Computational analysis of sequencing data

Each unique sequence was mapped to the human genome (Ensembl, build 50, downloaded from ftp://ftp.ensembl.org). Reads were mapped to the human genome using NCBI Megablast (ftp://ftp.ncbi.nlm.nih.gov/blast/) with the following setting: -W 12 –D 2 –p 100. Only sequences that aligned to the genome exactly starting from the first nucleotide were retained. In order to exclude sequences originating from repetitive elements, reads that aligned to the genome at more than five loci were discarded. In addition, reads that aligned to loci annotated as RNA genes (rRNA, scRNA, snRNA, snoRNA, tRNA) by UCSC were identified using a script included within the miRDeep software package, and excluded from microRNA analysis. In a similar way, all reads that aligned to Rfam RNA genes (ftp://ftp.sanger.ac.uk/pub/databases/Rfam/, version 9.1) were identified using megablast and discarded.

Based on the megablast output, potential miRNA precursors were excised from the genome using scripts included within the miRDeep software. The precursor secondary structures were predicted using RNAfold (I.L. Hofacker 1993). The miRDeep signatures were generated by aligning the filtered reads to the potential precursors per miRDeep instructions. Default miRDeep options were used to analyze samples independently. Consistent with the original description of the algorithm, we used a minimum log-odds score of 1 for identifying known and candidate novel microRNA. The numbers of unique reads at each stage of the microRNA discovery pipeline are summarized in Supplement Table 1.

For all reads predicted to originate from miRNA precursors, the loci (chromosome name, strand, mature sequence start, and mature sequence end) were retrieved from the megablast output. Because the megablast output only references the positive strand, reads that mapped to the minus strand were assigned corresponding new mature sequence start and end positions. The loci information were matched to the miRDeep output and stored in a local database. At this stage, all predicted miRNA reads from all 31 samples were consolidated for annotation purposes, and duplicate loci were removed. For every genomic locus, coding gene and non-coding RNA gene annotations were retrieved from the Ensembl database. All loci overlapping with Ensembl known miRNA annotations were categorized as known miRNA. The remaining reads were categorized as potential candidate novel miRNA. Each candidate novel miRNA locus was further examined in the Ensembl genome browser and discarded if it fell within coding regions.

Finally, for all loci predicted by miRDeep in at least one sample, we counted the mature sequence in all other samples. Only reads found 20 times in at least one sample were considered as known and candidate miRNA. We identified 619 total loci. Of these, 333 corresponded to known microRNAs, and the remaining 286 loci were categorized as candidate novel microRNAs.

Sensitivity and specificity using sequencing and quantitative PCR

We determined the sensitivity and specificity of our microRNAs discovery method by comparing sequencing data to quantitative PCR data of 360 known microRNA measured by multiplexed real-time PCR for the detection of microRNAs in the normal B cell replicates. Of these 360 known microRNAs, 88 microRNA were detected by our sequencing analysis, of which 64 were deemed as true positives based on quantitative PCR, resulting in a sensitivity of 72.7%. Of the 272 microRNA that were assayed by real-time PCR, and not detected by sequencing, 230 were also found to be absent by quantitative PCR, for a specificity of 84.6%. Detection by sequencing analysis was defined as discovery by miRDeep and presence of reads in at least half the normal B Cell samples. Detection by quantitative PCR was defined as measurement of Ct less than 36 in at least half of the normal B Cell samples.

Sensitivity and Specificity using miRDeep

In addition, computationally-based sensitivity and specificity estimates as recommended by miRDeep were made for two representative samples (Fig. S1).

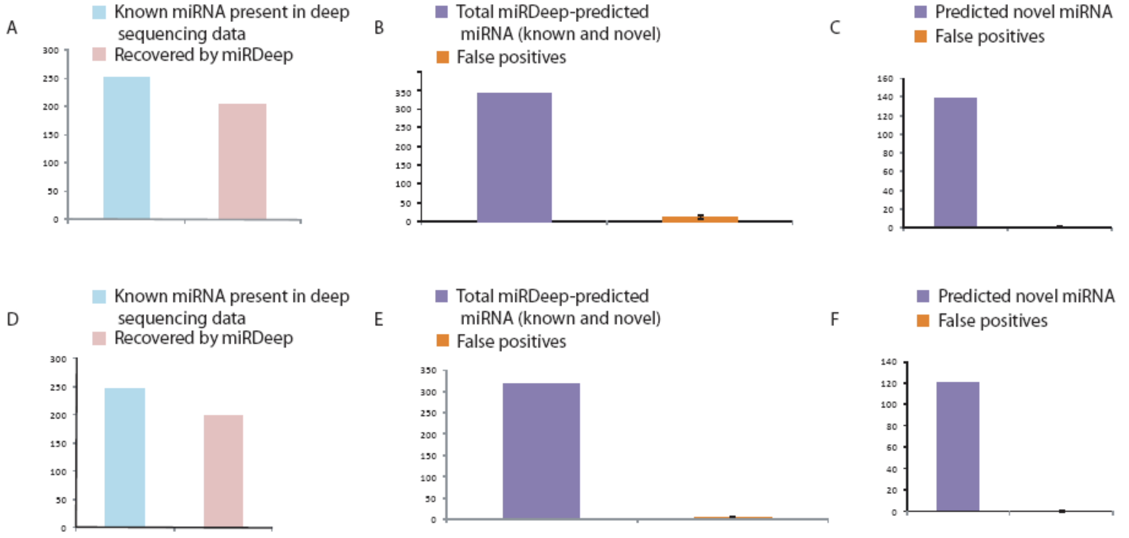

Figure S1. Sensitivity and false positive estimates for Plasma Cell samples PC137 (A-C) and PC44 (D-F).
A and D: Known microRNA mature sequences present in each deep sequencing data set (azure) compared to the final number recovered by miRDeep analysis (red). By this measure, the sensitivity is ~80%.
B and E: False positives (orange) for known and novel miRNA predicted by miRDeep (purple). Signal-to-noise ratios were 29:1 (B) and 60:1 (E).
C and F: False positives (orange) for only novel miRNA predicted by miRDeep (purple). Signal-to-noise ratios were 100:1 in both cases.

Promoter analysis of Novel miRNA cluster

A 5k bp sequence upstream of the novel mir-2355/2356 cluster was retrieved to predict a potential promoter region upstream of the cluster. An initial promoter prediction was made using the Neural Network Promoter Prediction for eukaryote algorithm (Reese, 2001) with a minimal promoter score of 0.8. Three potential promoter regions were predicted (See list below).

Predicted promoter regions: the locus of the predicted transcription start site (TSS) is shown in bold, and the predicted TATA box is underlined.

Start	End	Chr	strand	Score	Promoter Sequence
136886904	136886953	9	minus	0.99	GCTCCAGCTATGAAAAAGGGGGCTCCCTCTACCCCACTGTG**G**ACGGCCAC
136885967	136886016	9	minus	0.85	CTGTGATTGGCATAAAAGATGGTGCTTCCATCTTGCTAGC**A**GACTCTCCT
136884457	136884506	9	minus	0.93	TATGTCTGCATCTAAAGCACCCACTGTTTGTTACATGGGCAC**A**CCTCCAA

To determine which of these three candidate sequences might serve as a core promoter, we mapped the ChIP-Seq data of trimethylated lysine residue K4 of histone 3 (H3K4me3) from germinal center B cells and memory B cells (data not shown). A total of 8 reads were mapped uniquely to the 5kb region upstream of the cluster in the Germinal Center B cell ChIP-Seq data. Furthermore, all of the 8 reads mapped to the flanking region of the first predicted promoter shown in the list above. However the ChIP-Seq data from memory B cells showed only a single read that mapped to the 5kb region flanking the first predicted promoter, consistent with the cluster being expressed more highly in GC than memory B Cells. It is known that trimethylation of H3K4me3 on either side of the transcriptional start site strongly correlates with active promoters, which raises the possibility that the first predicted promoter acts on this novel cluster.

Semi-quantitative PCR was then performed to validate the relative enrichment of H3K4me3 observed in the GC B cell ChIP-Seq data. We used primers to amplify the −226 to +49 bp region relative to the TSS (FW 5’GCAGTGGGAGCTCAGGTG3’ REV 5’TCCCAGATTCCAAAAGAGGA3’) of the first predicted core promoter from the chromatin DNA immunoprecipitated with either the anti-H3K4me3 antibody or the anti-CD20 antibody as a control. Consistently, an enhanced association of this region with H3K4me3 was seen in GC B cells when compared with memory B cells (**Fig. 3D**). The relative enrichment of the active promoter indicator H3K4me3 in GC B cells when compared to memory B cells may explain the higher expression of the novel cluster in GC B Cells.

Analysis of other ncRNA

To determine the other ncRNA in the deep sequencing data, known snRNA, snoRNA, and rRNA sequences were compiled from the Ensembl database (Nov. 2009, version 50). tRNA sequences were downloaded from Genomic tRNA Database (http://lowelab.ucsc.edu/GtRNAdb/Hsapi19/, UCSC, Hg19), and piRNA sequences were downloaded from piBank (http://pirnabank.ibab.ac.in/ ,piBank Version 2.0, Institute of Bioinfomatics and Applied Biotechnology, India). This information was merged and formatted to create a single searchable database (formatdb can be found within the NCBI Blastall software package). All non-redundant reads were aligned to this database using megablast. Sequences that aligned perfectly (full length, 100% identity) were annotated accordingly. Finally, these sequences found to map to ncRNA were counted in all 31 samples, and redundancies were removed. In addition, human tiny RNA (tiRNA) associated with transcriptional start sites were downloaded from the FANTOM site (http://fantom.gsc.riken.jp/download/Supplemental_Materials/Taft_et_al_2009/) and searched within our deep sequencing data using an in-house Perl script.

Analysis and comparison of sequencing to quantitative PCR

For each sample, sequencing read counts were median-centered to 500 and log2 transformed. Quantitative PCR expression data were linearly transformed (2^{-CT}), normalized to RNU48 expression, median-centered to 500, and log2 transformed. Consistent with previous descriptions (Linsen et al, Nature Methods 2009), we found that direct correlation of log2-transformed sequencing and qPCR data was relatively low, because in many cases, the expression of a particular microRNA was better measured using one assay than another. However, the relative expression of miRNA between two or more samples was found to be quite well-preserved between the two measurement methods. This is evident from Fig. 2 of the paper.

To investigate whether we could apply the same analysis to predicted known and novel microRNA present on our custom quantitative PCR array, we examined the microRNAs from deep sequencing and generated reproducible results that could be measured using qPCR data for one sample, U266 (Fig. S2).

Following normalization and log2 transformation as described above, we compared the expression of U266 (S25) to the average of 4 other samples (S12, S16, S21, S25). We plotted the sequencing and qPCR relative values against one another and found that the correlation between the relative expressions was comparable for known and novel microRNAs.

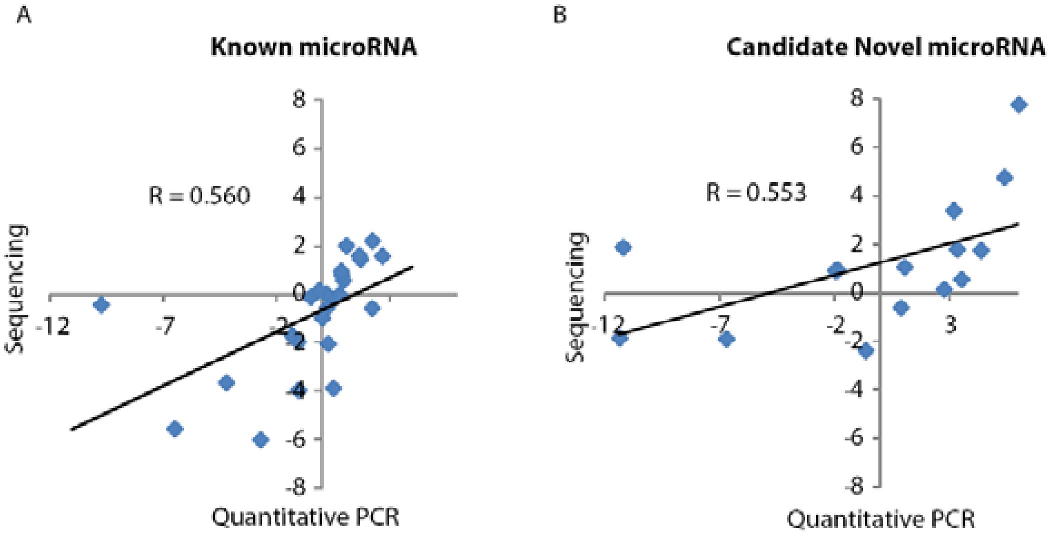

Figure S2

Fig. 2A shows the relative expression of known microRNAs measured in sample S25 by quantitative (real-time) PCR compared to deep sequencing. The measured correlation of the data is 0.56.

Fig. 2B shows the relative expression of microRNAs for candidate novel microRNAs expressed at comparable levels. We obtained similar results, with a correlation of 0.55.

Measuring the effect of knocking down Drosha on the cluster mir-2355

1 million KMS12 cells were electroporated with siRNA at 50 nM in 100μl nucleofector volume using the Lonza system (Nucleofector V, program T-001). The control siRNA was obtained from IDT. The siRNA targeting Drosha was a duplex of 5'rCrGrArGrUrArGrGrCrUrUrCrGrUrGrArCrU rUrArUrArUG A3' (top) and 5' rUrCrArUrArUrArArGrUrCrArCrGrArArGrCrCrUrArCrUrCrGrUrU 3' (bottom). Two days post-transfection, RNA was isolated using TriZol and reverse transcribed. Expression of hsa-pri-17 was measured with real-time PCR assay Hs03295901_pri (Applied Biosystems). Expression of cluster mir-2355 was measured by Sybr green real-time PCR with the annealing temperature set to 56°C. The primer sequences were 5'GGGTGGCTTGGCTTAGAAGG3' (FW) and GCCAGCGCTTAGCAGAAGTG 3' (REV).

Measuring the effect of cluster mir-2355 on the TGFβ_1 Pathway

microRNA and microRNA seed sequence mutant expression constructs were created for hsa-mir-2355a and hsa-mir-2355b precursors by annealing and ligating 100bp oligos (sequences shown below) into the Xho1 and NotI sites of the pL/CMV/eGFP vector (choo-choo cloning, mcLab). This pL/CMV/eGFP vector was generated by ligating a fragment containing the CMV promoter and the EGFP ORF into the BamHI and XhoI sites of the previously described lentiviral backbone pL[21]. The expression of the microRNA from the 3'UTR of EGFP in the resulting vector was confirmed by Taqman real time PCR in transfected 293T cells.

ID	name	Primer Sequences (5' to 3')
1	pre-miR-2355a	ctgtacaagtagctcgagggaggTGTGATATCATGGTTCCTGGGAggtatgatatcgtggttcctgggaggtgtgatcccgtgctcccgcggccgccaga
2	pre-miR-2355a_rc	tctggcggccgcgggagcacgggatcacacctcccaggaaccacgatatcataccTCCCAGGAACCATGATATCACAcctccctcgagctacttgtacag
3	pre-miR-2355a_mut	ctgtacaagtagctcgagggaggTGTCAAATCATGGTTCCTGGGAggtatgatatcgtggttcctgggaggtgtgatcccgtgctcccgcggccgccaga
4	pre-miR-2355a_mut_rc	tctggcggccgcgggagcacgggatcacacctcccaggaaccacgatatcataccTCCCAGGAACCATGATTTGACAcctccctcgagctacttgtacag
5	pre-miR-2355b	ctgtacaagtagctcgagggaggTGTGATATCGTGCTTCCTGGGAcgtgtgatgctgtgcttcctgggaggtgtgatcccacactcgcgcggccgccaga
6	pre-miR-2355b_rc	tctggcggccgcgcgagtgtgggatcacacctcccaggaagcacagcatcacacgTCCCAGGAAGCACGATATCACAcctccctcgagctacttgtacag
7	pre-miR-2355b_mut	ctgtacaagtagctcgagggaggTGTCAAATCGTGCTTCCTGGGAcgtgtgatgctgtgcttcctgggaggtgtgatcccacactcgcgcggccgccaga
8	pre-miR-2355b_mut_rc	tctggcggccgcgcgagtgtgggatcacacctcccaggaagcacagcatcacacgTCCCAGGAAGCACGATTTGACAcctccctcgagctacttgtacag

TGFβ1 pathway activity was monitored by measuring activity of the pE2.1 vector, which contains a luciferase gene under the regulation of the TGFβ-inducible promoter PAI-1, relative to control reporter (phRL/SV40). 293T cells at 50% confluency on 24-well plates were co-transfected with 1μg of the microRNA expression vector, 200ng of the TGFβ reporter vector, and 20ng of the control reporter using FUGENE (Roche). 24 hours post-transfection, transfected cells were induced by replacing normal growth media with 0.5 ml serum-free DMEM containing TGF-β at 100 pM. 24 hours post-induction, Firefly and Renilla luciferase were measured by the Dual Glo luciferase assay (Promega).

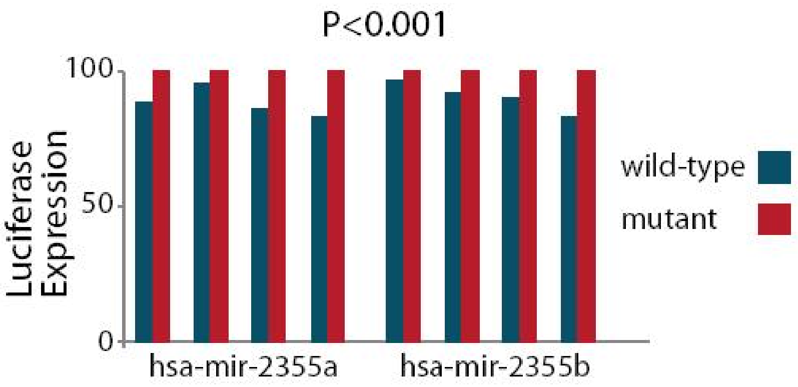

Figure S3

Dual Luciferase assay measuring TGF-β pathway activity after over-expression of hsa-mir-2355a, hsa-mir-2355b, and their respective seed sequence mutants. Introduction of mutations in the seed sequence abolishes the downregulation observed with higher expression of hsa-miR-2355a and hsa-miR-2355b (P<0.001).

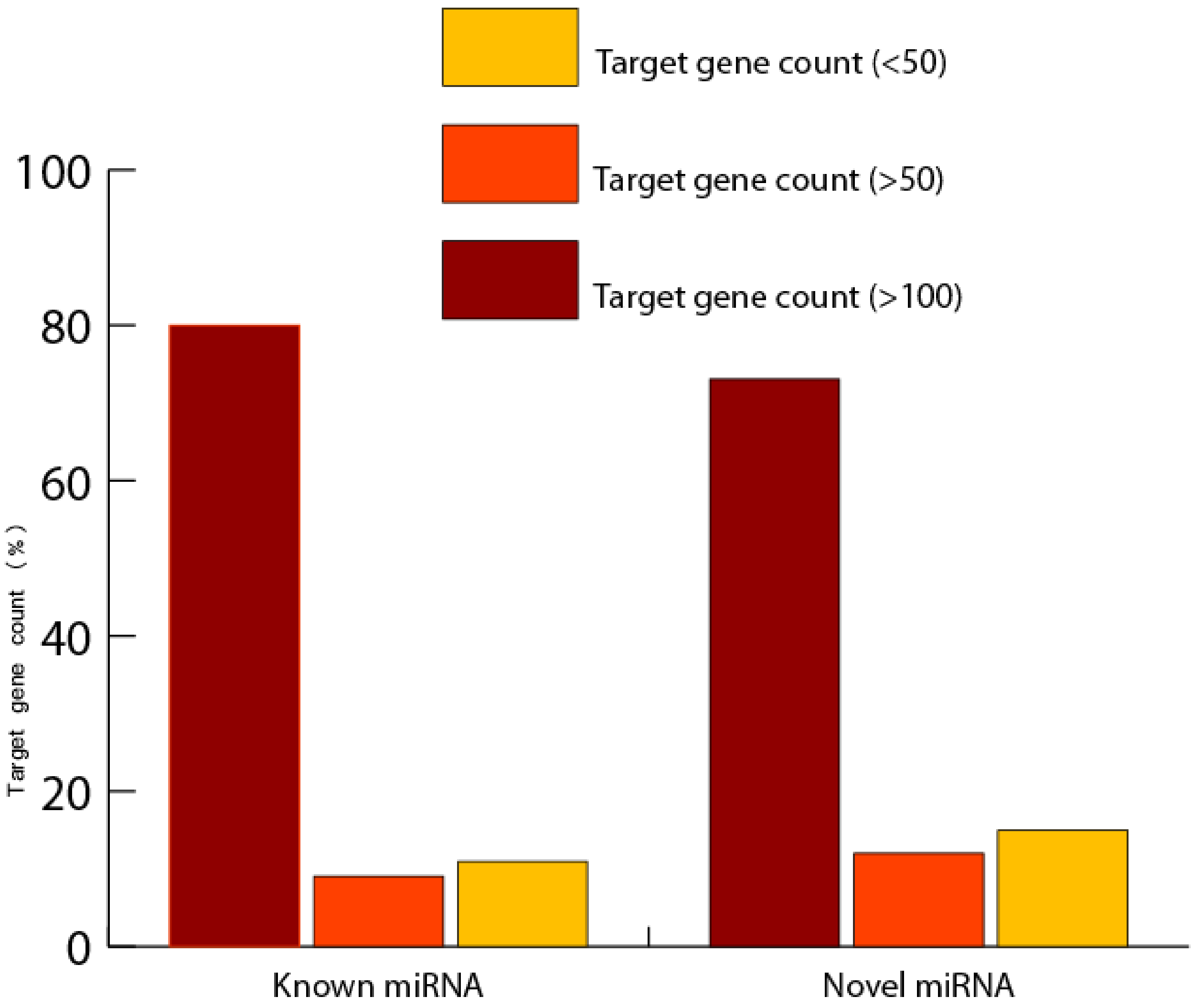

Figure S4

Distribution of the number of predicted target genes for novel and known miRNA. 75% of novel and 80 % of known miRNA had more than 100 predicted target genes, while 11 % of novel and 9 % of known miRNA had more than 50 predicted target genes.

Predictor MicroRNAs for distinguishing DLBCL subgroups.
hsa-miR-128
hsa-miR-129-3p
hsa-miR-152
hsa-miR-155
hsa-miR-185
hsa-miR-193a-5p
hsa-miR-196b
hsa-miR-199b-3p
hsa-miR-20b
hsa-miR-23a
hsa-miR-27a
hsa-miR-28-5p
hsa-miR-301a
hsa-miR-331-3p
hsa-miR-365
hsa-miR-625
hsa-miR-9
hsa-mir-2282 (Novel, AGATTGTTTCTTTTGCCGTGCA)
hsa-mir-2287 (Novel, GGCTCCTTGGTCTAGGGGTA)
hsa-mir-2290 (Novel, TGGGGATTTGGAGAAGTGGTGA)
hsa-mir-2311 (Novel, ACTGGACTTGGTGTCAGATGG)
hsa-mir-2348 (Novel, CAGGAAGGATTTAGGGACAGGC)
hsa-mir-2412 (Novel, TAGTGAGTTAGAGATGCAGAGC)
hsa-mir-2432 (Novel, TGGTCTGCAAAGAGATGACTGTG)
hsa-mir-2450 (Novel, CAAAAACTGCAGTTACTTTTGT)

Friedlander, M. R., W. Chen, et al. (2008). "Discovering microRNAs from deep sequencing data using miRDeep." *Nat Biotechnol* **26**(4): 407–415.

I.L. Hofacker, W. F., P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster (1993). "Fast Folding and Comparison of RNA Secondary Structures." *Monatshefte f. Chemie* **125**: 167–188.

Reese, M.G., (2001). ``Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome", *Comput Chem*, 26(1):51–6, 2001.