the Genome Sciences Centre for her contribution in editing the draft of this manuscript. This project was funded, in part, by the Canadian Stem Cell Network with cofunding from Stem Cell Technologies, Inc as well as the BC Cancer Foundation. M.A.M. is a Michael Smith Foundation for Health Research Senior Scholar. R.D.M. receives stipends from the Michael Smith Foundation for Health Research and the Canadian Institutes for Health Research. M.O.C. is a recipient of a StemCell Technologies-sponsored Canadian Institutes of Health Research (CIHR) Industrial Postdoctoral Fellowship. F.K. is supported by the Deutsche Forschungsgemeinschaft Germany (grant no. Ku 2288/1-1).

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:** 601–608.

Aravin, A. and Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.* **579:** 5830–5840.

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.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

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

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:** S2–S7.

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:** 1375–1377.

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:** 215–217.

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:** 2956–2964.

Bhattacharya, B., Cai, J., Luo, Y., Miura, T., Mejido, J., Brimble, S.N., Zeng, X., Schulz, T.C., Rao, M.S., and Puri, R.K. 2005. Comparison of the gene expression profile of undifferentiated human embryonic stem cell lines and differentiating embryoid bodies. *BMC Dev. Biol.* **5:** 22. doi: 10.1186/1471-213X-5-22.

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:** 947–956.

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:** 1957–1966.

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* **18:** 316–327.

Cummins, J.M., He, Y., Leary, R.J., Pagliarini, R., Diaz, A.L., Jr., Sjoblom, T., Barad, O., Bentwich, Z., Szafranska, A.E., Labourier, E., et al. 2006. The colorectal microRNAome. *Proc. Natl. Acad. Sci.* **103:** 3687–3692.

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:** 2875–2883.

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. doi: 10.1371/journal.pone.0000219.

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:** 75–79.
Griffiths-Jones, S. 2006. miRBase: The microRNA sequence database. *Methods Mol. Biol.* **342:** 129–138.
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. doi: 10.1186/gb-2007-8-6-r113.
Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31:** 3429–3431.
Houbaviy, H.B., Murray, M.F., and Sharp, P.A. 2003. Embryonic stem cell-specific MicroRNAs. *Dev. Cell* **5:** 351–358.
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:** 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–W344. doi: 10.1093/nar/gkm368.
John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human microRNA targets. *PLoS Biol.* **2:** e363. doi: 10.1371/journal.pbio.0020363.
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. doi: 10.1371/journal.pbio.0050057.
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:** 1137–1140.
Lakshmi, S. and Agrawal, S. 2007. piRNABank: A web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* (in press). doi: 10.1093/nar/gkm696.
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:** 1401–1414.
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:** 95–98.
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:** 376–380.
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–D735. doi: 10.1093/nar/gkj077.
Ng Kwang Loong, S. and Mishra, S.K. 2007. Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification. *RNA* **13:** 170–187.
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:** 3338–3344.
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. doi: 10.1371/journal.pcbi.0020033.
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:** 269–276.
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:** 6130–6135.
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:** 3407–3425.
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:** 1193–1207.
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:** 404–413.
Shivdasani, R.A. 2006. MicroRNAs: Regulators of gene expression and cell differentiation. *Blood* **108:** 3646–3653.
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:** 1343–1356.
Song, L. and Tuan, R.S. 2006. MicroRNAs and cell differentiation in mammalian development. *Birth Defects Res. C Embryo Today* **78:** 140–149.
Strauss, W.M., Chen, C., Lee, C.T., and Ridzon, D. 2006. Nonrestrictive developmental regulation of microRNA gene expression. *Mamm. Genome* **17:** 833–840.
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:** 488–498.
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:** 380–385.
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:** 1732–1743.
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. doi: 10.1186/1471-2105-6-310.
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. doi: 10.1186/gb-2007-8-6-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:** 246–254.
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:** 1419–1425.

Received September 25, 2007; accepted in revised form November 27, 2007.