Supplemental methods - Li et al.

DLBCL samples and cell lines

Fifty nine frozen DLBCL samples were selected for further analysis from the San Antonio Cancer Institute (SACI) tumor bank. Approval was obtained from the University of Texas Health Science Center at San Antonio institutional review board for these studies. All DLBCL samples were reviewed for diagnostic accuracy by a hematopathologist (R.S.R.). An additional 26 DLBCL cell lines and one Mediastinal Large B-cell lymphoma (MLBCL) were also selected for analysis, including: OCI-Ly1, OCI-Ly3, OCI-Ly4, OCI-Ly7, OCI-Ly8, OCI-Ly18, OCI-Ly19, SU-DHL-4, SU-DHL-5, SU-DHL-6, SU-DHL-7, SU-DHL-8, SU-DHL-9, SU-DHL-10, SU-DHL-16, NU-DHL-1, NU-DUL-1, USC-DHL-1, WSU-NHL, HT, Toledo, RC-K8, Pfeiffer, FL-318, Farage, Karpas-422, Karpas-1106p (MLBCL). Cell lines were maintained in RPMI 1640 supplemented with 10% FCS as described.

Immunohistochemistry

DLBCL sample frozen sections were subjected to immunostaining with antibodies to CD3, CD10, CD20, CD68, CD138, BCL6, MUM1, Ki-67, and BCL2. Cases were considered positive if 30% or more of the tumor cells were stained by an antibody. DLBCL samples were categorized as germinal center B-cell-like (GCB) or non-germinal center B-cell-like (non-GCB) according to the algorithm described by Hans et al. (Blood 2004; 103:275-82). The extent of infiltration of the DLBCL cases by T-cells and macrophages was also determined by semi-quantitative measurement of CD3 and CD68 staining. For CD3 staining, values of 1+, 2+, or 3+ corresponded to <5% T-cells, 5-15% T-cells, and 16-25% T-cells, respectively. CD68 staining values of 1+, 2+, or 3+ corresponded to <1% macrophages, 1-5% macrophages, and 6-10% macrophages, respectively.

Antibodies used for immunohistochemical stains.

Antibody	Clone	Source	Antigen Retrieval	Dilution
CD3	PS1	Novocastra Laboratories Ltd, Newcastle upon Tyne, UK	EDTA, pH 8.0	1:25
CD10	56C6	Novocastra Laboratories Ltd, Newcastle upon Tyne, UK	EDTA, pH 8.0	1:50
CD20	L26	Dako, Carpinteria, CA, USA	EDTA, pH 8.0	1:2000
CD68	KP1	Dako, Carpinteria, CA, USA	EDTA, pH 8.0	1:2000
CD138	B-A38	Serotec, Oxford, UK	EDTA, pH 8.0	1:200
BCL6	PG-B6p	Dako, Carpinteria, CA, USA	EDTA, pH 8.0	1:25
MUM1	MUM1p	Dako, Carpinteria, CA, USA	EDTA, pH 8.0	1:100
Ki-67	K2	Ventana Medical Systems, Tuscon, AZ, USA	EDTA, pH 8.0	Pre-diluted by vendor
BCL2	124	Dako, Carpinteria, CA, USA	EDTA, pH 8.0	1:50

The antigen retrieval "EDTA, pH 8.0" method refers to incubation at 100^{o}C for 60 min. in 0.1mM EDTA.

***BCL2/JH* Translocation PCR**
Genomic DNA isolated from the frozen DLBCL samples and DLBCL cell lines was subjected to PCR analysis for detection of the t(14;18). Assays were performed in duplicate using *BCL2* and immunoglobulin heavy chain gene joining region (*JH*) primers to detect (14;18) translocation major and minor breakpoint regions as previously described (Gulley, ML et al., Cancer 1992;69:1600-1606).

ArrayCGH:

Platform design: The miRTile platform was designed using the eArray interface (Agilent Technologies). In brief, the genomic coordinates from 474 miRNA genes (miRBase release V9.0, October 2006) were extracted and used to search for all oligonucleotide probes available in the Agilent catalogue located within a ~ 35kb area surrounding each precursor miRNA sequence. Probes were available to cover all but 3 miRNA loci (miR511-1, miR-511-2 and miR-550-2) included in the miRBase release V9.0. Thus, our platform effectively tiled the loci for 471 miRNAs. In addition, probes mapping to the entire genomic locus of 17 mRNA genes that are known to play a role in microRNA biogenesis were also identified and included in this platform (Supplementary Table 2). These high density tilling arrays were printed in the 4 X 44K slide format including 41631 oligoprobes (~55bp long, 44-59bp) with a total genomic coverage of 11.6 Mb encompassing 309 unique chromosomal regions Backbone probes between microRNA gene loci were included in the array at approximately 1 Mb intervals.

Hybridization and data retrieval: High molecular weight DNA from DLBCLs and sex-matched normal control DNAs (Promega) were differentially labeled with Cy3 and Cy5 and co-hybridization to the array CGH platform, as recommended by the manufacturer (Agilent Technologies). To test the performance of our platform, two independent pairs of normal male and female DNAs were used in dye swap experiments (totaling four hybridizations). Subsequently, the arrays were scanned and the intensity of the hybridization signals obtained using the Feature Extraction software (Agilent Technologies).

Data analysis: The log10 data of the 44K CGH probes (LogRatio column in the individual text files) of all array CGH samples were organized into a single data file. The data were read into to the dChip software (www.dchip.org, version 6/7/08) and the log10 ratios were converted to log2 ratios. The probe chromosome and position information were used to visualize the data in dChip according to probe genome positions. ***Data quality check***. We inspected the sample quality visually according the log2 ratio curves and excluded 13 samples with noisy or very low signals throughout the genome. The 77 remaining samples were used in the following analyses. First, for each probe, we computed the median log2 ratio across all samples and subtracted this median value from all samples. This step removed regional systematic gain or loss patterns, which are likely due to the bias in the hybridization process or the control channel sample. Second, data quality verification was performed for the each probe included in the array. We observed the existence of "spike probes", which showed significantly outlying higher or lower signals comparing to their neighboring probes in multiple samples (defined as absolute

value of log 2 ratio >= 1 in more than 20% of samples). Thus, we computed a correlation score across all samples between each probe data vector and the average data vector of all "spike probes"; a threshold >0.5 was then used to exclude these underperforming probes (n= 4158) from the downstream analysis. Thus, the actual number of probes in miRTile is 37113 and its resolution (mean probe spacing) in the focally covered areas is 319bp. ***Obtaining the miRNA-level data***. To obtain the miRNA genes copy number data we first organized the probes into miRNA regions. Each region extends the positions of the precursor miRNAs by ~ 17 Kb in the upstream and downstream directions to include all the CGH probes in that region; two or more such extended areas could be merged into one miRNA region (due to the common clustering of miRNA genes) containing multiple miRNA genes. The log2 ratio values of all the probes in a region were averaged to obtain the miRNA region-level log2 ratios and copy number summary plots displayed the proportion of the samples that have copy number gain (>=2.5) or loss (<=1.5) for all miRNA regions. ***Finding significantly altered regions – permutation analysis.*** To find statistically significant altered miRNA loci in all samples we tested the null hypothesis that no one miRNA is enriched by amplifications or deletions. Under this null hypothesis, a permutated dataset was constructed as follows. For each sample, we randomly shuffled the miRNA genomic positions while preserving the copy number correlations of most neighboring miRNAs. The permutation was performed for each sample independently creating a permuted dataset. For the permuted dataset, a gain (or loss) proportion score was computed for each miRNA, and the largest score across the genome was recorded. The permutation was carried out 1000 times, and the 1000 largest scores formed a distribution. The 95^{th} percentile of the 1000 largest scores (genome-wide p-value threshold of 0.05) was used to call significant regions of gain or loss in the original dataset. ***Finding significantly different regions between sample groups***. We used Fisher's exact test (p-value cutoff of 0.05) to identify differential copy number loss and gain between sample groups (e.g. nodal vs. extra-nodal tumors, GCB vs non-GCB and primary tumors vs. cell lines) as determined by the miRNA-level copy number data (Supplementary Table 3). Copy >= 2.3 and <= 1.7 were considered gain and loss respectively. For male chromosome X (the normal copy number is 1), these thresholds are divided by half. Gain and loss comparisons were performed separately.

MicroRNA expression analysis: Training set

Platform: Genome-wide determination of miRNA expression was performed using the Human miRNA Microarray Kit (V1, Agilent), with probe sets for 470 human and 64 viral microRNA genes (miRBase release v 9.1, February 2007), approximately 15,000 features/slide and 16-20 replicate probes/miRNA

Hybridization and data retrieval: Total RNA was isolated from 21 frozen DLBCL samples using Trizol and hybridized (100ng, single channel, Cy3-labeled) to microarray slides (8 complete microarrays/slide) and washed following manufacturer's guidelines. The arrays were subsequently scanned and the intensity of the hybridization signals obtained using the Feature Extraction software (Agilent Technologies).

Data Analysis: The gene-level signal values for the miRNAs were extracted from individual text files generated by the Feature Extraction software. Data for 470 human miRNAs were used in the downstream analysis. To adjust for hybridization differences between arrays, normalization was performed to scale all the signal values in one array by a multiplicative factor, so that the 75^{th} percentile of the signal values in each array is the same. Subsequently, to focus our analyses towards those miRNAs likely to be more relevant in DLBCL biology, we filtered the miRNA genes whose signal values were higher than 50 in more than 50% of the samples and used these genes for clustering analyses. The hierarchical clustering function in the dChip software was used, which standardizes the gene expression values for each gene first, and then computes the gene correlation and sample correlations as distances used in an average-linkage clustering. As shown in figure 2 and supplementary table 7, data from 21 DLBCL (training set) were used in this process, which led to identification of three major subsets of DLBCL, named MiRNA Groups (MG-) A, B and C. We next used one-way ANOVA to obtain genes from the initial 98 miRNAs list that significantly distinguish the three sample groups ($p<0.01$, 38 genes are obtained; $p<0.001$, 16 genes are obtained). This smallest gene set was the basis for the quantitative real-time RT-PCR studies described below (validation set).

Fluorescence in situ hybridization (FISH): was used to validate our copy number calling algorithm. Bacterial artificial chromosome (BAC) probes mapping to miRNA loci that commonly exhibited gains (miR-26a-2 on chromosome 12q13, clones RP11-571M6 and RP11-155I23) or losses (miR-16-1 on chromosome 13q14, clone RP11-34F20) were indentified using electronic mapping (NCBI MapViewer), and hybridized to selected DLBCL cell lines and primary tumors (Interphase-FISH)

Stem-loop quantitative real-Time PCR for mature microRNAs.
Total RNA isolated from thirty-two additional DLBCL samples (validation set) and, together with 10 tumors included in the training set described above, analyzed by quantitative real-time RT-PCR (TaqMan MicroRNA Assays, Applied Biosystems) The following mature miRNAs were included in these studies: miR-17-3p, miR-17-5p, miR-18a, miR-19a, miR-19b, miR-20a, miR-20b, miR-92, miR-106a, miR-24, miR-130a, miR-100, miR-199a, miR-15a, miR-16, miR-29a, miR-29c, miR-140. Results were normalized to the expression levels of three different small nucleolar RNAs RNU6, relative expression defined by the $2^{-\Delta\Delta CT}$ method and results loaded into the dCHIP software for hierarchical clustering analysis.

Kaplan-Meier survival probability curves: Kaplan-Meier survival curves for 30 DLBCL patients treated with anthracycline-based regimens and assigned to MG-A (n=8, 5 events, 3 censored), MG-B (n=11, 4 events, 7 censored) and MG-C (n= 11, 5 events, 6 censored) were calculated. The significance of the differences between these survival curves was determined by the log-rank test. However, given the relatively small sample size and the possibility of inappropriate distribution, we also used permutation methods to compute the p-value of the observed test statistic, without assuming the chi-square distribution. Specifically, during each permutation we randomly assign the 30 samples

into the MG-A, -B, and -C groups of 8, 11, and 11 patients, respectively. Then, we computed the log-rank test statistic for the permuted dataset. After 50,000 permutations, **these test statistics** determined the empirical distribution of the test random variable under the null hypothesis that the MG groups have the same survival distributions. A permutation p-value was determined as the proportion of the permutated test statistics that was larger than the actually observed.

Integration of copy number and expression data.
To compute the correlation between miRNA expression values and copy number, the miRNA-level arrayCGH regions were matched to the expressed miRNA genes. Only genes with expression signal >= 50 in more than 30% samples were included in the downstream analyses. For each gene, we used 0 - 1.7, 1.7 - 2.3 and 2.3 - 100 to divide copy numbers into three bins (loss, diploid, gain). When there were more than two bins and each bin had two or more samples, the gene was considered for further analyses. Subsequently, the expression distribution of each gene in the three copy number bins was measured and the top ranked genes, ANOVA p-value < 0.01 when testing the expression differences among the two or three bins, shown in supplementary figure 10. A cohort of 17 DLBCL analyzed by both genome-wide miRNA expression profiling and arrayCGH was suitable for these studies.

Real-time RT-PCR for MYC targets.
Quantitative real-time RT-PCR was used to define the expression of three MYC target genes, *CAD* (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase) *PGK1*, (phosphoglycerate kinase 1) and *TFAM* (transcription factor A, mitochondrial) in 24 primary DLBCL samples. Values were normalized by the expression of Cyclophilin A. All reactions were run in triplicate and relative expression defined as $2^{-\Delta\Delta CT}$. Significance of the differences in expression between tumors assigned to MG-A and MG-C was determined with the Mann-Whitney Test.

Oligonucleotides used are listed below:

PGK1-F 5’ GATCATTGGTGGTGGAATGG 3’
PGK1-R 5’ AGTAGCTTGGCCAGTCTTGG 3’
TFAM-F 5’AATGGATAGGCACAGGAAACC 3’
TFAM-R 5’CAAGTATTATGCTGGCAGAAGTC 3’
CAD-F 5’ GCGGTGCTGCTATGAATGTG 3’
CAD-R 5’ ATGCTCAGAGATGGCGATGG 3’