Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs


1. Duke Institute for Genome Sciences and Policy, Duke University, Durham, NC, USA
2. University of North Carolina, Chapel Hill NC, USA
3. The University of Hong Kong, Queen Mary Hospital, Hong Kong, China
4. Indiana University, Indianapolis IN, USA
5. Duke University Medical Center and 6V.A. Medical Center, Durham NC 27710, USA
7. Emory University, Atlanta GA, USA
8. Imperial College, London, UK
9. Northwestern University, Chicago IL, USA
10. ProHealth Care Regional Cancer Center, Waukesha WI, USA
11. Baylor University Medical Center, Dallas TX, USA
12. Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond VA, USA.
13. Autoimmunity Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda MD, USA

* Contributed equally to this work

Address correspondence to: Sandeep S. Dave, M.D.
101 Science Drive, Duke University, Box 3382, Durham, North Carolina 27710, USA
(919) 681-1922 (office); (919) 668-4777 (Fax);
sandeep.dave@duke.edu (email)
Keywords:
microRNA, miRNA, B cell differentiation, microarray, leukemia, lymphoma, deep sequencing, EBV-transformation, naive B cells, germinal center B cells, memory B cells, plasma cells, high throughput sequencing, acute lymphoblastic lymphoma, mantle cell lymphoma, follicular lymphoma, marginal zone lymphoma, diffuse large B cell lymphoma, activated B cell, Burkitt lymphoma, Hodgkin lymphoma, chronic lymphocytic leukemia, multiple myeloma, transfer RNA, piwi-interacting RNA, small nuclear RNA, ribosomal RNA, and small nucleolar RNA

Running Title: Small RNAs in normal and malignant B cells
Abstract

A role for microRNAs has been recognized in nearly every biological system examined thus far. A complete delineation of their role must be preceded by the identification of all microRNAs present in any system. We elucidated the complete small RNA transcriptome of normal and malignant B cells through deep sequencing of 31 normal and malignant human B cell samples that comprise the spectrum of B cell differentiation and common malignant phenotypes. We identified the expression of 333 known microRNAs, which is over twice the number previously recognized in any tissue type. We further identified the expression of 286 candidate novel microRNAs in normal and malignant B cells. These microRNAs were validated at a high rate (92%) using quantitative PCR and we demonstrated their application in the distinction of clinically relevant subgroups of lymphoma. We further demonstrated that a novel microRNA cluster, previously annotated as a hypothetical gene LOC100130622, contains 6 novel microRNAs that regulate the TGF-beta pathway. Thus, our work suggests that over a third of the microRNAs present in most cellular types are currently unknown and that these microRNAs may regulate important cellular functions.
Introduction

The role of microRNAs is being increasingly recognized in a number of diverse cellular processes, including oncogenesis\textsuperscript{1,2} and immune cell function\textsuperscript{3,4}. Efforts towards a comprehensive characterization of microRNA-expression in cellular systems are limited by the continuing discovery of new microRNAs. Over a hundred new microRNAs have been identified in humans in the past two years\textsuperscript{5}. Thus, microRNA platforms used to measure microRNA expression in any biological system are frequently obsolete by the time the studies are published.

Bioinformatic analyses of the genome suggest that there might be thousands of microRNAs encoded in the genome\textsuperscript{6}. However, thus far only about 700 unique microRNAs have been identified in humans\textsuperscript{7}. Traditional methods for discovering microRNAs have relied on cloning and Sanger sequencing\textsuperscript{8,9}. The vast dynamic range of microRNAs (5 log or higher) can result in the domination of the data by a few highly expressed microRNAs, limiting the identification of less abundant microRNAs. Massively parallel, high-throughput sequencing (deep sequencing) provides the means to identify the expression of millions of small RNA transcripts in a single tissue type, thereby providing a sensitivity of detection that is several orders of magnitude higher than conventional methods.

Mature B cell differentiation is of interest, not only for its role in adaptive immunity, but also because malignancies derived from B cells are common. B cell malignancies reflect defined stages of normal B cell differentiation and constitute the majority of leukemias and lymphomas. Therefore, we carefully chose a total of 31 samples from normal human B cell subsets and human B cell malignancies to represent the spectrum of normal and malignant B cells (Table 1). We reasoned that deep sequencing of a large number of biologically related cases would provide the opportunity to discover new microRNAs, and that shared expression of these microRNAs in separate cases would minimize the rate of false discovery.

In this study, we addressed whether deep sequencing could be used to elucidate the small RNA transcriptome of normal human B cells and human B cell tumors. We analyzed over 328,934,149 separate sequences (6 billion bases) to identify the expression of small non-coding RNAs including microRNA, tRNA, piRNA, snRNA, rRNA and snoRNA. In all, we identified 333
mature microRNAs, a number that is over twice as high as previously reported in a single tissue type\textsuperscript{17}. We also identified the expression of 286 candidate novel microRNAs, which satisfied the same structural and expression criteria\textsuperscript{19} that were used to identify known microRNAs. Through deep sequencing of biological replicates, as well as real-time PCR of known and candidate novel microRNAs, we established deep sequencing as a viable method for reproducibly measuring quantitative microRNA expression. It is anticipated that our work in delineating the expression of small RNAs, including the complete identification of their mature sequences, precursors, genome locations and conservation patterns of microRNAs, will enable a complete exploration of their functional role in normal and malignant B cells. We have thus developed a comprehensive framework that spans the spectrum of identifying new microRNAs using high throughput sequencing and validating them, and then applying real-time PCR in order to provide a clear path to clinical utility.

**Materials and Methods**

**Patient sample processing**

Patient samples were obtained using a protocol approved by the Institutional Review Board at Duke University Medical Center and the other collaborating institutions. Total RNA was extracted using the phenol-chloroform method to preserve microRNAs, using Ambion reagents.

**Selection of B cell subsets**

Tonsils from young patients undergoing routine tonsillectomy were disaggregated and separated by Ficoll. The mononuclear cell layer was harvested, washed in PBS, and resuspended in ACK lysing buffer to remove small numbers of red blood cells. After a wash and resuspension with 10 ml of PBS with 10% Bovine Serum Albumin, cells were counted and 200 million were stained with fluorochrome-tagged monoclonal antibodies to CD19, IgD, CD38 and CD27. The specific monoclonal antibodies employed were anti-CD19-PE-Cy5.5, anti-IgD-FITC, anti-CD27-PE, and anti-CD38-APC, all from BD Biosciences and BD Pharmingen (San Jose California). Cells were sorted using the MoFlo Cell sorter (Dako Cytomation, Colorado Springs, CO) into naïve B cells(CD19\textsuperscript{+}IgD\textsuperscript{+}CD27\textsuperscript{CD38\textsuperscript{-}}), germinal center B cells(CD19\textsuperscript{+}IgD\textsuperscript{-}CD27\textsuperscript{+}CD38\textsuperscript{+}), memory B
cells (CD19\textsuperscript{+}IgD\textsuperscript{-} CD27\textsuperscript{+}CD38\textsuperscript{dim}) and plasma cells (CD19\textsuperscript{dim}IgD\textsuperscript{-}CD27\textsuperscript{++}CD38\textsuperscript{+++}). Two replicates of each B cell subset were obtained from separate patients. The sample purity was verified by FACS and found to be over 90% in all cases.

**MiRNA profiling using real-time PCR**

MiRNA expression profiling was conducted using the Applied Biosystems 384-well multiplexed real-time PCR assay using 400ng of total RNA. Eight reactions, each containing 50ng of RNA and a multiplex looped primer pool with endogenous small nucleolar (sno)-RNA controls, were used to reverse-transcribe the miRNAs in parallel fashion. The completed reactions were loaded onto the 384-well plate per manufacturer’s instructions, and real-time PCR was run on the ABI 7900HT Prism instrument. For each 384-well plate, we used the automatically determined cycle-threshold (C\textsubscript{T}) using the SDS 2.2.1 software (Applied Biosystems). Consistent with manufacturer recommendations, a C\textsubscript{T} greater than 35 was treated as undetected. The probes deemed to be present were normalized to the average expression of a sno-RNA control. The expression values were calculated as $2^{-\Delta C_{T}}$, then median centered to 500 and log2-transformed.

**MicroRNA Library Preparation and Sequencing**

Total RNA (typically 5\mu g) from each sample was run on denaturing polyacrylamide-urea gels. The \textasciitilde 17-25 nucleotide RNAs were excised from the gel, ligated to sequencing adaptors on both ends and reverse transcribed. The resulting cDNA library was PCR-amplified for 15 cycles and gel purified on 6\% acrylamide gel. The gel-purified amplicon quality and quantity were analyzed on a 6\% acrylamide gel relative to oligonucleotides of known concentration and size. 120 \mu l of 1-4 pM library were loaded on to the Illumina cluster station, where DNA molecules were attached to high-density universal adaptors in the flow cells and amplified. The DNA clusters generated via this process were sequenced with sequencing-by-synthesis technology, where successive high-resolution images of the four-color fluorescence excitation dependent on the base incorporated during each cycle were captured. Sequencing reads were generated for each of the 31 samples and base calls were rendered using Illumina software. All the primary sequencing data and gene expression data is publicly available through the GEO archive through accession GSE22898
Activation of B cells

B cells were selected from peripheral blood of normal individuals using an antibody for CD19 and magnetic beads. These peripheral blood B cells were incubated with CpG and polyclonal antibodies to IgM and CD40 for 24 hours and then harvested.

Transformation of B cells with Epstein-Barr virus (EBV)

Peripheral blood B cells from normal individuals were incubated with high titers of the EBV virus and incubated for 10 days. Those cells that were replicating autonomously were cultured separately for 7 days and RNA was extracted.

Gene Expression Profiling

Gene expression profiling was performed on three replicates of each B cell subpopulation using standard Affymetrix protocols, as described previously\(^10\). Briefly, 1 μg of total RNA was reversed transcribed using an oligo(dT) primer and cDNA was synthesized. In vitro transcription using a T7 primer was used to generate labeled cRNA, which was fragmented and hybridized to Affymetrix whole-genome U133 plus 2.0 microarrays. The arrays were scanned and data normalized as described previously. The data have been deposited in the publicly available Gene Expression Omnibus database (GSE12366).

Tumor samples from 101 patients with diffuse large B cell lymphoma were obtained at the time of diagnosis and freshly frozen. These cases were profiled using Affymetrix Gene 1.0 ST arrays. The molecular subgroups were distinguished using a Bayesian approach described previously\(^10\).

GO Term Enrichment analysis

TargetScan\(^11\) (release 5.0) was used to predict conserved target genes of known microRNAs that were differentially expressed during B cell differentiation. For each pair wise comparison (naive
vs. germinal center, germinal center vs. plasma cell and germinal center vs. memory), significantly over-represented microRNA seed sequences were identified (chi-square test, p<0.001). Genes that were differentially expressed in these binary distinctions for which ontology data for molecular function existed were analyzed using GOEAST program\textsuperscript{12}. Gene ontology terms were considered significantly enriched at p <0.001.

**Comparison of RT-PCR and Sequencing Data**

Sequencing frequency for each microRNA was scaled to the total number of known microRNA sequences identified in the sample. These scaled frequencies were multiplied by 10\textsuperscript{6} and log2-transformed for the purpose of comparison between samples to the normalized RT-PCR data.

**Chromatin Immunoprecipitation**

Primary germinal center and memory B cells were crosslinked with 1% formaldehyde for 10 minutes at 37 °C and terminated with 0.125 M glycine for 5 min at room temperature. Cells were lysed in hypertonic lysis buffer (5mM Tris-HCl [pH 7.5] (Teknova), 85mM KCl, and 0.5% NP-40) on ice for 10 min. After centrifugation at 6,000 rpm for 5 min at 4 °C, the pellets of nuclei were lysed in ChIP lysis buffer (0.5% SDS, 1% NP-40, 0.5% sodium deoxycholate in 1XPBS) and sonicated (Ultrasonic CV-18) to shear chromatin DNA to about 500 bp fragments. 1/20 of each nuclear lysate was kept for input and the rest was incubated for 4 hours with Dynabeads (Invitrogen) that were coupled overnight with either the anti-H3K4me3 antibody (Millipore, 07-473) or the anti-CD20 antibody (Santa Cruz Biotechnology, sc-15361). Precipitated chromatin DNA was eluted with elution buffer (1% SDS, 1% NaHCO\textsubscript{3}) and purified using the Qiagen MinElute PCR purification kit. The library was subsequently sequenced, and semi-quantitative PCR was carried out using primers to amplify the putative promoter region of the microRNA cluster (-226 to +49 relative to the predicted transcription start site, which is 4439 bp upstream of the first microRNA gene in the novel cluster).

**Results**

**Deep Sequencing Identifies the Small RNA Transcriptome of Normal and Malignant B cells.**
Small RNA libraries from these 31 cases were subjected to massively parallel, high-throughput sequencing using the Illumina platform to generate a total of 328 million separate reads (summarized in Supplement Table 1). Our approach to analyzing the sequences and discovering microRNAs broadly follows a previously described method\textsuperscript{13} and is summarized in Figure 1A. All bioinformatics analyses were performed using a cluster of 1024 Linux computer nodes. Preprocessing was carried out using locally written Shell and Perl scripts.

From the raw sequences generated by high-throughput sequencing, the 3’- and 5’- adaptor sequences were trimmed. Low quality sequences were identified as those sequencing reads that contained stretches of consecutive identical bases or uncalled nucleotides (N) in the first 12 bases and sequencing reads shorter than 17 nucleotides. To minimize redundancy, reads were initially curtailed to the first 22 nucleotides and identical sequences were represented with a single fasta entry for analysis. Each unique sequence was mapped to the reference genome (Ensembl, build 50) and reads were filtered such that only perfect alignments (full length, 100% identity) were retained. Reads that aligned to more than five positions in the genome and reads that overlapped with the UCSC RNA genes were identified and excluded from microRNA analysis. Additional details regarding our methods are included in the supplement.

Sequencing reads that mapped to known non-coding RNAs other than microRNAs\textsuperscript{14} were annotated separately. For each read, flanking sequences were retrieved from the genome and potential secondary structures were predicted using RNAfold\textsuperscript{15}. Predictions of miRNA were made using miRDeep\textsuperscript{13} which evaluated the sequences that indicated microRNA precursor processing, and the thermodynamic energy of folding these precursors. The predicted probability and the thermodynamic energy of the microRNA precursor structures were consistent with previously described criteria for identifying microRNAs\textsuperscript{16}. An example of microRNA sequences that met the criteria for identifying a known microRNA (miR-20b) is shown in Figures 1B-1D. Sequences that occurred 20 or more times in at least one sample were consolidated and annotated for the 31 samples. Genomic loci that overlapped with microRNAs described in miRBase (version 13) were identified as known microRNAs. The remaining genomic loci were identified as encoding candidate novel microRNAs.
In order to carefully calibrate the performance of our sequencing and bioinformatic approach, we performed multiplexed real-time PCR for 360 known microRNAs on biologic replicates of the same normal B cell types. We found a high degree of overlap between the detection of microRNAs using both methods. For those 360 microRNAs that we measured using both multiplex real-time PCR and deep sequencing, we found that 64 of the 88 miRNAs were measured using both methods corresponding to a sensitivity of 72%, whereas 230/272 miRNAs were not detected by either method, corresponding to a specificity of 84%. Additional details regarding the calculation of sensitivity and specificity, as well as an alternative method of determining these statistics, which produced similar results, are described in Text S1 and Figure S1.

In all, we identified 619 microRNA precursor loci from normal and malignant B cells. The 333 microRNA precursor loci annotated in miRBase were identified as known microRNAs (Supplement Table 2), whereas the remaining 286 precursor loci were identified as candidate novel microRNAs (Supplement Table 3). Among these known microRNAs, there were 22 mature sequences (Supplement Table 2) that mapped to the ostensible microRNA* sequences and were expressed at levels that were comparable or, in some cases, higher than their complementary microRNAs. These previously unknown sequences likely reflect the influence of recently described pathways that allow both the microRNA and the microRNA* strands to independently exert their regulatory functions.

In addition to microRNAs, we identified the expression of other small RNAs, including transfer RNA (tRNA), and ribosomal RNA (rRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), piwi-interacting RNA (piRNA) and a recently described class of small RNAs, tiny RNA (tiRNA) in our data. The expression of these RNAs is shown in Figure 2A (primary data in Supplement Table 4).

**Deep sequencing reproducibly measures microRNA-expression in biologic replicates.**

We compared the expression of microRNAs identified by high-throughput sequencing and our analytic approach. In each of the biologic replicates (Figure 2B), we found that the measured microRNA expression in the biologic replicates were highly correlated (P<10^-6, correlation test
for each comparison), confirming that our approach produces robust results. We further examined the expression of 107 microRNAs that we had previously identified as being differentially expressed among B cell transitions\textsuperscript{21}. We compared the expression of these microRNAs as measured by real-time PCR and deep sequencing (Figures 2C-2E), and found highly concordant results with over 90\% of the differentially expressed microRNAs demonstrating identical patterns of expression regardless of the method of measurement (P<10\textsuperscript{-6}, chi-squared test).

Therefore, we concluded that our sequencing and analytic approach reproducibly identified microRNA expression in biologic replicates, and the differential microRNA expression measured by deep sequencing and real-time PCR produced similar results.

**Hundreds of known and novel microRNAs are expressed in normal and malignant B cells.**

Roughly two-thirds of the microRNAs measured in each cell type were known microRNAs (Figure 3A), whereas the rest were candidate novel microRNAs. In general, the expression of candidate novel microRNAs was lower than that of the known microRNAs, although there was considerable overlap between their expression levels. We also found that the correlation of microRNA expression measured by real-time polymerase chain reaction (RT-PCR) and sequencing was similar for both known and candidate novel microRNAs (Figure S2). The number of known and novel microRNAs detected in each sample type did not vary significantly between individual sample types or between normal and malignant cases. We examined the known and novel microRNAs present in the different normal B cell subsets and found that the majority of both known (Figure 3B) and novel microRNAs (Figure 3C) were shared across the B cell lineage.

The vast majority (96\%) of the candidate novel microRNAs were found in more than one sample, with only a small minority of microRNAs were expressed exclusively in a specific B cell subset or malignancy (Figure 3D). We further examined the expression of these microRNAs in all publicly available small RNA datasets from mammals. The combined number of publicly available samples (N=20) from 4 separate studies\textsuperscript{13,22,23} (GEO datasets GSE16579, GSE15190, GSE19473, GPL6583, GSE10829) is considerably smaller than that in our study. Nevertheless,
we confirmed the expression of 197 (71%) of the candidate novel microRNAs in those data (Figure 3E, Supplement Table 6). Many of the microRNAs that we identified in normal and malignant B cells were expressed at ten-fold or higher levels in these non B cell cases. These findings suggest that the microRNAs we have identified are broadly expressed and may have roles in a number of diverse tissue types.

We also examined the conservation of the novel microRNAs across four mammals (chimpanzee, rhesus macaque, mouse and dog). We found that over 80% of the identified microRNAs were conserved with two or more species (Figure 3F). Our data suggest that there may be species-specific differences in microRNA expression, as has been noted previously. In our data, 75% of novel and 80% of known microRNA had more than 100 predicted target genes with a perfectly conserved match in the 3’UTR of the gene in the following species: human, mouse, rat and dog. We found that 11% of novel and 9% of known miRNA had more than 50 predicted target genes. The remaining microRNA had fewer than 50 predicted target genes (Supplement figure 4). Analysis of ontology of predicted target genes of both known and candidate novel microRNAs expressed during B cell differentiation strongly implicated transcription factor activity as a regulated function, suggesting a potential role of microRNAs in the regulation of these processes (Supplement Table 5).

**Real-time PCR Independently Confirms the Expression of 92% of the Novel MicroRNAs**

For further validation, we selected 86 candidate novel microRNAs that were detectably measured in at the sequencing data from least one of four diffuse large B cell lymphoma (DLBCL) cases. Using stem-loop reverse transcription for quantitative PCR, we tested the expression of the 86 microRNAs in 101 primary tumors from patients with DLBCL and found that 79 (92%) were detectably measured by real-time PCR (Figure 3G) in at least 10% of these cases, suggesting that real-time PCR reproducibly identifies microRNAs that are expressed in lymphomas. We also used real-time PCR to measure the expression of 90 known microRNAs in the same 101 samples and found that over 90% of these were also detected in at least 10% of the cases using real-time PCR. We found that six of the seven RT-PCR constructs that targeted RNA hairpins that had low probability of being a microRNA resulted in no detectable signal. These
results suggest that our assays have high specificity for microRNAs and that the computational predictions based on our sequencing data correctly identified microRNAs.

MicroRNAs are Efficacious in the Distinction of Clinically Relevant Groups of Lymphoma

Gene expression profiling of patients with DLBCL has demonstrated that the tumors comprise at least two distinct diseases with different response rates to standard chemotherapy regimens\textsuperscript{27}. MicroRNAs have been shown to be robust biomarkers in malignancies\textsuperscript{28}. We hypothesized that microRNAs might be used to make this clinically important distinction for which gene expression profiling remains the gold standard. We performed gene expression profiling on 101 DLBCL cases and further subdivided these cases into the molecular subgroups. We found that 25 microRNAs with the highest t-statistic were equally efficacious as the gene expression profiling in differentiating the two groups of DLBCL with over 95% overlap between the classifications rendered by the two methods, using leave out one cross-validation (Figure 3H). Interestingly, 6 of these 25 predictor microRNAs (listed in the Supplement) were candidate novel microRNAs, suggesting a biological and clinical relevance for these candidate novel microRNAs in DLBCL tumors. The complete microRNA expression data are included in Supplement Table 8.

Deep Sequencing Reveals a New Cluster of 6 Novel MicroRNAs

Although microRNAs appear to be distributed throughout the genome, a number of microRNAs have been found in clusters such as miR-17-92 that are transcribed from a single primary transcript and cleaved into the individual microRNAs by the enzyme DROSHA. The discovery of the miR-17-92 cluster greatly accelerated the identification of the function of those microRNAs in a number of biological systems\textsuperscript{3,4}. We found 2 separate clusters of candidate novel microRNAs on chromosome 9 and chromosome 14 (within the IgH locus) respectively. The first cluster was previously annotated as a hypothetical gene \textit{LOC100130622}, and subsequently discarded from Refseq when no associated protein was identified. Our data demonstrate that this cluster, conserved only in primates, encodes 6 separate microRNAs (Figure 4A). In order to evaluate whether the microRNAs encoded in these clusters originate from the same primary transcript, we took KMS12 multiple myeloma cells which express these
microRNAs and used siRNA to knock-down the expression of the microRNA processing enzyme Drosha. This enzyme acts at the first step of microRNA processing by cleaving microRNA precursors from the primary transcript. We found that decreased Drosha expression was associated with increased accumulation of primary transcripts of both the miR-17-92 cluster as well as the novel miR-2355 cluster. (Figure 4B). MicroRNAs from this cluster were found to be expressed more highly in normal germinal center (GC) B cells compared to memory cells (Figure 4C). We computationally predicted the core promoter region at 4 kb upstream from the primary transcript. We performed chromatin immunoprecipitation for H3K4Me3, the histone marker associated with gene expression. DNA bound to this histone marker was found to be selectively enriched in primary GC cells which express this microRNA cluster highly, but not in primary memory B cells (Figure 4D). The microRNAs of this cluster all share the same seed sequence, suggesting that they target the same genes. Among the computationally predicted targets of this microRNA cluster, we identified SMAD2 and SMAD3 which are well known mediators of the TGF-beta signaling pathway. We noted that gene expression both SMAD2 and SMAD3 in our set of 101 DLBCLs were inversely correlated with this cluster (P<0.01, correlation test). Gene set enrichment analysis revealed that expression of the TGF-beta pathway in DLBCL samples varied inversely with the expression of the microRNA cluster, with a higher expression of the microRNA association with a lower expression of the pathway (P<10^-6, Figure 4E and Supplement), which has previously been noted to be important in the biology of these tumors.

Discussion
The advent of widely accessible whole genome sequencing has created a new urgency to annotate the human genome to provide the context for understanding the genetic variation underlying health and disease. Non coding RNAs have emerged as the “dark matter” of the genome that may play an important role in regulating the 1% of the total DNA that is expressed as protein. The ongoing identification of new microRNA genes in different tissue types is analogous to the discovery of protein-coding genes. Our work provides an exhaustive identification of the microRNAs in normal and malignant B cells, that is a prerequisite to the delineation of their role. Further, we have developed a comprehensive framework that spans the
identification of microRNAs from deep sequencing data to measuring their expression using real-time PCR and validating their expression in primary human tumors.

A number of microRNAs have previously been suggested as being tissue-specific. For instance, miR-223 has been suggested as a myeloid-specific microRNA33, whereas miR-9, miR-124, miR-128 have been suggested as brain-specific microRNAs8. However, our study found expression of all of these microRNAs in the B cell lineage. Our data indicate that the highly increased sensitivity of deep sequencing will challenge currently held notions regarding tissue-specificity of microRNAs, and suggest that the same microRNAs could play different context and lineage-specific roles in different cell types.

MicroRNA expression has a wide dynamic range from a few transcripts to hundreds of thousands of transcripts per cell. Deep sequencing is especially efficacious for the identification of microRNAs expressed at low abundance in a cell. The number of microRNA transcripts needed to exert their effects in a cell is not known, and it is likely to be context-dependent. Further, it has been demonstrated that some microRNAs containing the hexanucleotide motif AGUGUU can relocalize to the nucleus to potentially regulate the transcription of target genes37. These findings suggest that a few copies of a microRNA in a cell may be adequate to exert profound down-stream effects. We found the same hexanucleotide motif in 5 microRNAs expressed in B cells. It is also conceivable that some of the low-abundance microRNAs that we have identified in our study may be expressed at higher levels in other development stages or in other cell types. This notion is confirmed by our examination of the novel microRNAs in non B cell data. For instance, a number of the microRNAs that we discovered were also present at 10-fold or higher levels in cell lines derived from breast cancer and cervical cancer, suggesting that the microRNAs that we have discovered in B cells have broad biological significance.

MicroRNAs have shown promise as biomarkers in a number of malignancies20, 33, 34, 39. Diffuse large B cell lymphoma is the most common form of lymphoma and is known to comprise at least 2 molecularly distinct subgroups with different responses to standard therapy. There are 2 important clinical applications for the molecular sub-grouping of DLBCL patients. First, the prognostic information can inform the choices and expectations of patients and their physicians.
Second, the important molecular differences in these subgroups might dictate the use of different therapies in these patients. For instance, the benefit of receiving a proteosome inhibitor, bortezomib, appears to be predominantly limited to those patients who have ABC DLBCL\textsuperscript{40}. However, the current methods can only distinguish GCB from non GCB DLBCL in a limited fashion\textsuperscript{41}, and have yielded sometimes conflicting results\textsuperscript{42, 43}. Gene expression profiling remains the gold standard for the distinguishing the 2 molecular subgroups; however, this is not routinely performed in clinical laboratories. Our data suggest that microRNAs can render this distinction with equal efficacy. Since microRNAs are stable in tissue processed for routine paraffin sections using standard methods\textsuperscript{44}, microRNA-based assays would be much easier to implement in the clinical laboratory.

Emerging data have demonstrated a regulatory role for small RNAs in nearly every cellular system examined thus far. A comprehensive delineation of their role in any cellular system must be preceded by a complete identification of the small RNA transcriptome in that system. The microRNAs that we have identified will likely be included in future commercial microarray platforms that would be widely available. The expression and role of individual microRNAs can also be queried using existing methods as we have demonstrated. Thus, we anticipate that our data, which represents the most in-depth examination of small RNA sequences using second-generation methods, will serve as a foundation for the exploration of the role of small RNAs in a number of cellular systems.
Acknowledgements
We thank Drs. Joseph Nevins and Jen-Tsan Chi for a critical review of the manuscript. The authors thank Susan Sunay for assistance with and the Georgia Cancer Coalition for support for sample collection. S.S.D. was supported by the Doris Duke Charitable Foundation and the National Institutes of Health.

Authorship Contributions

Conducted Data Analysis: Dereje Jima, Jenny Zhang, Cassandra Jacobs, Sandeep Dave

Wrote paper: Sandeep Dave, Jenny Zhang, Dereje Jima

Disclosure of Conflict of Interest
The authors have no conflicts to disclose.
References


Table 1: Normal and Malignant B cells used to identify small RNA transcriptome

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample</th>
<th>B cell stage/ Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Naive B cell (FACS-sorted primary human)</td>
<td>Naive</td>
</tr>
<tr>
<td>S2</td>
<td>Naive B cell (FACS-sorted primary human)</td>
<td>Naive</td>
</tr>
<tr>
<td>S3</td>
<td>Germinal Center B cell (FACS-sorted primary human)</td>
<td>Germinal center</td>
</tr>
<tr>
<td>S4</td>
<td>Germinal Center B cell (FACS-sorted primary human)</td>
<td>Germinal center</td>
</tr>
<tr>
<td>S5</td>
<td>Memory B cell (FACS-sorted primary human)</td>
<td>Memory</td>
</tr>
<tr>
<td>S6</td>
<td>Memory B cell (FACS-sorted primary human)</td>
<td>Memory</td>
</tr>
<tr>
<td>S7</td>
<td>Plasma cell (FACS-sorted primary human)</td>
<td>Plasma cell</td>
</tr>
<tr>
<td>S8</td>
<td>Plasma cell (FACS-sorted primary human)</td>
<td>Plasma cell</td>
</tr>
<tr>
<td>S9</td>
<td>Activated B cell (primary human)</td>
<td>Naive</td>
</tr>
<tr>
<td>S10</td>
<td>EBV-transformed B cell (primary human)</td>
<td>Naive</td>
</tr>
<tr>
<td>S11</td>
<td>Lymphoblastic lymphoma (primary human tumor)</td>
<td>Precursor B cell</td>
</tr>
<tr>
<td>S12</td>
<td>Transformed follicular lymphoma (primary human tumor)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td></td>
<td>Diffuse Large B cell lymphoma</td>
<td></td>
</tr>
<tr>
<td>S13</td>
<td>Diffuse Large B cell lymphoma (Farage human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S14</td>
<td>Diffuse Large B cell lymphoma (BJAB human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S15</td>
<td>Diffuse Large B cell lymphoma (LY3 human cell line)</td>
<td>Activated phenotype</td>
</tr>
<tr>
<td>S16</td>
<td>Burkitt lymphoma (primary human tumor)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S17</td>
<td>Burkitt lymphoma (Ramos human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S18</td>
<td>Burkitt lymphoma (Daudi human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S19</td>
<td>HIV lymphoma (primary human tumor)</td>
<td>Unknown</td>
</tr>
<tr>
<td>S20</td>
<td>Mantle cell lymphoma (JVM2 human cell line)</td>
<td>Naive</td>
</tr>
<tr>
<td>S21</td>
<td>Mantle cell lymphoma (JVM13 human cell line)</td>
<td>Naive</td>
</tr>
<tr>
<td>S22</td>
<td>Mantle cell lymphoma (Mino human cell line)</td>
<td>Naive</td>
</tr>
<tr>
<td>S23</td>
<td>Hodkgin lymphoma (L1236 human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S24</td>
<td>Hodkgin lymphoma (L428 human cell line)</td>
<td>Germinal Center</td>
</tr>
<tr>
<td>S25</td>
<td>Multiple Myeloma (U266 human cell line)</td>
<td>Plasma cell</td>
</tr>
<tr>
<td>S26</td>
<td>Multiple Myeloma (H929 human cell line)</td>
<td>Plasma cell</td>
</tr>
<tr>
<td>S27</td>
<td>Multiple Myeloma (KMS12 human cell line)</td>
<td>Plasma cell</td>
</tr>
<tr>
<td>S28</td>
<td>Marginal Zone lymphoma (primary human tumor)</td>
<td>MALT type</td>
</tr>
<tr>
<td>S29</td>
<td>Marginal Zone lymphoma (primary human tumor)</td>
<td>Splenic type</td>
</tr>
<tr>
<td>S30</td>
<td>Chronic lymphocytic leukemia (primary human tumor)</td>
<td>IgVH Mutated phenotype</td>
</tr>
<tr>
<td>S31</td>
<td>Chronic lymphocytic leukemia (primary human tumor)</td>
<td>IgVH Unmutated phenotype</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1: Identifying MicroRNAs in Normal and Malignant B Cells

A Analytic pipeline for the discovery and validation of known and novel microRNAs in B Cells. Small RNAs from B cell subsets were subjected to massively parallel sequencing. 328,934,149 raw sequence reads filtered, mapped to the genome, analyzed by miRDeep, and annotated. Identified microRNA loci were cross-referenced with miRBase13 to distinguish known from candidate novel microRNA.

B Frequency of reads for miR-20b and their alignment to the genome. The sequence surrounded by a rectangle corresponds to the most abundant and longest mature microRNA sequence that perfectly matched the genome.

C Outputs of the RNAshapes program that computed the abstract structure of the microRNA precursor, the probability of the predicted structure, and the minimum free energy of folding for the precursor structure for miR-20b.

D Energetically favorable folding of the miR-20b precursor, with the microRNA mature sequence and the microRNA* sequence highlighted in yellow.

Figure 2: Distribution of small RNAs in normal and malignant B cells

A Expression of small RNAs, including microRNAs (miRNA), transfer RNA (tRNA), piwi-interacting RNA (piRNA), small nuclear RNA (snRNA), ribosomal RNA (rRNA), small nucleolar RNA (snoRNA), and tiny RNA (tiRNA) in each sample are shown over a two-fold range. Samples are listed in the same order as in Table 1.

B Expression of microRNAs measured by sequencing across two biological replicates for Naïve, Germinal Center, Memory, and Plasma B Cells. The P-values were computed using a correlation-test against the null hypothesis that the data were not correlated; p<10^-6 in all 4 cases.

C Relative expression of microRNAs differentially expressed in the naive to germinal center B cell transition as measured by real-time PCR (RT-PCR) and sequencing.

D Relative expression of microRNAs differentially expressed in the germinal center B cell to plasma transition as measured by real-time PCR and sequencing.

E Relative expression of microRNAs differentially expressed in the germinal center B to memory B cell transition as measured by real-time PCR and sequencing.
Figure 3: Identifying novel microRNAs through high throughput sequencing

A  Distribution of known and candidate novel microRNAs in normal and malignant B cells. Samples are listed in the same order as Table 1.
B  Distribution of known microRNA expression among the normal B cell subsets. The majority of microRNAs are shared between the B cell types.
C  Distribution of candidate novel microRNA expression among the normal B cell subsets. The majority of microRNAs are shared between the B cell types.
D  Proportion of samples in which candidate novel microRNAs were identified. The vast majority of microRNAs are expressed in more than one sample.
E  Proportion of candidate novel microRNAs identified in twenty additional sequencing datasets.
F  Conservation of novel human microRNAs in Chimpanzee, Rhesus Monkey, Mouse, and Dog. Ninety percent of the 286 novel microRNAs were found to be conserved across humans and one or more additional species, and over 80% were found to be conserved across humans and two or more additional species.
G  Validation of 86 novel microRNAs candidates by an alternative method of measurement, real-time PCR. Maximum measured values for microRNAs expressed in at least 10% of the samples are shown in log2-scale. Detection threshold is the expression level corresponding to CT ≤35.
H  MicroRNA profiling distinguishes the molecular subgroups of diffuse large B cell lymphoma (DLCBL) that were defined based on gene expression profiles. ABC refers to activated B cell like DLBCL. GCB refers to germinal center B cell like DLCBL. Unclassified cases are those that did not meet criteria for either group. Black bars indicate candidate novel microRNAs that were differentially expressed between the two groups.

Figure 4: Discovery and Functional Validation of a Novel microRNA Cluster

A  Genomic locus of the mir-2355 cluster (chromosome 9q34.3), conservation across three primates and one additional mammal (mouse), and predicted secondary structure of each microRNA hairpin, with the mature sequences highlighted in yellow. The sequences are conserved only in primates.
B Knock-out of the primary microRNA transcript processing enzyme, Drosha, by RNA interference results in accumulation of a known microRNA cluster primary transcript (pri-hsa-mir-17) and the novel microRNA cluster primary transcript (pri-hsa-mir-2355).

C Expression of hsa-miR-2355 is higher in germinal center B cells compared to memory B cells.

D Chromatin immunoprecipitation with an antibody to H3K4me3 of germinal center B cells and memory B cells shows enrichment of DNA from the predicted core promoter region of the miR-2355 cluster in germinal center B cells compared to memory B Cells. Three sets are shown: input (positive control for PCR using chromatin prior to immunoprecipitation), immunoprecipitation with an antibody for H3K4me3, and immunoprecipitation with an antibody for CD20 (negative control).

E GSEA enrichment score and distribution of TGF-beta pathway genes along the rank of transcripts differentially expressed in high versus low expressors of the miR-2355 cluster. Expression of the TGF-beta pathway genes were inversely correlated with expression of the microRNA.
FIGURE 1

A

Normal B cell subsets and B cell malignancies (N=31)

Size Fractionate microRNA and perform High Throughput Sequencing

Total raw sequences: 328,934,149

Map sequences to human genome

Total genomic loci: 12,044,212

Identify sequences corresponding to repeats, and other non-coding RNAs

Identify tRNA, piRNA, tRNA, snRNA, rRNA, snoRNA

Total non-redundant sequences: 3,651,054

Extract potential microRNA precursors from the genome and predict microRNA structures

Total predicted structures: 5,071,057

Align reads to predicted microRNA Precursor Structures

Total number of reads aligned to predicted structure: 6,437,87

Candidate microRNA prediction

Total predicted loci = 619

Known MicroRNAs Total=333

Candidate Novel MicroRNAs Total = 286

B

Reads

Mature MicroRNA Sequence

3
CAAGTGCTCATAGTGCAGGTA

4
CAAGTGCTCATAGTGCAGGTA

5
CAAGTGCTCATAGTGCAGGTA

7
CAAGTGCTCATAGTGCAGGTA

15
CAAGTGCTCATAGTGCAGGTA

99
CAAGTGCTCATAGTGCAGGTA

Star Sequence

A CTGATGATAGGGCAGTCCAGA

A CTGATGATAGGGCAGTCCAGA

A CTGATGATAGGGCAGTCCAGA

Reference Sequence: Chromosome X, 133131547 - 133131568 (miR-20b) -1 strand

Structure: ((((((((((((((((((((((((((((((((((((((( ((((((((((((((((

Probability of Predicted Structure: 100%

Minimum Free Energy of Folding: -30.00 Kcal

C

Deep Sequencing Output

Reads

4
17
14

RNAshapes Output

agtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagtgaagttgcatgactcctact
Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs