Supplementary Information for

DIFFUSE LARGE B-CELL LYMPHOMA OUTCOME PREDICTION BY GENE EXPRESSION PROFILING AND SUPERVISED MACHINE LEARNING

Margaret A. Shipp, Ken N. Ross, Pablo Tamayo, Andrew P. Weng, Jeffery L. Kutok, Ricardo C. T. Aguiar, Michelle Gaasenbeek, Michael Angelo, Michael Reich, Geraldine S. Pinkus, Tane S. Ray, Margaret A. Koval, Kim W. Last, Andrew Norton, T. Andrew Lister, Jill Mesirov, Donna S. Neuberg, Eric S. Lander, Jon C. Aster, and Todd R. Golub

December 6, 2001

Contents:

Section I: Expanded Methods

This document provides supplementary and detailed analysis information not included in the paper. Other sources of information and the original data sets can be found in our web site www-genome.wi.mit.edu/MPR/lymphoma.

Primary Lymphoma Specimens and Clinical Information

Frozen diagnostic nodal tumor specimens from 58 DLBCL patients and 19 FL patients were selected for these initial studies. A summary of the clinical data for the patients can be found in the ***List of all samples*** section of the document. The histopathology and immunophenotype of each tumor specimen was reviewed to confirm diagnosis and uniform involvement with tumor. Treatment records of all 58 DLBCL patients were reviewed to confirm that patients had received adequate doses of CHOP-like combination chemotherapy[1] for 6 or more cycles or until documented disease progression and to document outcome and clinical IPI risk group[14]. All tumor samples were obtained from diagnostic lymph node biopsies prior to treatment. The samples were snap frozen in liquid nitrogen and stored at -80°C. DLBCL study patients had representative IPI-risk profiles and disease-free and overall survivals (OS). The IPI was not determined in 2 patients because of missing LDH levels in these patients. DLBCL study patients (predicted 5 year OS 54%, median follow-up 58 months) were divided into 2 discrete categories: 1) 29 patients who achieved CR and remained free of disease plus 3 additional patients who died of other causes (total 32 “cured” patients); and 2) 23 patients who died of lymphoma plus 3 additional patients who remained alive with recurrent refractory or progressive disease (total “fatal/refractory” 26 patients).

Microarray Hybridization

For a detailed protocol, see http://www-genome.wi.mit.edu/MPR/. Total RNA was extracted from each frozen tumor specimen and converted to double-stranded cDNA as previously described[2]. Briefly, tissue samples were homogenized (Polytron, Kinematica, Lucerne) in guanidinium isothiocyanate and RNA was isolated by centrifugation over a CsCl gradient. RNA integrity was assessed either by northern blotting or by gel electrophoresis. The amount of starting total RNA for each reaction varied between 10 and 12 µg. First strand cDNA synthesis was generated using a T7-linked oligo-dT primer, followed by second strand synthesis. An in vitro transcription reaction was done to generate the cRNA containing biotinylated UTP and CTP, which was subsequently chemically fragmented at 95°C for 35 minutes. Ten micrograms of the fragmented, biotinylated cRNA was hybridized in MES buffer (2-[N-Morpholino]ethansulfonic acid) containing 0.5 mg/ml acetylated bovine serum albumin (Sigma, St. Louis) to Affymetrix (Santa Clara, CA) HU6800 oligonucleotide arrays[3] at 45°C for 16 hours. HuGeneFL arrays contain 5920 known genes and 897 expressed sequence tags. Arrays were washed and stained with streptavidin-phycoerythrin (SAPE, Molecular Probes). Signal amplification was performed using a biotinylated anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) at 3 µg/ml. This was followed by a second staining with SAPE. Normal goat IgG (2 mg/ml) was used as a blocking agent. Scans were performed on Affymetrix scanners and the expression value for each gene was calculated using Affymetrix

GENECHIP software. Minor differences in microarray intensity were corrected using a linear scaling method as detailed in the next section.

Preprocessing and Re-scaling

The raw expression data as obtained from Affymetrix's GeneChip is re-scaled to account for different chip intensities. Each column (sample) in the data set was multiplied by ***1/slope*** of a least squares linear fit of the sample vs. the reference (the first sample in the data set). This linear fit is done using only genes that have 'Present' (P) calls in both the sample being re-scaled and the reference. (The P calls are calculated by Affymetrix’s GENECHIP software and each P call represents a gene with RNA “Present” as determined by the average difference analysis of expression measurements from a gene’s set of probes on the microarray.) The sample chosen as reference is a typical one (i.e. one with the number of "P" calls closer to the average over all samples in the data set).

A ceiling of 16,000 units was chosen for all experiments because it is at this level that we observe fluorescence saturation of the scanner; values above this cannot be reliably measured. We set a lower threshold for the expression levels to 20 units to minimize noise effects while avoiding missing any potentially informative marker genes.

These numbers are Affymetrix’s scanner “average difference” units. After this preprocessing, gene expression values were subjected to a variation filter that excluded genes showing minimal variation across the samples being analyzed. The variation filter tests for a fold-change and absolute variation over samples (comparing max/min and max-min with predefined values and excluding genes not obeying both conditions). For maximum/minimum fold variation, we excluded genes with less than 3-fold variation and, for maximum-minimum absolute variation, we excluded genes with less than 100 units absolute variation.

Supervised Learning

This is the methodology for building a supervised classifier that we followed:

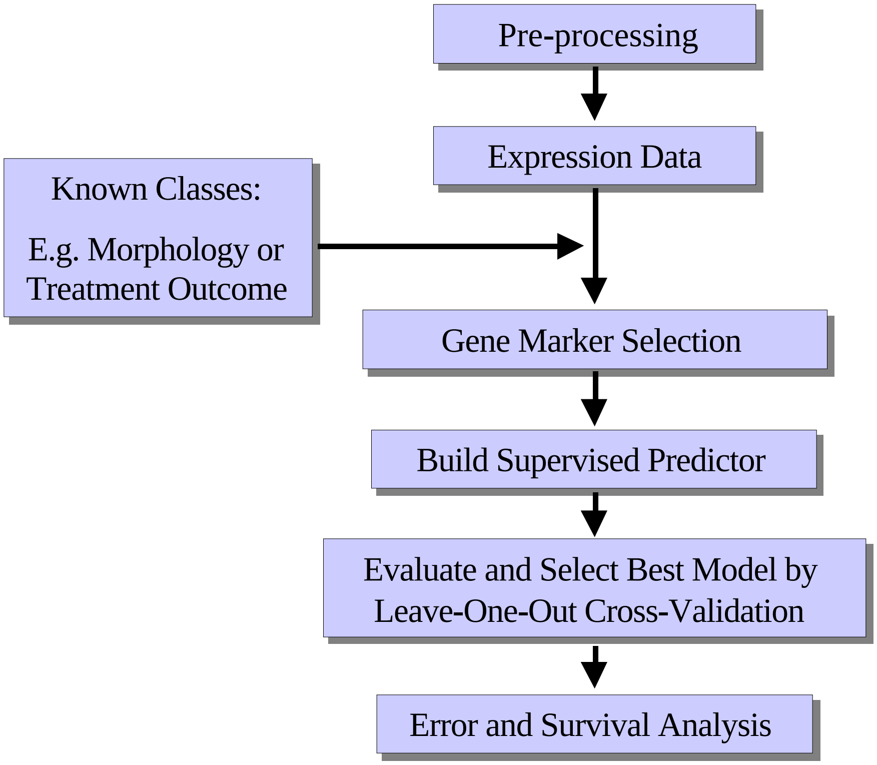
a) define a target class based on morphology, tumor class or treatment outcome clinical information;

b) select the “marker” genes with the highest correlation with the target class using a class separation statistic (signal-to-noise ratio). A permutation test is also applied to the top ranked genes to assess their class-correlation statistical significance;

c) build a classifier in cross-validation (leave-one-out) by removing one sample and then using the rest as a training set;

d) several models are built using different numbers of marker genes and the final chosen model is the one that minimizes the total error in cross-validation;

e) evaluate prediction results, compute confusion matrices and produce Kaplan-Meier survival plots.


Pre-processing
Expression Data
Known Classes:
E.g. Morphology or Treatment Outcome
Gene Marker Selection
Build Supervised Predictor
Evaluate and Select Best Model by Leave-One-Out Cross-Validation
Error and Survival Analysis

This methodology was used with the following algorithms: weighted voting (WV), k-nearest neighbors (KNN), and support vector machines (SVM). The details for each algorithm are described below.

Gene Marker Selection

Genes correlated with a particular class distinctions (e.g. class 0 and class 1) were identified by sorting all of the genes on the array according the signal-to-noise statistic[3,5] $(\mu_{class0} - \mu_{class1})/(\sigma_{class0} + \sigma_{class1})$ where μ and σ represent the mean and standard deviation of expression, respectively, for each class. Permutation of the column (sample) labels was performed to compare these correlations to what would be expected by chance (see the next section). These marker genes were used to build the k-nearest neighbor and weighted voting classifiers. SVM used different methods to select marker genes.

Permutation Test and Neighborhood Analysis for Marker Genes

A permutation test[5] was used to calculate whether the top marker genes with respect to a biologically meaningful phenotype (e.g. morphology) were statistically significant. To do this we compared the top signal-to-noise scores for top marker genes and compared them with the corresponding ones for random permutation versions of the class labels (phenotype). Typically 500 random permutations were used to build histograms for the top marker, the second best etc. Based on this histogram we determined the 50% (median), 5% and 1% significance levels and compared them with the values obtained for the real data set.

This procedure is motivated by considering the following question: what is the likelihood that the set of markers genes, for example selected by signal-to-noise or any other distance or correlation measure, of a phenotype of interest represent chance correlations and not any biological significant match? If one moves down the list of markers, how many could one consider as being significantly correlated and not the results of chance correlations?

In detail the permutation test procedure is as follows:

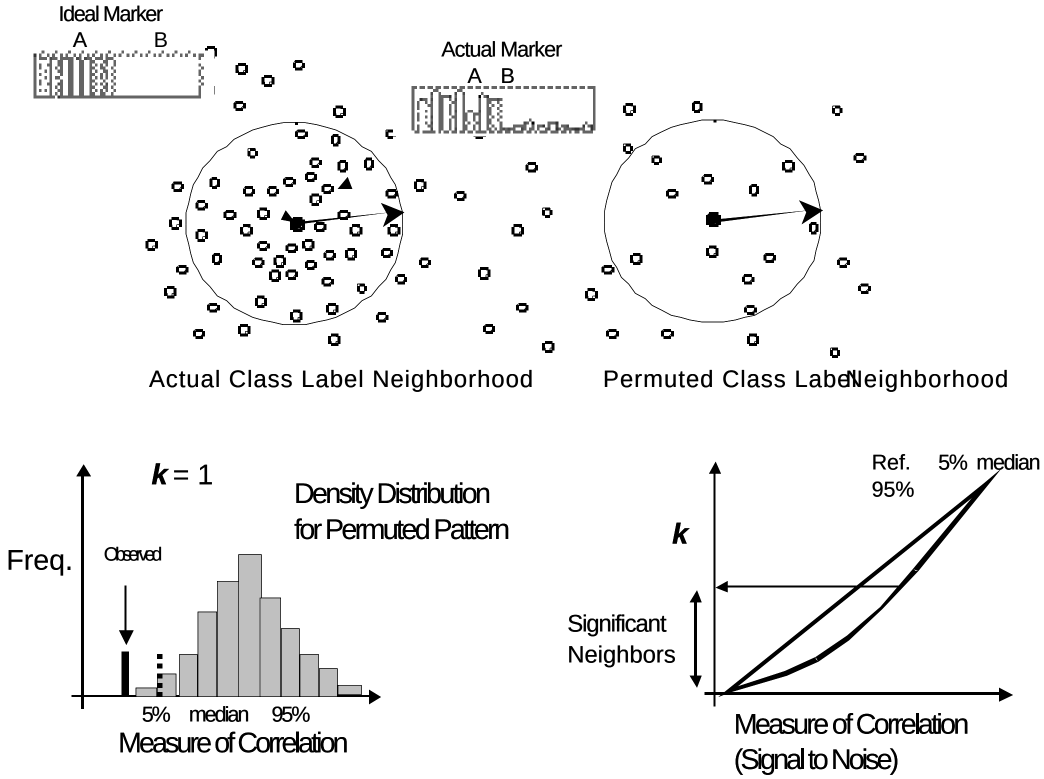
- Generate signal-to-noise $(\mu_{class0} - \mu_{class1})/(\sigma_{class0} + \sigma_{class1})$ scores for all genes that pass a variation filter using the actual class labels (phenotype) and sort them accordingly. The best match (k=1) is the gene “closer” or more correlated to the phenotype using the signal-to-noise as a distance function. In fact one can imagine the reciprocal of the signal-to-noise as a “distance” between the phenotype and each gene as shown in the figure (see next page).
- Generate 500 random permutations of the class labels (phenotype). For each case of randomized class labels generate signal-to-noise scores and sort genes accordingly.
- Build a histogram of signal-to-noise scores for each value of k. For example, one for all the 500 top markers (k=1), another one for the 500 second best (k=2) etc. These histograms represent a reference statistic for the best match, second etc. and for a given value of k different genes contribute to it. Notice that the correlation structure of the data is preserved by this procedure. Then for each value of k one determines the 50% (median), 5% and 1% significance levels. See the bottom diagrams in the figure.
- Compare the actual signal-to-noise scores with the different significance levels obtained for the histograms of permuted class labels for each value of k. This test helps to assess the statistical significance of gene markers in terms of target class-correlations.

In the results section the values for permutation tests of marker genes are reported in tables with this format:

Distinction	Distance	Perm 1%	Perm 5%	Median 50%	Feature	Desc
class 0	0.96694607	1.0144908	0.8333578	0.6280173	M93119_at	INSM1 Insulinoma-associated 1
class 0	0.9096911	0.8600172	0.7669801	0.5740431	M30448_s_at	Casein kinase II beta subunit
class 0	0.90010124	0.85051423	0.7251496	0.5494933	S82240_at	RhoE
class 0	0.832689	0.84354156	0.7071885	0.5292253	U44060_at	Homeodomain protein (Prox 1)
class 0	0.83225346	0.8009565	0.68034023	0.5169537	D80004_at	KIAA0182 gene
…………	…….	…….	…….	…….	…….	…….
class 1	1.6520017	0.9831643	0.84544426	0.6230137	X86693_at	High endothelial venule
class 1	1.2436218	0.88150144	0.7559189	0.5795857	M93426_at	PTPRZ Protein tyrosine phosphatase, receptor-type, zeta polypeptide
class 1	1.2317128	0.86047184	0.70928395	0.5539352	U48705_rna1_s_at	Receptor tyrosine kinase DDR gene
class 1	1.2259983	0.8433512	0.68909335	0.5358038	X86809_at	Major astrocytic phosphoprotein PEA-15
class 1	1.214929	0.8281318	0.6849929	0.5217813	U45955_at	Neuronal membrane glycoprotein M6b mRNA, partial cds
class 1	1.2095517	0.79365546	0.6711517	0.510208	U53204_at	Plectin (PLEC1) mRNA
…….	…….	…….	…….	…….	…….	…….

The **distinction** represents the class for which the markers are high (and low in the other classes). **Distance** is the signal to noise to the actual phenotype. **Perm. 1%, 5%** and **50%** and the corresponding percentiles (significance levels) in the histograms of random permutation signal to noise scores for a given value of k. **Feature** is the gene accession number and **Description** the gene name and annotation. Permutation test results are reported in the gene markers sections: ***Expression Profiles of DLBCL and FL*** and ***Expression Profiles of Cured and Fatal/Refractory Disease***.

Neighborhood Analysis: Assessing Statistical Significance of Gene-Class Correlations


Ideal Marker
A
B
Actual Marker
A B
Actual Class Label Neighborhood
Permuted Class Label Neighborhood
k = 1
Density Distribution for Permuted Pattern
Freq.
Observed
5%
median
95%
Measure of Correlation
Ref.
95%
5%
median
k
Significant Neighbors
Measure of Correlation (Signal to Noise)

Additional Notes:

- This test helps to assess the statistical significance of gene markers in terms of class-gene correlations but if a group of genes fails to pass the test that by itself does not necessarily imply that they cannot be used to build an effective classifier[6,7]. For example, in contrast with the case of morphological distinctions, for treatment outcome prediction the top marker genes do not show overwhelming statistical significance ("weak" markers) and yet they are effective when used in combination by the classifiers to provide statistically significant predictions.
- The choice of the signal-to-noise is somewhat ***ad hoc*** but not unreasonable as a choice of class distance. The reason the signal-to-noise ratio was chosen instead of a t-statistic or other class distance measures was mainly historical and empirical: it performed slightly

better in a previous study of gene expression feature selection combined with a weighted voting classifier.

- We deal with the problem of multiple hypotheses by performing a permutation test and use quantiles of the empirical distributions of rank signal-to-noise values to assess significance. This is a distribution-free approach that preserves the correlation structure of genes.
- The advantages of performing a permutation test are multiple:
  - It is a direct empirical to test the significance of the matching of a given phenotype to a particular set of genes (data set).
  - It doesn't assume a particular functional form for the distribution or correlation structure of genes.
  - As the permutation test is done on the entire distribution of genes (as scored by signal-to-noise from the phenotype) the gene-to-gene correlation structure is preserved and therefore one doesn't need to explicitly compensate for multiple hypothesis testing (for example by Bonferroni, Sidak's or some other procedure that makes strong assumptions about the distribution, correlations or independence of genes).
- Another more geometrical and sometimes more intuitive way to look at this procedure is to consider the figure above as a hypothetical projection of normalized gene expression space where each dimension represents an experiment and each data point a gene. The entire data set of filtered genes will be represented by a collection of data points distributed in that space. Each gene is represented by a point and the closer two points are the more correlated they are (i.e. across the set of experiments being considered). Now imagine projecting a point that corresponds to an ideal marker gene that perfectly represents the phenotype of interest. This is for example a marker gene that is high and constant in one of the classes and low and constant in the other. This gene will be a perfect classifier to distinguish the two classes. We are interested in finding marker genes that are if not equal at least similar to this ideal marker. This can be accomplished by computing a distance or correlation measure between the class labels (phenotype) and the genes. In this sense we are looking at the "neighborhood" of a phenotype in gene expression space trying to find "close" neighbors. A permutation test in this context is equivalent to moving the ideal gene point randomly (as the labels are permuted) and studying the distribution of neighbors each time it lands to a new reference point in expression space. By building a histogram of distance distributions to these random locations one can assess how "typical" is the actual neighborhood of the actual phenotype. For example if only once in a thousand random tries we found a set of top 10 markers as correlated as in the actual neighborhood then we will consider those markers to be significant.

Algorithms

Weighted Voting

The weighted voting algorithm[3,5] makes a weighted linear combination of relevant "marker" or "informative" genes obtained in the training set to provide a classification scheme for new samples. Target classes (classes 0 and 1) were initially defined based on morphology or treatment outcome. Class distinction was represented by an idealized expression pattern according to whether a sample belonged to class 0 or class 1 (e.g. follicular or large B-cell). The selection of features (marker genes) is accomplished by computing the signal-to-noise statistic $\boldsymbol{S_x}$ (described above). The class predictor is uniquely defined by the initial set of samples and marker genes. In addition to computing $\boldsymbol{S_x}$, the algorithm also finds the decision boundaries (half way) between the class means: $\boldsymbol{b_x} = (\mu_{class0} + \mu_{class1})/2$ for each gene. To predict the class of a test sample ***y***, each gene ***x*** in the feature set casts a vote: $\boldsymbol{V_x = S_x (g_x^y - b_x)}$ and the final vote for class 0 or 1 is ***sign*** $(\Sigma_x \boldsymbol{V_x})$***.*** The strength or ***confidence*** in the prediction of the winning class ***is*** $(\boldsymbol{V_{win}}\text{-}\boldsymbol{V_{lose}})/(\boldsymbol{V_{win}}+\boldsymbol{V_{lose}})$ (i.e., the relative margin of victory for the vote). For our lymphoma outcome "cured" versus "fatal/refractory" experiments, the weighted models were evaluated by 58-fold leave-one-out cross-validation[3,5] whereby a training set of 57 samples was used to predict the class of a randomly withheld sample. This was repeated for all samples and the cumulative error rate was recorded. Thereafter, the total number of prediction errors in cross-validation was calculated and a final model chosen which minimized cross-validation errors. Detailed prediction results are in the sections: DLBCL versus FL Prediction and DLBCL Outcome Prediction

k-Nearest Neighbors (KNN)

We developed a weighted implementation of the KNN algorithm[8] that predicts the class of a new sample by calculating the Euclidean distance (d) of this sample to the ***k*** "nearest neighbor" standardized samples in "expression" space in the training set, and by selecting the predicted class to be that of the majority of the ***k*** samples (the method is defined in terms of Euclidean distances over standardized vectors so it is equivalent to using inner products: $\mathbf{a} \,.\, \mathbf{b} / |a||b|$). We performed the marker gene selection process by which we feed the KNN algorithm only the features with higher correlation with the target class. This feature selection is done by sorting the features according to the signal-to-noise statistic[3,5] $(\mu_{class0} - \mu_{class1})/(\sigma_{class0} + \sigma_{class1})$. In our version of the algorithm, the weight of each of the ***k*** neighbors was weighted according to 1/d. For our lymphoma outcome "cured" versus "fatal/refractory" experiments, the KNN model was evaluated by sequentially removing one sample at a time and using the remainder of samples as the training set. This was repeated for all samples and the cumulative error rate was recorded. The detailed results of applying this algorithm to the lymphoma outcome prediction can be found in the DLBCL Outcome Prediction section.

Support Vector Machines

The Support Vector Machine (SVM) for classification minimizes the generalization error rather than the training error. The basic idea behind SVMs is to construct an optimal separating hyperplane by mapping the gene expression data to a high-dimensional space[9,10]. Linear separation in this higher dimensional space corresponds to a nonlinear decision boundary in the original space. A new feature selection algorithm was developed to scale the input features to minimize the ratio of the radius around the support vectors and the margin (Weston et al.[11]).

The Weston et al algorithm for feature selection used in the SVM is basically a compromise between filtering methods and wrapper methods for feature selection. Filtering methods, like our signal-to-noise ratio, rely on a preprocessing step that occurs before the model is created and operate by trying to remove irrelevant features. Wrapper methods search through the space of feature subsets using the estimated accuracy from the prediction algorithm (in this case, on a held out subset of the data) as a measure of the goodness of a particular feature subset. Generally wrapper methods provide better performance than filtering methods but they are much more computationally expensive because the prediction algorithm must be evaluated on each feature subset. The Weston et al. feature selection algorithm is based upon an approximation of the wrapper method that uses a gradient descent method to minimize the expectation of the leave-one-out error. The expectation of the leave-one-out error is bounded by the ratio of the radius around a subset of the training data called support vectors to the distance between the two nearest points of opposite classes. Using a gradient descent algorithm, the feature selection method scales the input features to minimize the ratio described above and iteratively eliminates the features corresponding to a small-scale parameter.

The detailed results from using the SVM to predict outcome are in the DLBCL Outcome Prediction section.

Proportional Chance Criterion

In order to compute p-values for non-survival predictions, for example the p-val=10^{-9} for the DLBCL vs. FL classifier reported in the paper (71 out of 77 samples correctly classified) we used a “proportional chance criterion” to evaluate the probability that a random predictor will produce a confusion matrix with the same row and column counts as the gene expression predictor. This approach considers the question of how well classes are discriminated by formulating a likelihood ratio to estimate chance classification. For example, for a binary class (A vs. B) problem, if α is the prior probability of a sample being in class A and $\boldsymbol{p}$ is the true proportion of samples in class A then $\boldsymbol{C_p = p}\,\alpha + (\boldsymbol{1\text{-}p})\,(\boldsymbol{1}\text{-}\alpha)$ is the proportion of the overall sample that is expected to receive correct classification by chance alone. Then if $\boldsymbol{C_{model}}$ is the proportion of correct classifications achieved by the gene expression predictor one can estimate its significance by using a Z statistic of the form: $(\boldsymbol{C_{model} - C_p})/\mathrm{Sqrt}(\boldsymbol{C_p}\,(\boldsymbol{1\text{-}C_p})/\boldsymbol{n})$, where $\boldsymbol{n}$ is the total sample count. For more details see chapter VII of Huberty’s ***Applied Discriminant Analysis***[6].

Survival Analysis and Kaplan-Meier Plots

The Kaplan-Meier survival analysis plots[12] are computed using the S-Plus (http://www.insightful.com/products/) statistical software package: S-Plus 2000, Guide to Statistics Volume 2, chapter 9. The p-values for the prediction of outcome groups are computed using a log-rank test (Mantel-Haenszel method, chapter 9 in the same reference). The Kaplan Meier plots and associated rank test p-values are included in the DLBCL Outcome Prediction and the ***In Silico*** Model Validation sections.

Analysis of Lymphochip Microarray Data

Detailed descriptions of the procedure used to perform an ***In Silico*** validation that explored the connection between the cell-of-origin classification described by Alizadeh et al.[13] and the lymphoma outcome predictor developed by this paper are contained in the section titled ***In Silico*** Model Validation.

Unsupervised Learning: Hierarchical Clustering

Hierarchical Clustering is a method for performing unsupervised learning (i.e., learning models for classifying data where the true class for the data samples is assumed to be unknown prior to model training) useful for dividing data into natural groups. Data is clustered hierarchically by organizing the data into a tree structure based upon the degree of similarity between features. We used the Cluster and TreeView software[4] (available from http://www.microarrays.org/) to perform average linkage clustering, which organizes all of the data elements into a single tree with the highest levels of the tree representing the discovered classes.

Immunohistochemical Staining

Five representative 0.6 mm cores were obtained from diagnostic areas of each paraffin-embedded formalin-fixed DLBCL and inserted in a grid pattern in a single recipient paraffin block using a tissue arrayer (Beecher Instruments, Silver Spring, MD). Five micron sections cut from this “tissue array” were stained for PKCb using an immunoperoxidase method. Briefly, slides were deparaffinized and pre-treated in 1 mM EDTA, pH 8.0, for 20 minutes at 95°C. All further steps were performed at room temperature in a hydrated chamber. Slides were pre-treated with Peroxidase Block (DAKO, USA) for 5 minutes to quench endogenous peroxidase activity, and a 1:5 dilution of goat serum in 50 mM Tris-Cl, pH 7.4, for 20 minutes to block non-specific binding sites. Primary antibody (murine monoclonal antibody specific for PKCb (Serotec, UK)) was applied at a 1:1000 dilution in 50 mM Tris-Cl, pH 7.4 with 3% goat serum for 1 hour. After washing, secondary goat anti-mouse horseradish peroxidase-conjugated antibody (Envision detection kit, DAKO, USA) was applied for 30 minutes. After further washing, immunoperoxidase staining was developed using a DAB chromogen kit (DAKO, USA) per the manufacturer. Following counterstaining with hematoxylin, immunoperoxidase staining within the malignant cell population of each core was scored in a blinded fashion with respect to clinical outcome and expression profile results by three experienced hematopathologists (JCA, AW, JLK). The intensity of staining on each core was graded from 0 (no staining) to 3 (maximal staining), and an average staining intensity (the mean of all five cores) was generated for each tumor. The p-value for the association between immunostaining intensities and the array-based transcript levels was evaluated by using median to divide measured intensities into two levels and then using the Fisher exact test to evaluate the degree of association between the quantized measurements.

Section II: Datasets and Clinical Attributes

This section of the document describes the samples, clinical attributes and data sets in detail. Two data sets were formed out of the samples listed below: (1) a combined diffuse large B-cell lymphoma (DLBCL) and follicular lymphoma (FL) for identifying tumors within a single (B-cell) lineage, and (2) a set made up of just the DLBCL samples to distinguish the cured versus fatal/refractory cases. These data sets are available on our website (http://www-genome.wi.mit.edu/MPR/lymphoma). The following table shows a list of samples analyzed for this paper and associated clinical information. This table can also be downloaded from the supplemental information website.

List of all samples

Sample	FULL IPI	SURTIME	STATUS	OUTCOME
DLBC1	Low	72.9	Alive w/o disease	0
DLBC2	Low	143.1	Alive w/o disease	0
DLBC3	Low intermediate	144.2	Alive w/o disease	0
DLBC4	High intermediate	61	Alive w/o disease	0
DLBC5	Low	86.5	Alive w/o disease	0
DLBC6	Low	84.2	Alive w/o disease	0
DLBC7	High intermediate	112.5	Alive w/o disease	0
DLBC8	Low	133.2	Alive w/o disease	0
DLBC9	Low	22.1	Alive w/o disease	0
DLBC10	Low intermediate	182.4	Alive w/o disease	0
DLBC11	Low	66.4	Alive w/o disease	0
DLBC12	.	146.8	Alive w/o disease	0
DLBC13	Low intermediate	62.9	Alive w/o disease	0
DLBC14	Low intermediate	50.9	Alive w/o disease	0
DLBC15	Low	26.3	Alive w/o disease	0
DLBC16	.	48.6	Alive w/o disease	0
DLBC17	High intermediate	55.9	Alive w/o disease	0
DLBC18	Low	12.6	Dead w/o disease	0
DLBC19	Low intermediate	50.2	Dead w/o disease	0
DLBC20	High intermediate	58	Alive w/o disease	0
DLBC21	Low intermediate	66.4	Alive w/o disease	0
DLBC22	Low	65.7	Alive w/o disease	0
DLBC23	Low	50.2	Alive w/o disease	0
DLBC24	Low	26.9	Dead w/o disease	0
DLBC25	Low	34.4	Alive w/o disease	0
DLBC26	Low	26	Alive w/o disease	0
DLBC27	Low	30	Alive w/o disease	0
DLBC28	Low intermediate	31.7	Alive w/o disease	0
DLBC29	Low	32.2	Alive w/o disease	0
DLBC30	Low	19.2	Alive w/o disease	0
DLBC31	Low	33	Alive w/o disease	0
DLBC32	Low	21.4	Alive w/o disease	0
DLBC33	Low	15.7	Dead w/disease	1
DLBC34	High intermediate	11.6	Dead w/disease	1
DLBC35	High intermediate	3.4	Dead w/disease	1
DLBC36	Low	36.6	Dead w/disease	1

DLBC37	High intermediate	5.0	Dead w/disease	1
DLBC38	Low	9.5	Dead w/disease	1
DLBC39	High	3.2	Dead w/disease	1
DLBC40	Low intermediate	4.9	Dead w/disease	1
DLBC41	High intermediate	12	Dead w/disease	1
DLBC42	High intermediate	4.9	Dead w/disease	1
DLBC43	High intermediate	60.4	Dead w/disease	1
DLBC44	Low intermediate	16.3	Dead w/disease	1
DLBC45	High intermediate	16.4	Dead w/disease	1
DLBC46	High intermediate	9.5	Dead w/disease	1
DLBC47	High intermediate	15.6	Dead w/disease	1
DLBC48	High intermediate	17.8	Dead w/disease	1
DLBC49	Low intermediate	56.9	Dead w/disease	1
DLBC50	Low	13.3	Dead w/disease	1
DLBC51	Low intermediate	12.3	Dead w/disease	1
DLBC52	Low	44.6	Alive w/disease	1
DLBC53	High intermediate	4.6	Dead w/disease	1
DLBC54	High	7.5	Dead w/disease	1
DLBC55	High intermediate	19.3	Dead w/disease	1
DLBC56	Low	30.1	Dead w/disease	1
DLBC57	Low	33.6	Alive w/disease	1
DLBC58	High intermediate	13.9	Dead w/disease	1
FSCC1				
FSCC2				
FSCC3				
FSCC4				
FSCC5				
FSCC6				
FSCC7				
FSCC8				
FSCC9				
FSCC10				
FSCC11				
FSCC12				
FSCC13				
FSCC14				
FSCC15				
FSCC16				
FSCC17				
FSCC18				
FSCC19				

Clinical Information Definitions:

Sample – The coded identifier for the sample where a sample id of the form DBLC# represents a sample from a patient with diffuse large B-cell lymphoma and a sample id of the form FSCC# represents a sample from a patient with follicular lymphoma.

FULL IPI – Full International Prognosis Index[14] (high, high intermediate (hint), low intermediate (lint), or low).

SURTIME – The patient’s survival time in months from diagnosis to the latest follow-up.

STATUS – The patient’s current (at the last follow-up) disease status (alive or dead with or without disease).

OUTCOME – DLBCL study patients were divided into two discrete categories. A “0” signifies patients who achieved complete remission and remain free of disease (alive without disease) or patients who achieved complete remission and died of other causes (dead without disease). A “1” signifies patients who died of lymphoma (dead with disease) or remain alive with recurrent refractory or progressive disease (alive with disease).

Section III: Detailed Analysis Results

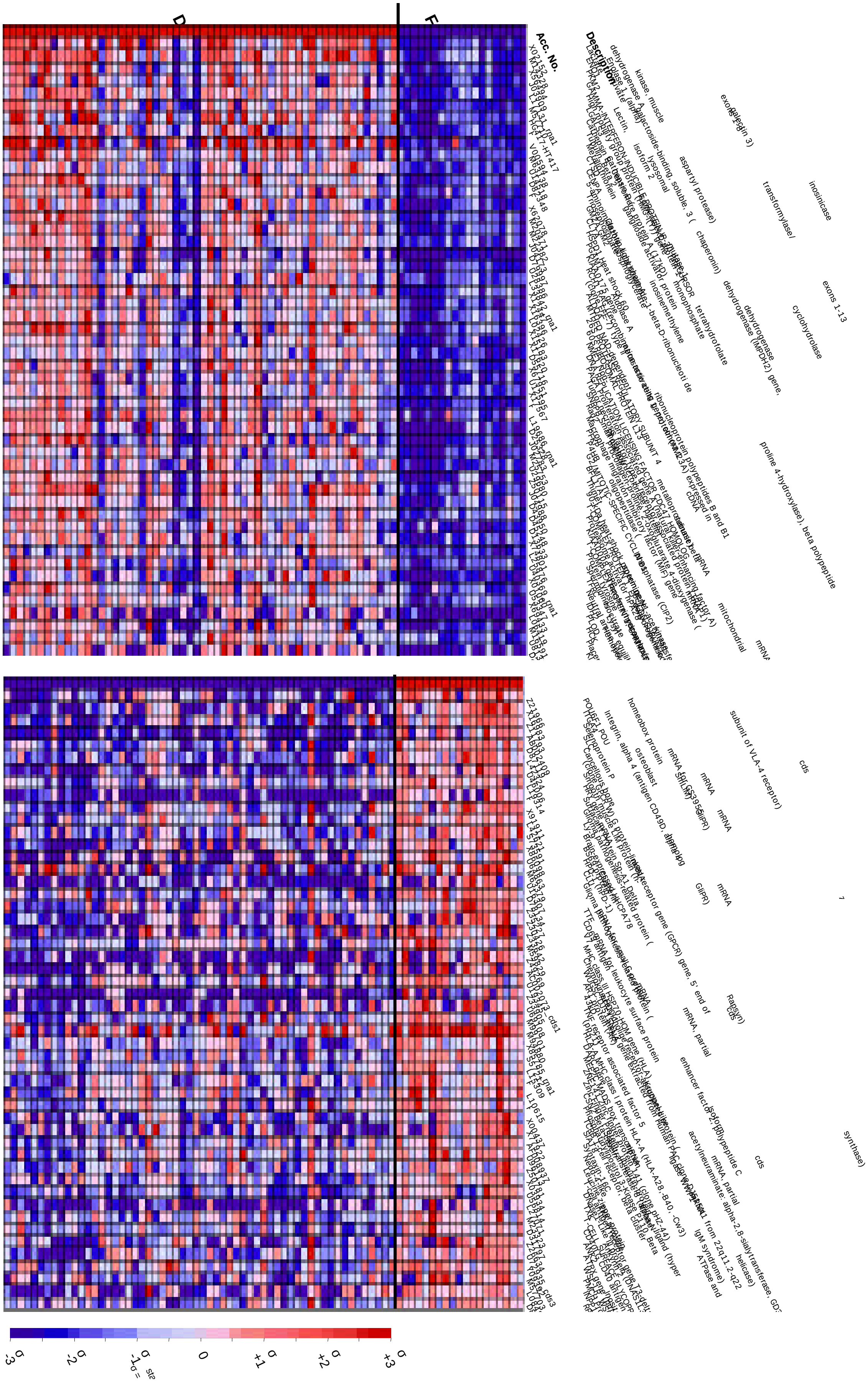
This section presents the results of applying the methods of section I to the data sets of section II. A brief comment precedes each table of results.

DLBCL versus FL Distinction

Within this section, we expand on the Diffuse Large B-Cell Lymphoma (DLBCL) versus Follicular Lymphoma (FL) analysis of the paper. The first subsection presents a pink-o-gram showing the expression profiles of the top 50 genes for DLBCL and FL and the permutation tests associated with those genes. In the next subsection, we show the results from predicting the DLBCL versus FL distinction.

Expression Profiles of DLBCL and FL

This section expands on Figure 1 from the paper. This picture shows the top 50 markers per class for the DLBCL versus FL distinction as sorted by their signal-to-noise ratios (using mean) as described in ***Gene Marker Selection*** section. The genes that were expressed at higher levels in DLBCL are shown on top while the genes that were more highly expressed in FL are shown on the bottom. Red indicates a high relative expression while blue represents a low relative expression. Each column is a sample and each row is a gene (with the first rows of the DLBCL and FL sections showing an idealized expression profile). Expression profiles for the 58 DLBCL samples are on the left while the profiles for the 19 FL samples are on the right. The pink-o-gram and table below show the top 50 markers for each tumor class. The table below the pink-o-gram shows the permutation test values (see ***Permutation Test and Neighborhood Analysis for Marker Genes***) for the top 50 markers for each tumor class. Standard preprocessing was used for the data where expression values were thresholded to 20 from below and 16000 from above and a variation filter removed non-changing genes (genes were filtered out if either maximum/minimum<3 (3-fold variation) or maximum-minimum<100 absolute units).


Acc. No.
Description
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ

Distinction	Distance	Perm 1%	Perm 5%	Perm 50%	Feature	Description
DLBCL	1.25	0.725783	0.621883	0.498348	X02152_at	LDHA Lactate dehydrogenase A
DLBCL	1.12	0.631373	0.568993	0.463286	M14328_s_at	ENO1 Enolase 1, (alpha)
DLBCL	1.06	0.6084	0.550469	0.444176	X56494_at	PKM2 Pyruvate kinase, muscle
DLBCL	1.02	0.600469	0.529616	0.430845	J03909_at	GAMMA-INTERFERON-INDUCIBLE PROTEIN IP-30 PRECURSOR
DLBCL	1.02	0.591185	0.522587	0.41867	L17131_rna1_at	High mobility group protein (HMG-I(Y)) gene exons 1-8
DLBCL	0.99	0.579695	0.504343	0.410885	M57710_at	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
DLBCL	0.98	0.56105	0.500048	0.402829	HG417-HT417_s_at	Cathepsin B
DLBCL	0.98	0.55919	0.494779	0.398693	HG1980-HT2023_at	Tubulin, Beta 2
DLBCL	0.94	0.551724	0.493136	0.393305	V00594_s_at	Metallothionein isoform 2
DLBCL	0.94	0.549168	0.489562	0.388675	M63138_at	CTSD Cathepsin D (lysosomal aspartyl protease)
DLBCL	0.94	0.544988	0.483049	0.383786	U14518_at	CENPA Centromere protein A (17kD)
DLBCL	0.93	0.539093	0.478518	0.379424	D82348_at	5-aminoimidazole-4-carboxamide-1-beta-D-ribonucleoti de transformylase/inosinicase
DLBCL	0.93	0.535533	0.475344	0.375516	HG2279-HT2375_at	Triosephosphate Isomerase
DLBCL	0.93	0.529515	0.472461	0.373096	X62078_at	GM2A GM2 ganglioside activator protein
DLBCL	0.92	0.528798	0.468316	0.370132	M20471_at	CLTA Clathrin light chain A
DLBCL	0.91	0.522234	0.465377	0.366263	M22382_at	HSPD1 Heat shock 60 kD protein 1 (chaperonin)
DLBCL	0.9	0.522206	0.458481	0.364112	J04173_at	PGAM1 Phosphoglycerate mutase 1 (brain)
DLBCL	0.9	0.518696	0.457256	0.360795	D79997_at	KIAA0175 gene
DLBCL	0.89	0.517013	0.456813	0.359301	U28386_at	RCH1 RAG (recombination activating gene) cohort 1
DLBCL	0.89	0.514614	0.454409	0.355618	L33842_rna1_at	(clone FFE-7) type II inosine monophosphate dehydrogenase (IMPDH2) gene, exons 1-13
DLBCL	0.89	0.514131	0.45277	0.354263	X12447_at	ALDOA Aldolase A
DLBCL	0.88	0.513357	0.452086	0.352307	X16396_at	MTHFD NAD-dependent methylene tetrahydrofolate dehydrogenase cyclohydrolase
DLBCL	0.88	0.509015	0.450906	0.349564	L02426_at	26S PROTEASE REGULATORY SUBUNIT 4
DLBCL	0.87	0.507412	0.448108	0.347601	X15183_at	60S RIBOSOMAL PROTEIN L13
DLBCL	0.87	0.506322	0.447786	0.345691	X17620_at	NME1 Non-metastatic cells 1, protein (NM23A) expressed in
DLBCL	0.87	0.506011	0.447455	0.344497	D55716_at	DNA REPLICATION LICENSING FACTOR CDC47 HOMOLOG
DLBCL	0.86	0.505555	0.443096	0.342214	X67951_at	PAGA Proliferation-associated gene A (natural killer-enhancing factor A)
DLBCL	0.86	0.5053	0.43887	0.341477	U12595_at	Tumor necrosis factor type 1 receptor associated protein (TRAP1) mRNA, partial cds
DLBCL	0.86	0.502769	0.438058	0.339241	X17567_s_at	SNRPB Small nuclear ribonucleoprotein polypeptides B and B1
DLBCL	0.86	0.501931	0.436479	0.338219	HG4074-HT4344_at	Rad2
DLBCL	0.85	0.50185	0.435299	0.336796	L19686_rna1_at	Macrophage migration inhibitory factor (MIF) gene
DLBCL	0.85	0.501628	0.434148	0.335113	D25328_at	PFKP Phosphofructokinase, platelet
DLBCL	0.84	0.50034	0.432889	0.334007	J02783_at	P4HB Procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase;

						thyroid hormone binding protein p55)
DLBCL	0.84	0.500014	0.432308	0.332955	M25753_at	G2/MITOTIC-SPECIFIC CYCLIN B1
DLBCL	0.84	0.496897	0.431165	0.332026	U29680_at	Bcl-2 related (Bfl-1) mRNA
DLBCL	0.83	0.495355	0.430388	0.331187	Z50115_s_at	Thimet oligopeptidase (metalloproteinase)
DLBCL	0.83	0.493936	0.428397	0.329768	J04988_at	90-kDa heat-shock protein gene, cDNA
DLBCL	0.83	0.492029	0.426338	0.328631	D43950_at	T-COMPLEX PROTEIN 1, EPSILON SUBUNIT
DLBCL	0.82	0.491698	0.425498	0.328014	D45248_at	Proteasome activator hPA28 subunit beta
DLBCL	0.82	0.490525	0.423757	0.325838	D13633_at	KIAA0008 gene
DLBCL	0.82	0.490265	0.421918	0.324744	X74801_at	T-COMPLEX PROTEIN 1, GAMMA SUBUNIT
DLBCL	0.81	0.488898	0.419354	0.323613	L25876_at	Protein tyrosine phosphatase (CIP2)mRNA
DLBCL	0.8	0.488739	0.418746	0.322749	U40369_rna1_at	Spermidine/spermine N1-acetyltransferase (SSAT) gene
DLBCL	0.8	0.488599	0.416889	0.320644	X01060_at	TFRC Transferrin receptor (p90, CD71)
DLBCL	0.8	0.487413	0.415948	0.319924	U53347_at	Neutral amino acid transporter B mRNA
DLBCL	0.8	0.487102	0.415472	0.319036	X69433_at	IDH2 Isocitrate dehydrogenase 2 (NADP+), mitochondrial
DLBCL	0.79	0.486837	0.415192	0.318241	L06419_at	PLOD Lysyl hydroxylase
DLBCL	0.79	0.485817	0.41238	0.317421	M16591_s_at	HCK Hemopoietic cell kinase
DLBCL	0.78	0.483999	0.412347	0.316479	U81375_at	Placental equilibrative nucleoside transporter 1 (hENT1) mRNA
DLBCL	0.78	0.482738	0.412115	0.316097	D29958_at	KIAA0116 gene, partial cds
Follicular	0.8	0.625413	0.564843	0.442353	Z21966_at	POU6F1 POU homeobox protein
Follicular	0.75	0.560816	0.526607	0.407332	X16983_at	ITGA4 Integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)
Follicular	0.69	0.548743	0.500653	0.393084	Z11793_at	Selenoprotein P
Follicular	0.69	0.532611	0.484738	0.381398	AB002409_at	SLC
Follicular	0.66	0.52049	0.474044	0.370661	D87119_at	Cancellous bone osteoblast mRNA for GS3955
Follicular	0.63	0.512199	0.469188	0.365806	L42324_at	(clone GPCR W) G protein-linked receptor gene (GPCR) gene, 5' end of cds
Follicular	0.61	0.504205	0.461574	0.360505	U46006_s_at	Smooth muscle LIM protein (h-SmLIM) mRNA
Follicular	0.61	0.502025	0.458774	0.354998	L19314_at	HRY gene
Follicular	0.6	0.500113	0.454945	0.350754	HG3928-HT4198_at	Surfactant Protein Sp-A1 Delta
Follicular	0.59	0.494389	0.449126	0.346475	X91911_s_at	Glioma pathogenesis-related protein (GliPR) mRNA
Follicular	0.59	0.492023	0.444349	0.343027	L42621_at	Ly-9 mRNA
Follicular	0.58	0.482037	0.44216	0.339331	S73591_at	Brain-expressed HHCPA78 homolog [human, HL-60 acute promyelocytic leukemia cells, mRNA, 2704 nt]
Follicular	0.57	0.480941	0.438895	0.336835	X86098_at	BS69 protein
Follicular	0.57	0.477009	0.433887	0.333261	U64863_at	HPD-1 (hPD-1) mRNA
Follicular	0.57	0.472307	0.432979	0.331301	M63379_at	CLU Clusterin (complement lysis inhibitor; testosterone-repressed prostate message 2; apolipoprotein J)
Follicular	0.56	0.471385	0.429803	0.32881	U16307_at	Glioma pathogenesis-related protein (GliPR) mRNA
Follicular	0.56	0.464764	0.426799	0.325733	D78134_at	YWHAZ Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide
Follicular	0.55	0.463519	0.424091	0.323913	Z35227_at	TTF mRNA for small G protein
Follicular	0.55	0.458776	0.420806	0.320333	Z30426_at	CD69 CD69 antigen (early T cell activation antigen)

Follicular	0.55	0.454894	0.419801	0.319149	Z33642_at	V7 mRNA for leukocyte surface protein
Follicular	0.55	0.448948	0.418436	0.318177	M59829_at	MHC class III HSP70-HOM gene (HLA)
Follicular	0.54	0.448891	0.416759	0.316995	Z49269_at	Chemokine HCC-1
Follicular	0.54	0.447631	0.415283	0.314132	AC002073_cds1_at	WUGSC:DJ515N1.2 gene extracted from Human PAC clone DJ515N1 from 22q11.2-q22
Follicular	0.54	0.447108	0.414237	0.312085	U19345_at	AR1 protein (AR) mRNA
Follicular	0.54	0.446023	0.413136	0.310901	Z33905_at	43kD acetylcholine receptor-associated protein (Rapsyn)
Follicular	0.54	0.437523	0.412024	0.310226	U69108_at	TNF receptor associated factor 5 mRNA, partial cds
Follicular	0.53	0.436616	0.409926	0.308044	M99701_at	(pp21) mRNA
Follicular	0.53	0.435312	0.406132	0.30729	M94880_f_at	HLA-A MHC class I protein HLA-A (HLA-A28,-B40, -Cw3)
Follicular	0.53	0.434787	0.40372	0.306417	X85785_rna1_at	DARC gene
Follicular	0.53	0.434068	0.403106	0.305256	S57212_s_at	MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)
Follicular	0.53	0.432514	0.401956	0.303967	L15309_at	ZNF141 Zinc finger protein 141 (clone pHZ-44)
Follicular	0.52	0.431058	0.399777	0.302146	HG3635-HT3845_f_at	Zinc Finger Protein, Kruppel-Like
Follicular	0.52	0.430707	0.39714	0.300316	L10615_s_at	CSN2 Beta-casein
Follicular	0.52	0.427892	0.396804	0.29958	HG3254-HT3431_at	Phosphatidylinositol 3-Kinase P110, Beta Isoform
Follicular	0.51	0.427627	0.396715	0.298372	X00437_s_at	TCRB T-cell receptor, beta cluster
Follicular	0.51	0.425774	0.395783	0.297631	X77922_s_at	SIAT8 Sialyltransferase 8 (alpha-N-acetylneuraminate: alpha-2,8-sialytransferase, GD3 synthase)
Follicular	0.51	0.424852	0.393411	0.296946	AF008937_at	Syntaxin-16C mRNA
Follicular	0.51	0.424642	0.390192	0.294195	U96113_at	Nedd-4-like ubiquitin-protein ligase WWP1 mRNA, partial cds
Follicular	0.5	0.423499	0.388411	0.293715	Z50781_at	Leucine zipper protein
Follicular	0.5	0.423051	0.387732	0.292031	X03934_at	T-cell antigen receptor gene T3-delta
Follicular	0.5	0.422899	0.387475	0.291387	U56814_at	DNase1-Like III protein (DNAS1L3) mRNA
Follicular	0.5	0.422814	0.387361	0.290189	L27071_at	TXK TXK tyrosine kinase
Follicular	0.5	0.422569	0.38503	0.289444	M23323_s_at	T-CELL SURFACE GLYCOPROTEIN CD3 EPSILON CHAIN PRECURSOR
Follicular	0.5	0.421817	0.383622	0.288181	D31797_at	CD40LG CD40 antigen ligand (hyper IgM syndrome)
Follicular	0.49	0.420856	0.383083	0.287359	Z26634_at	ANK2 Ankyrin 2 (neuronal)
Follicular	0.49	0.420585	0.382664	0.287277	U72935_cds3_s_at	ATRX gene (putative DNA dependent ATPase and helicase) extracted from Human putative DNA dependent ATPase and helicase (ATRX) gene
Follicular	0.49	0.419956	0.379825	0.286827	Y09392_s_at	WSL-LR, WSL-S1 and WSL-S2 proteins
Follicular	0.49	0.41545	0.379744	0.286322	M57703_s_at	PMCH Pro-melanin-concentrating hormone
Follicular	0.49	0.415292	0.379152	0.28568	L08488_at	INPP1 Inositol polyphosphate-1-phosphatase
Follicular	0.49	0.415218	0.378907	0.285068	D83597_at	RP105

DLBCL versus FL Prediction

The following table shows the prediction results when using weighted voting to predict the DLBCL versus FL distinction. The distinction was predicted using GeneCluster using weighted voting and the standard preprocessing (expression values were thresholded at a minimum of 20 and a maximum of 16,000 and genes were filtered out if either maximum/minimum<3 (3-fold variation) or maximum-minimum<100). The table shows the cross-validation testing results from the weighted voting predictor that used a mean based signal-to-noise feature selection of 30 genes.

Sample	Predicted Class	Observed Class	Error?
DLBC1	0	0	
DLBC2	0	0	
DLBC3	0	0	
DLBC4	0	0	
DLBC5	0	0	
DLBC6	0	0	
DLBC7	0	0	
DLBC8	0	0	
DLBC9	0	0	
DLBC10	0	0	
DLBC11	0	0	
DLBC12	0	0	
DLBC13	0	0	
DLBC14	0	0	
DLBC15	1	0	*
DLBC16	0	0	
DLBC17	0	0	
DLBC18	0	0	
DLBC19	0	0	
DLBC20	0	0	
DLBC21	1	0	*
DLBC22	0	0	
DLBC23	1	0	*
DLBC24	0	0	
DLBC25	0	0	
DLBC26	1	0	*
DLBC27	0	0	
DLBC28	0	0	
DLBC29	1	0	*
DLBC30	0	0	
DLBC31	0	0	
DLBC32	0	0	
DLBC33	0	0	
DLBC34	0	0	
DLBC35	0	0	
DLBC36	0	0	
DLBC37	0	0	

DLBC38	0	0	
DLBC39	0	0	
DLBC40	0	0	
DLBC41	0	0	
DLBC42	0	0	
DLBC43	0	0	
DLBC44	0	0	
DLBC45	0	0	
DLBC46	0	0	
DLBC47	0	0	
DLBC48	0	0	
DLBC49	0	0	
DLBC50	0	0	
DLBC51	0	0	
DLBC52	0	0	
DLBC53	0	0	
DLBC54	0	0	
DLBC55	0	0	
DLBC56	1	0	*
DLBC57	0	0	
DLBC58	0	0	
FSCC1	1	1	
FSCC2	1	1	
FSCC3	1	1	
FSCC4	1	1	
FSCC5	1	1	
FSCC6	1	1	
FSCC7	1	1	
FSCC8	1	1	
FSCC9	1	1	
FSCC10	1	1	
FSCC11	1	1	
FSCC12	1	1	
FSCC13	1	1	
FSCC14	1	1	
FSCC15	1	1	
FSCC16	1	1	
FSCC17	1	1	
FSCC18	1	1	
FSCC19	1	1	

The following table shows the confusion matrix from predicting the DLBCL versus follicular distinction using the weighted voting model.

		Weighted Voting		
		DLBCL	Follicular	
True	DLBCL	52	6	58
	Follicular	0	19	19
		52	25	77

The model predicts 71 out of 77 samples correctly and it is clearly highly significant (P-val < 1.4 x 10^{-9}, see the calculation below and the ***Proportional Chance Criterion***)

C_{pro} = (52/77)*(58/77) + (25/77)*(19/77) = 0.589

P_{cc} = (52+19)/77 = 0.922

Z = (0.922-0.589)/sqrt(0.922*(1-0.922)/77)

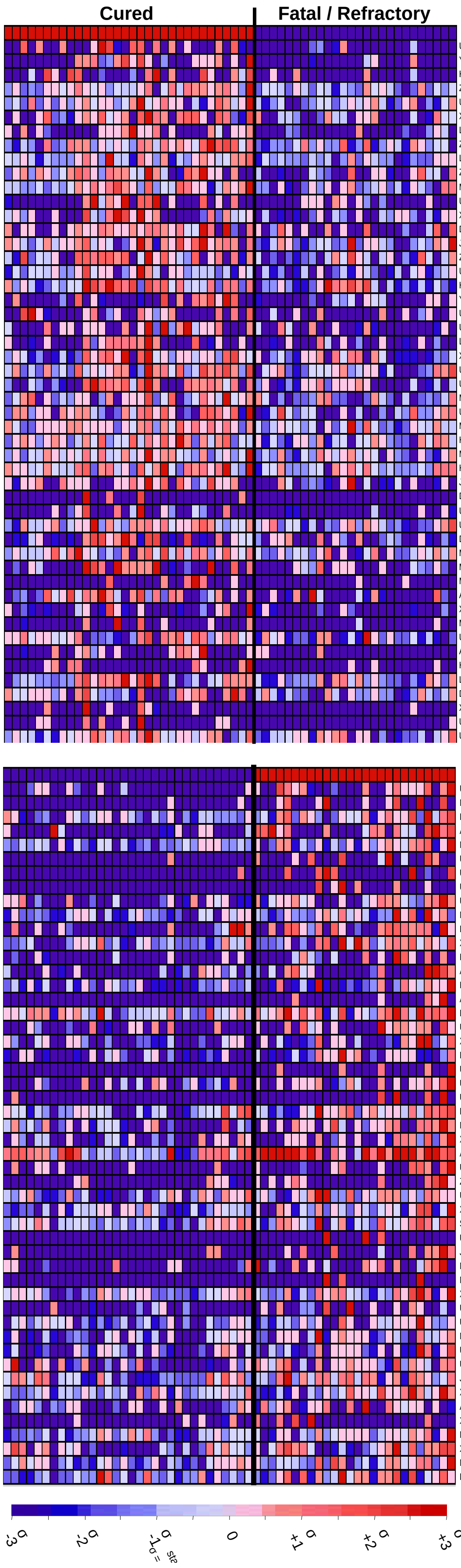
Pval = 1.4 x 10^{-9}

DLBCL Cured versus Fatal/Refractory Distinction

Within this section, we expand on the Diffuse Large B-Cell Lymphoma (DLBCL) outcome (cured versus fatal / refractory disease) analysis of the paper. First, we begin with the pink-o-gram showing the expression profiles of the top 50 genes for DLBCL and FL and the permutation tests associated with those genes. In the next subsection, we show the results from predicting the DLBCL outcome using several different prediction methods.

Expression Profiles of Cured and Fatal/Refractory Disease

This section expands on Figure 2 from the paper. This picture shows the top 50 markers per class for the DLBCL cured versus fatal / refractory distinction as sorted by their signal-to-noise ratios (using mean) as described in ***Gene Marker Selection*** section. The genes that were expressed at higher levels in cured disease are shown on top while the genes that were more highly expressed in fatal disease are shown on the bottom. Red indicates a high relative expression while blue represents a low relative expression. Each column is a sample and each row is a gene (with the first rows of the cured and fatal / refractory sections showing an idealized expression profile). Expression profiles for the 32 cured DLBCL samples are on the left while the profiles for the 26 fatal / refractory samples are on the right. The table below shows the top 50 markers for each tumor class including the permutation test values (see ***Permutation Test and Neighborhood Analysis for Marker Genes***). Standard preprocessing was used for the data where expression values were thresholded to 20 from below and 16000 from above and a variation filter removed non-changing genes (genes were filtered out if either maximum/minimum<3 (3-fold variation) or maximum-minimum<100 absolute units).



Acc. No.	Description
U43519	DRP2 Dystrophin related protein 2
Y09836	3'UTR of unknown protein
HG2314-HT2410	Uncharacterized
Z15114	PRKCG Protein kinase C, gamma
U12767	Mitogen induced nuclear orphan receptor (MINOR)
X77307	5-HYDROXYTRYPTAMINE 2B RECEPTOR
L40377	Cytoplasmic antiproteinase 2 (CAP2)
Z30644	Chloride channel (putative) 2163bp
L07765	CES2 Carboxylestease 2 (liver)
Z83802	Axonemal dynein heavy chain (partial, ID hdhc3)
M24736	SELE Selectin E (endothelial adhesion molecule 1)
U04898	RORA RAR-related orphan receptor A
X70683	SOX4 SRY (sex determining region Y)-box 4
D87468	KIAA0278 gene, partial cds
J03242	IGF2 Insulin-like growth factor 2 (somatomedin A)
Z95624	DNA sequence from cosmid U237H1 contains Ras like GTPase and ESTs
U96781_cds1	ATP2A1 gene
HG3227-HT3404	Guanine Nucleotide-Binding Protein Hsr1
Y10262	EYA3 gene
U79271	Clones 23920 and 23921 mRNA sequence
U82535	Fatty acid amide hydrolase mRNA
L04751	CYP4A11 Cytochrome P450, subfamily IVA, polypeptide 11
X17254	GATA1 Transcription factor Eryf1
U48437	Amyloid precursor-like protein
U49082	Transporter protein (g17)
M37763	Neurotrophin-3 (NT-3)
U18548	GPR12 G protein coupled-receptor
M17863	IGF2 Insulin-like growth factor 2 (somatomedin A)
HG2415-HT2511	Transcription Factor E2f-2
M35531	FUT1 Fucosyltransferase 1 (galactoside 2-alpha-L-fucosyltransferase, Bombay phenotype included)
HG4660-HT5073	Microtubule-Associated Protein 1b
J05036	CTSE Cathepsin E
D50402	NRAMP1 Natural resistance-associated macrophage protein 1
U05861	DDH1 Dihydrodiol dehydrogenase
U50743	Na,K-ATPase gamma subunit
D87937	Alpha(1,2)fucosyltransferase, 5'UTR partial sequence
M14123_xpt2	Gag 1 protein from Human endogenous retrovirus HERV-K10./ntype=DNA /annot=CDS
M63904	GNA15 Guanine nucleotide binding protein (G protein), alpha 15 (Gq class)
M86383	Nicotinic acetylcholine receptor alpha 3 subunit precursor
AFFX-HUMRGE/M10098_3	AFFX-HUMRGE/M10098_3 (endogenous control)
X54925	MMP1 Matrix metalloproteinase 1
M16441_cds1	Lymphotoxin gene extracted from Human tumor necrosis factor and lymphotoxin genes
U00930	Clone CE29 8.1 (CAC)n/(GTG)n repeat-containing mRNA
AFFX-LysX-5	Lysozyme precursor
HG3264-HT3441	Af-6 (Gb:U02478)
L11672	ZNF91 Zinc finger protein 91 (HPF7, HTF10)
D25217	KIAA0027 gene
X13955	MYL4 Myosin, light polypeptide 4, alkali; atrial, embryonic
U29195	NPTX2 Neuronal pentraxin II
U40279	Beta-2 integrin alphaD subunit (ITGAD) gene, exons 25-30, and partial cds

U83908	Nuclear antigen H731 mRNA
M99435	TRANSDUCIN-LIKE ENHANCER PROTEIN 1
L20971	PDE4B Phosphodiesterase 4B, cAMP-specific
AC002450	Uncharacterized
M18255_cds2	PRKACB gene (protein kinase C-beta-1)
U09550	Oviductal glycoprotein
U38864	Zinc-finger protein C2H2-150
U02609	Transducin-like protein
U65093	Msg1-related gene 1 (mrg1)
M55531	SLC2A5 Solute carrier family 2 (facilitated glucose transporter), member 5
D16480	HADHA alpha subunit
X89750	TGIF protein
HG4638-HT5050	Spliceosomal Protein Sap 49
AFFX-HUMISGF3A/M97935_MA	Interferon-stimulated gene factor 3 (ISGF3)
HG4322-HT4592	Tubulin, Beta
AFFX-HUMISGF3A/M97935_5	Interferon-stimulated gene factor 3 (ISGF3)
D87735	CAG-isl 7 {trinucleotide repeat-containing sequence} [human, pancreas, mRNA Partial, 701 nt]
U97502_rna1	Butyrophilin (BT3.3) gene
X58399	L2-9 transcript of unrearranged immunoglobulin V(H)5 pseudogene
D86969	KIAA0215
U50360	Calcium, calmodulin-dependent protein kinase II gamma mRNA, partial cds
U60415	BHLH-PAS protein JAP3
U60276	HASNA-I
D86959	KIAA0204 gene
M12625	Lecithin-cholesterol acyltransferase mRNA, with 5' and 3' flanking DNA sequences
X17644	GSPT1 G1 to S phase transition 1
AFFX-HUMGAPDH/M33197_M	AFFX-HUMGAPDH/M33197_M (endogenous control)
U52111_rna3	ALD gene (adrenoleukodystrophy protein) extracted from Human Xq28 genomic
Z12173	GNS Glucosamine (N-acetyl)-6-sulfatase (Sanfilippo disease IIID)
U48730	Transcription factor Stat5b (stat5b)
X06318	PRKCB1 Protein kinase C, beta 1
S74221	IK
U20230	Guanyl cyclase C
J05448	POLR2C RNA polymerase II, polypeptide C (33kD)
M27281	VEGF Vascular endothelial growth factor
M55683	CRTM Cartilage matrix protein
X63468	GTF2E1 General transcription factor TFIIE alpha subunit, 56 kD
U19180	BAGE B melanoma antigen
U26727	CDKN2A Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
M36067	LIG1 Ligase I, DNA, ATP-dependent
U71598	Zinc finger protein zfp2 (zf2)
U73167_cds5	SM15 gene (human interferon-related protein SM15 (U09585) extracted from Human cosmid LUCA14
J03798	SMALL NUCLEAR RIBONUCLEOPROTEIN SM D1
X91249	WHITE PROTEIN HOMOLOG
AB002314	KIAA0316
X93017	Ncx2 gene (exon 2)
M55905	ME2 Malic enzyme 2, mitochondrial
X07109	(clones lambda-hPKC-beta[15,802]) protein kinase C-beta-2 (PRKCB2) mRNA
M92439	130 KD LEUCINE-RICH PROTEIN
D89077	Src-like adapter protein mRNA

Distinction	Score	Perm 1%	Perm 5%	Perm 50%	Feature	Description
cured	0.51	0.701828	0.606881	0.489807	U43519_at	DRP2 Dystrophin related protein 2
cured	0.48	0.63085	0.556084	0.451116	Y09836_at	3'UTR of unknown protein
cured	0.45	0.585544	0.532573	0.432913	HG2314-HT2410_at	Uncharacterized
cured	0.44	0.557965	0.514223	0.419056	Z15114_at	PRKCG Protein kinase C, gamma
cured	0.42	0.54197	0.499829	0.408782	U12767_at	Mitogen induced nuclear orphan receptor (MINOR)
cured	0.4	0.533506	0.492673	0.399329	X77307_at	5-Hydroxytryptamine 2B Receptor
cured	0.39	0.525141	0.488428	0.391837	L40377_at	Cytoplasmic antiproteinase 2 (CAP2)
cured	0.38	0.523665	0.485158	0.38647	Z30644_at	Chloride channel (putative) 2163bp
cured	0.38	0.513542	0.477072	0.381914	L07765_at	CES2 Carboxylestease 2 (liver)
cured	0.37	0.512298	0.471612	0.375938	Z83802_at	Axonemal dynein heavy chain (partial, ID hdhc3)
cured	0.37	0.505015	0.468794	0.372087	M24736_s_at	SELE Selectin E (endothelial adhesion molecule 1)
cured	0.36	0.503627	0.462368	0.368665	U04898_at	RORA RAR-related orphan receptor A
cured	0.36	0.502394	0.459726	0.36443	X70683_at	SOX4 SRY (sex determining region Y)-box 4
cured	0.35	0.496602	0.457745	0.360582	D87468_at	KIAA0278 gene, partial cds
cured	0.35	0.494674	0.455238	0.357121	J03242_s_at	IGF2 Insulin-like growth factor 2 (somatomedin A)
cured	0.34	0.484418	0.447009	0.348859	Z95624_at	DNA sequence from cosmid U237H1 contains Ras like GTPase and ESTs
cured	0.34	0.491168	0.449814	0.351119	U96781_cds1_at	ATP2A1 gene
cured	0.34	0.492444	0.452359	0.353374	HG3227-HT3404_at	Guanine Nucleotide-Binding Protein Hsr1
cured	0.34	0.480884	0.444184	0.346359	Y10262_s_at	EYA3 gene
cured	0.34	0.480849	0.443607	0.344516	U79271_at	Clones 23920 and 23921 mRNA sequence
cured	0.34	0.478695	0.442053	0.341813	U82535_at	Fatty acid amide hydrolase mRNA
cured	0.33	0.476235	0.441441	0.339551	L04751_at	CYP4A11 Cytochrome P450, subfamily IVA, polypeptide 11
cured	0.33	0.473773	0.440106	0.337639	X17254_at	GATA1 Transcription factor Eryf1
cured	0.33	0.47319	0.436045	0.334579	U48437_at	Amyloid precursor-like protein 1 mRNA
cured	0.33	0.469373	0.435	0.33388	U49082_at	Transporter protein (g17)
cured	0.33	0.468894	0.43085	0.330915	M37763_at	Neurotrophin-3 (NT-3) gene
cured	0.33	0.468919	0.432797	0.332559	U18548_at	GPR12 G protein coupled-receptor gene
cured	0.32	0.466871	0.430702	0.329294	M17863_s_at	IGF2 Insulin-like growth factor 2 (somatomedin A)
cured	0.32	0.465294	0.428315	0.328131	HG2415-HT2511_at	Transcription Factor E2F-2
cured	0.32	0.464442	0.427678	0.327347	M35531_at	FUT1 Fucosyltransferase 1 (galactoside 2-alpha-L-fucosyltransferase, Bombay phenotype included)
cured	0.32	0.461678	0.426236	0.325049	HG4660-HT5073_at	Microtubule-Associated Protein 1b
cured	0.32	0.459537	0.423169	0.324146	J05036_s_at	CTSE Cathepsin E
cured	0.32	0.457099	0.41793	0.32094	D50402_at	NRAMP1 Natural resistance-associated macrophage protein 1 (might include Leishmaniasis)
cured	0.32	0.459077	0.422821	0.322058	U05861_at	DDH1 Dihydrodiol dehydrogenase
cured	0.31	0.456314	0.417538	0.319928	U50743_at	Na,K-ATPase gamma subunit
cured	0.31	0.454757	0.416749	0.318132	D87937_at	Alpha(1,2)fucosyltransferase, 5'UTR partial sequence
cured	0.31	0.454722	0.41496	0.317387	M14123_xpt2_at	Gag 1 protein from Human endogenous retrovirus HERV-K10
cured	0.31	0.451535	0.41376	0.316178	M63904_at	GNA15 Guanine nucleotide binding protein (G protein), alpha 15 (Gq class)

cured	0.31	0.449922	0.41228	0.314932	M86383_s_at	Nicotinic acetylcholine receptor alpha 3 subunit precursor
cured	0.31	0.448932	0.411437	0.313067	AFFX-HUMRGE/M10098_3_at	AFFX-HUMRGE/M10098_3_at (endogenous control)
cured	0.31	0.448903	0.410967	0.312634	X54925_at	MMP1 Matrix metalloproteinase 1 (interstitial collagenase)
cured	0.31	0.445025	0.410691	0.311763	M16441_cds1_at	Lymphotoxin gene extracted from Human tumor necrosis factor and lymphotoxin genes
cured	0.31	0.444606	0.408006	0.310486	U00930_at	Clone CE29 8.1 (CAC)n/(GTG)n repeat-containing mRNA
cured	0.31	0.443729	0.407523	0.308979	AFFX-LysX-5_at	Lysozyme precursor
cured	0.3	0.44295	0.406483	0.307633	HG3264-HT3441_at	Af-6 (Gb:U02478)
cured	0.3	0.442854	0.40632	0.306814	L11672_at	ZNF91 Zinc finger protein 91 (HPF7, HTF10)
cured	0.3	0.442814	0.404929	0.305609	D25217_at	KIAA0027 gene
cured	0.3	0.442637	0.40453	0.305216	X13955_s_at	MYL4 Myosin, light polypeptide 4, alkali; atrial, embryonic
cured	0.29	0.442214	0.403909	0.3042	U29195_at	NPTX2 Neuronal pentraxin II
cured	0.29	0.442122	0.402953	0.302759	U40279_at	Beta-2 integrin alphaD subunit (ITGAD) gene, exons 25-30, and partial cds
fatal / ref.	0.52	0.629499	0.590196	0.473969	U83908_at	Nuclear antigen H731 mRNA
fatal / ref.	0.5	0.598192	0.550471	0.440446	M99435_at	Transducin-Like Enhancer Protein 1
fatal / ref.	0.49	0.586605	0.52294	0.422172	L20971_at	PDE4B Phosphodiesterase 4B, cAMP-specific
fatal / ref.	0.47	0.571745	0.510494	0.410309	AC002450_at	Uncharacterized
fatal / ref.	0.45	0.551235	0.498252	0.400033	M18255_cds2_s_at	PRKACB gene (protein kinase C-beta-1)
fatal / ref.	0.44	0.546654	0.48438	0.393251	U09550_at	Oviductal glycoprotein
fatal / ref.	0.4	0.542818	0.478461	0.3866	U38864_at	Zinc-finger protein C2H2-150
fatal / ref.	0.4	0.541602	0.47429	0.378177	U02609_at	Transducin-like protein
fatal / ref.	0.39	0.53989	0.471154	0.374855	U65093_at	Msg1-related gene 1 (mrg1)
fatal / ref.	0.39	0.538255	0.466382	0.369673	M55531_at	SLC2A5 Solute carrier family 2 (facilitated glucose transporter), member 5
fatal / ref.	0.37	0.537615	0.463894	0.36446	D16480_at	HADHA alpha subunit
fatal / ref.	0.37	0.533264	0.460969	0.361161	X89750_at	TGIF protein
fatal / ref.	0.37	0.528345	0.457629	0.359085	HG4638-HT5050_at	Spliceosomal Protein Sap 49
fatal / ref.	0.36	0.528265	0.456796	0.356477	AFFX-HUMISGF3A/M97935_MA_at	Interferon-simulated gene factor 3 (ISGF3)
fatal / ref.	0.35	0.528039	0.453817	0.354938	HG4322-HT4592_at	Beta Tubulin
fatal / ref.	0.35	0.527696	0.44931	0.35136	AFFX-HUMISGF3A/M97935_5_at	Interferon-simulated gene factor 3 (ISGF3)
fatal / ref.	0.35	0.522245	0.445128	0.347713	D87735_at	CAG-isl 7 {trinucleotide repeat-containing sequence} [human, pancreas, mRNA Partial, 701 nt]
fatal / ref.	0.35	0.519589	0.443869	0.345508	U97502_rna1_at	Butyrophilin (BT3.3) gene
fatal / ref.	0.35	0.509272	0.438467	0.341222	X58399_at	L2-9 transcript of unrearranged immunoglobulin V(H)5 pseudogene
fatal / ref.	0.34	0.508104	0.436378	0.339558	D86969_at	KIAA0215 gene
fatal / ref.	0.34	0.505898	0.433669	0.337453	U50360_s_at	Calcium, calmodulin-dependent protein kinase II gamma mRNA, partial cds
fatal / ref.	0.34	0.505311	0.432566	0.335795	U60415_at	BHLH-PAS protein JAP3
fatal / ref.	0.34	0.504822	0.429215	0.334266	U60276_at	HASNA-I
fatal / ref.	0.33	0.502244	0.425958	0.329912	D86959_at	KIAA0204 gene
fatal / ref.	0.33	0.503425	0.427205	0.332262	M12625_at	Lecithin-cholesterol acyltransferase mRNA, with 5' and 3' flanking DNA sequences

fatal / ref.	0.33	0.502184	0.423554	0.328482	X17644_s_at	GSPT1 G1 to S phase transition 1
fatal / ref.	0.33	0.497344	0.422055	0.326836	AFFX-HUMGAPDH/M33197_M_at	AFFX-HUMGAPDH/M33197_M_at (endogenous control)
fatal / ref.	0.33	0.496499	0.420607	0.325051	U52111_rna3_at	ALD gene (adrenoleukodystrophy protein) extracted from Human Xq28 genomic DNA
fatal / ref.	0.33	0.49615	0.418816	0.322953	Z12173_at	GNS Glucosamine (N-acetyl)-6-sulfatase (Sanfilippo disease IIID)
fatal / ref.	0.33	0.493655	0.415445	0.320633	U48730_at	Transcription factor Stat5b (stat5b)
fatal / ref.	0.33	0.495202	0.417776	0.321183	X06318_at	PRKCB1 Protein kinase C, beta 1
fatal / ref.	0.33	0.490236	0.414764	0.319562	S74221_at	IK
fatal / ref.	0.32	0.490049	0.414141	0.317909	U20230_at	Guanyl cyclase C
fatal / ref.	0.32	0.488861	0.411753	0.316488	J05448_at	POLR2C RNA polymerase II, polypeptide C (33kD)
fatal / ref.	0.32	0.486796	0.411539	0.314923	M27281_at	VEGF Vascular endothelial growth factor
fatal / ref.	0.32	0.486495	0.410496	0.31434	M55683_at	CRTM Cartilage matrix protein
fatal / ref.	0.32	0.483128	0.410335	0.312946	X63468_at	GTF2E1 General transcription factor TFIIE alpha subunit, 56 kD
fatal / ref.	0.32	0.481082	0.408804	0.311451	U19180_at	BAGE B melanoma antigen
fatal / ref.	0.32	0.480847	0.408178	0.309727	U26727_at	CDKN2A Cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
fatal / ref.	0.32	0.479592	0.406983	0.309455	M36067_at	LIG1 Ligase I, DNA, ATP-dependent
fatal / ref.	0.32	0.479563	0.40648	0.308443	U71598_at	Zinc finger protein zfp2 (zf2)
fatal / ref.	0.32	0.478498	0.404581	0.307401	U73167_cds5_at	SM15 gene (human interferon-related protein SM15 (U09585)) extracted from Human cosmid LUCA14
fatal / ref.	0.32	0.473722	0.40426	0.306035	J03798_at	Small Nuclear Ribonucleoprotein SM D1
fatal / ref.	0.31	0.473424	0.403465	0.30549	X91249_at	White Protein Homolog
fatal / ref.	0.31	0.470856	0.402851	0.303643	AB002314_at	KIAA0316
fatal / ref.	0.31	0.470518	0.401797	0.302593	X93017_at	Ncx2 gene (exon 2)
fatal / ref.	0.31	0.468978	0.401125	0.301642	M55905_at	ME2 Malic enzyme 2, mitochondrial
fatal / ref.	0.31	0.468884	0.400301	0.300672	X07109_at	(clones lambda-hPKC-beta[15,802]) protein kinase C-beta-2 (PRKCB2)
fatal / ref.	0.31	0.467592	0.398794	0.29949	M92439_at	130 KD Leucine-rich Protein
fatal / ref.	0.31	0.46715	0.39795	0.2989	D89077_at	Src-like adapter protein mRNA

Evaluated individually, no genes exceed the 1% and 5% significance levels with respect to outcome but a number pass at the 50% levels. Despite this, we show in the next section how combinations of these markers can be used to build models that can accurately predict lymphoma outcome.

DLBCL Outcome Prediction

This section expands on the outcome prediction results from the paper. The main outcome results in the paper used a thirteen-gene weighted-voting (WV) predictor so we present those results first. Results from using other types of predictors (k-nearest neighbors (KNN) and support vector machines (SVM)) are presented later in this section.

The following table presents results from using weighted voting and cross-validation to predict lymphoma treatment outcome (cured versus fatal / refractory). Outcome was predicted using GeneCluster using this project’s standard preprocessing (expression values were thresholded at a minimum of 20 and a maximum of 16,000 and genes with maximum/minimum<3 (3-fold variation) or maximum-minimum<100 were filtered). The table shows the cross-validation testing results on the 58 DLBCL outcome samples from the weighted-voting predictor that used a mean based signal-to-noise feature selection of 13 genes.

Sample	Predicted Class	Observed Class	Error?	IPI Number	Survival (months)	Truncated Survival
DLBC1	1	0	*	1	72.9	60
DLBC2	0	0		1	143.1	60
DLBC3	0	0		2	144.2	60
DLBC4	0	0		3	61	60
DLBC5	0	0		1	86.5	60
DLBC6	0	0		1	84.2	60
DLBC7	0	0		3	112.5	60
DLBC8	0	0		1	133.2	60
DLBC9	0	0		1	22.1	22.1
DLBC10	0	0		2	182.4	60
DLBC11	0	0		1	66.4	60
DLBC12	0	0		.	146.8	60
DLBC13	1	0	*	2	62.9	60
DLBC14	0	0		2	50.9	50.9
DLBC15	0	0		1	78.5	60
DLBC16	0	0		.	48.6	48.6
DLBC17	0	0		3	55.9	55.9
DLBC18	0	0		1	12.6	12.6
DLBC19	0	0		2	50.2	50.2
DLBC20	0	0		3	58	58
DLBC21	0	0		2	66.4	60
DLBC22	0	0		1	65.7	60
DLBC23	0	0		1	50.2	50.2
DLBC24	0	0		1	26.9	26.9
DLBC25	0	0		1	34.4	34.4
DLBC26	0	0		1	26	26
DLBC27	0	0		1	30	30
DLBC28	0	0		2	31.7	31.7
DLBC29	0	0		1	32.2	32.2
DLBC30	1	0	*	1	19.2	19.2
DLBC31	0	0		1	33.1	33.1
DLBC32	0	0		1	21.4	21.4
DLBC33	1	1		1	15.7	15.7
DLBC34	1	1		3	11.6	11.6
DLBC35	1	1		3	3.4	3.4
DLBC36	1	1		1	36.6	36.6
DLBC37	0	1	*	3	5	5
DLBC38	0	1	*	1	9.5	9.5
DLBC39	0	1	*	4	3.2	3.2
DLBC40	1	1		2	4.9	4.9
DLBC41	1	1		3	12	12
DLBC42	1	1		3	4.9	4.9
DLBC43	1	1		3	60.4	60
DLBC44	0	1	*	2	16.3	16.3
DLBC45	0	1	*	3	16.4	16.4
DLBC46	1	1		3	9.5	9.5

DLBC47	1	1		3	15.6	15.6
DLBC48	1	1		3	17.8	17.8
DLBC49	1	1		2	56.9	56.9
DLBC50	0	1	*	1	13.3	13.3
DLBC51	1	1		2	12.3	12.3
DLBC52	0	1	*	1	44.6	44.6
DLBC53	1	1		3	4.6	4.6
DLBC54	0	1	*	4	7.5	7.5
DLBC55	1	1		3	19.3	19.3
DLBC56	0	1	*	1	30.1	30.1
DLBC57	0	1	*	1	33.6	33.6
DLBC58	0	1	*	3	13.9	13.9

Within the table, class 0 represents the cured cases while class 1 represents the fatal/refractory cases. The following confusion matrix summarizes the leave-one-out cross-validation prediction results.

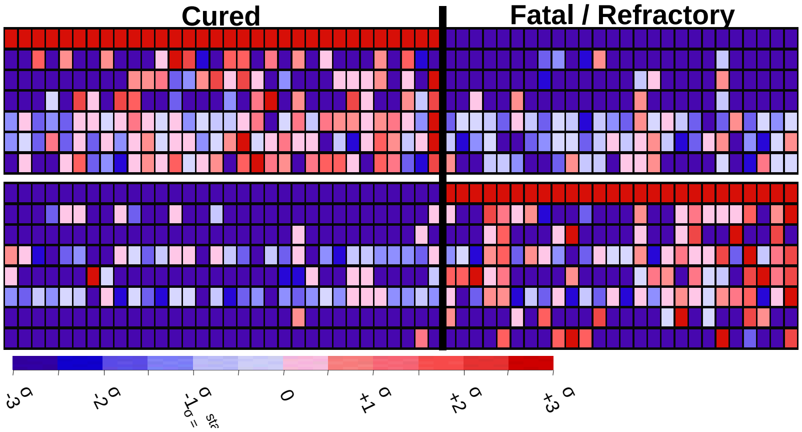
		Predicted Class	
		Cured	Fatal/Refractory
Observed Class	Cured	29	3
	Fatal/Refractory	11	15

Leave one out cross-validation builds a predictor and picks a set of features to use during each of the leave-one-out tests. The following table lists the features and the number of times they get used in the 58 leave-one-out predictors. Seven of the genes were common to all 58 cross-validation models; 4 additional genes were included in 54 or more models, 3 additional genes were included in 20-34 models, and 5 additional genes were used in 3-8 of the models.

Affymetrix Identifier	Number of Cross Validation Models Using Gene	Unigene ID	Description
U43519_at	58	Hs.159291	DRP2 Dystrophin related protein 2
M18255_cds2_s_at	58	Hs.77202	PRKACB gene (protein kinase C-beta-1)
U83908_at	58	Hs.100407	Nuclear antigen H731
Y09836_at	58	Hs.82503	3'UTR of unknown protein
M99435_at	58	Hs.28935	Transducin-like Enhancer Protein 1
AC002450_at	58		Uncharacterized
L20971_at	58	Hs.188	PDE4B Phosphodiesterase 4B, cAMP-specific
HG2314-HT2410_at	57		Uncharacterized
Z15114_at	57	Hs.2890	PRKCG Protein kinase C, gamma
U09550_at	55	Hs.1154	Oviductal glycoprotein
U12767_at	54	Hs.80561	Mitogen induced nuclear orphan receptor (MINOR)
U38864_at	34	Hs.108139	Zinc-finger protein C2H2-150
X77307_at	23	Hs.2507	5-HYDROXYTRYPTAMINE 2B RECEPTOR
L40377_at	20	Hs.41726	Cytoplasmic antiproteinase 2 (CAP2)
U65093_at	8	Hs.82071	Msg1-related gene 1 (mrg1)
Z30644_at	7	Hs.123059	Chloride channel (putative) 2163bp
U02609_at	5	Hs.114416	Transducin-like protein
Z83802_at	4		Axonemal dynein heavy chain (partial, ID hdhc3)

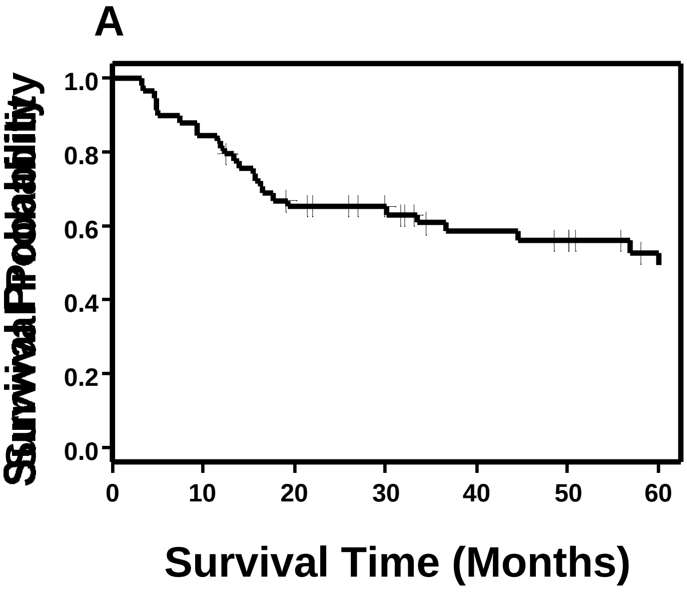
M55531_at	3	Hs.33084	SLC2A5 Solute carrier family 2 (facilitated glucose transporter), member 5

The top thirteen genes in the above table are highlighted in bold and are the genes that we consider to make up our “thirteen gene” model even though other genes are occasionally used in the cross-validation model. The expression values for these genes are shown in the pink-o-gram shown below. Within the pink-o-gram, red indicates a gene is expressed at a relatively high level while blue indicates that a gene is expressed a relatively low level. Each column is a sample while each row in a gene with its corresponding label shown on the right. Expression profiles for the 32 cured DLBCLs are on the left while the expression profiles for the 26 fatal / refractory tumors are on the right. The top row on each half of the pink-o-gram represent idealized cured and fatal / refractory genes.

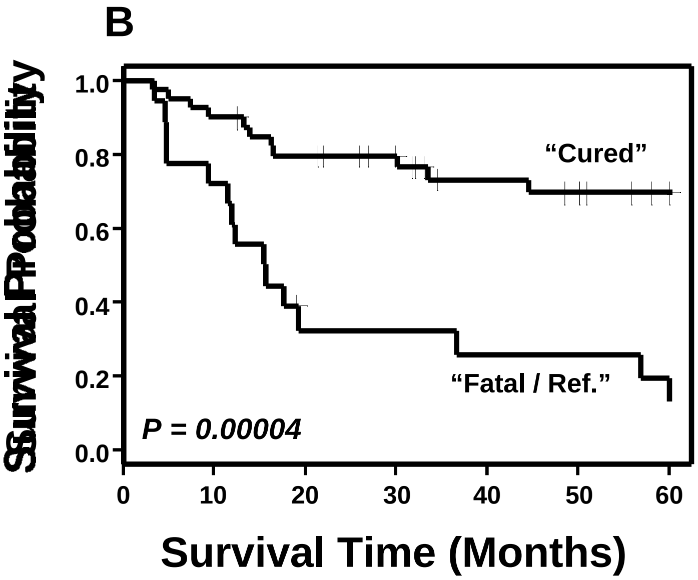


Acc. No.	Description	No. Models Using Gene
U43519	DRP2 Dystrophin related protein 2	58
Y09836	3'UTR of unknown protein	58
HG2314-HT2410	Uncharacterized	57
Z15114	PRKCG Protein kinase C, gamma	57
U12767	MINOR / NOR1	54
X77307	5-Hydorxytryptamine 2B Receptor	23
U83908	Nuclear antigen H731	58
M99435	Transducin-like enhancer protein 1	58
L20971	PDE4B Phosphodiesterase 4B	58
AC002450	Uncharacterized	58
M18255	PRKACB gene (protein kinase C-beta-1)	58
U09550	Oviductal glycoprotein mRNA	55
U38864	Zinc-finger protein C2H2-150	34

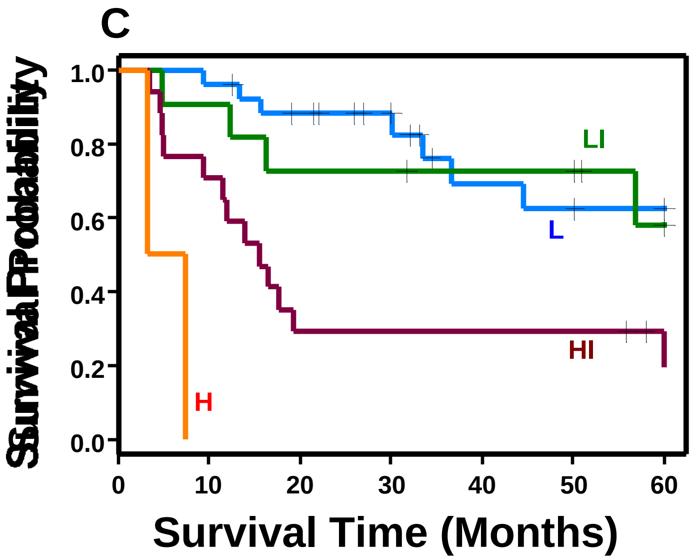
The following figure shows the Kaplan-Meier plot for the 5-year overall survival (OS) for the entire study group (all survival times greater than 5 years were truncated to 5 years). Thirty-three of the 58 DLBCL study patients remained alive after a median of 58 months of follow-up. The observed 5-year overall survival for the group as a whole was 54%.



The following figure shows the Kaplan-Meier plot for the 5-year overall survival for the predicted “cured” and “fatal/refractory” risk groups where the predicted groups were defined by the 13-gene model described above. The group of patients in the predicted “cured” group had 70% survival after five years versus only 12% 5-year survival for predicted “fatal/ref.” group (log-rank p-value = 0.00004).

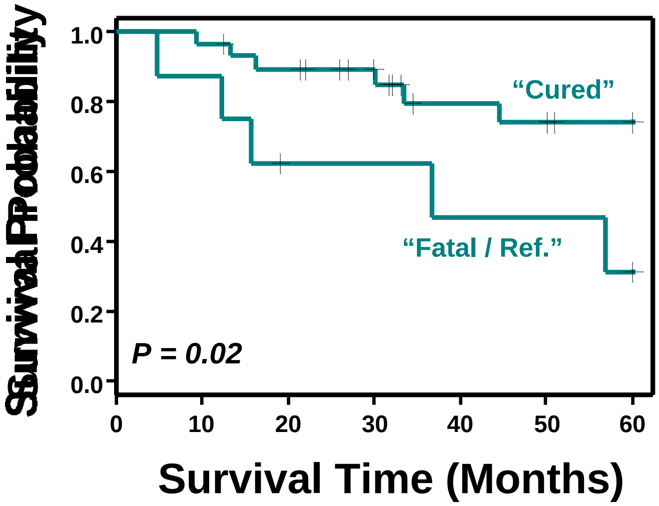


The following figure shows the Kaplan-Meier plot for the 5-year overall survival for the patients in low (L), low-intermediate (LI), high-intermediate (HI) and high (H)-risk categories as defined by the IPI[14] (IPI: L – 26 patients; LI – 11 patients; HI – 17 patients; H – 2 patients).



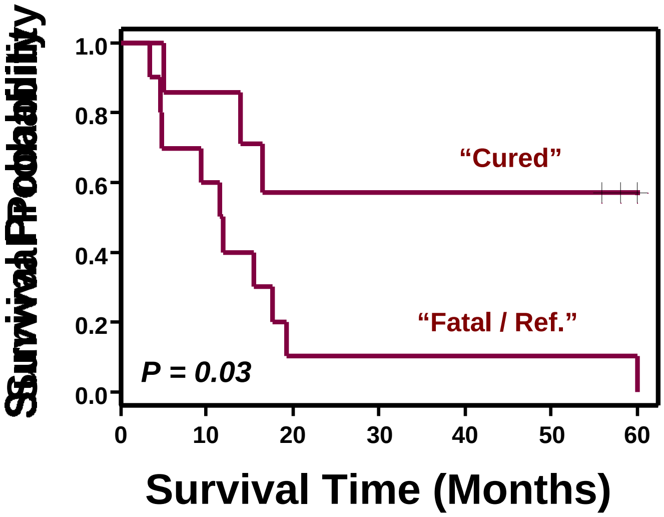
The following figure shows the Kaplan-Meier plot for the 5-year overall survival for the combined IPI L/LI-risk patients for the predicted “cured” and “fatal/refractory” risk groups where the predicted groups were defined by the 13-gene model described above. The group of L and LI patients in the predicted “cured” group had 75% 5-year overall survival versus only 32% of patients in the predicted “fatal/ref.” group had 5-year overall survival (nominal log-rank p-value = 0.02).

D



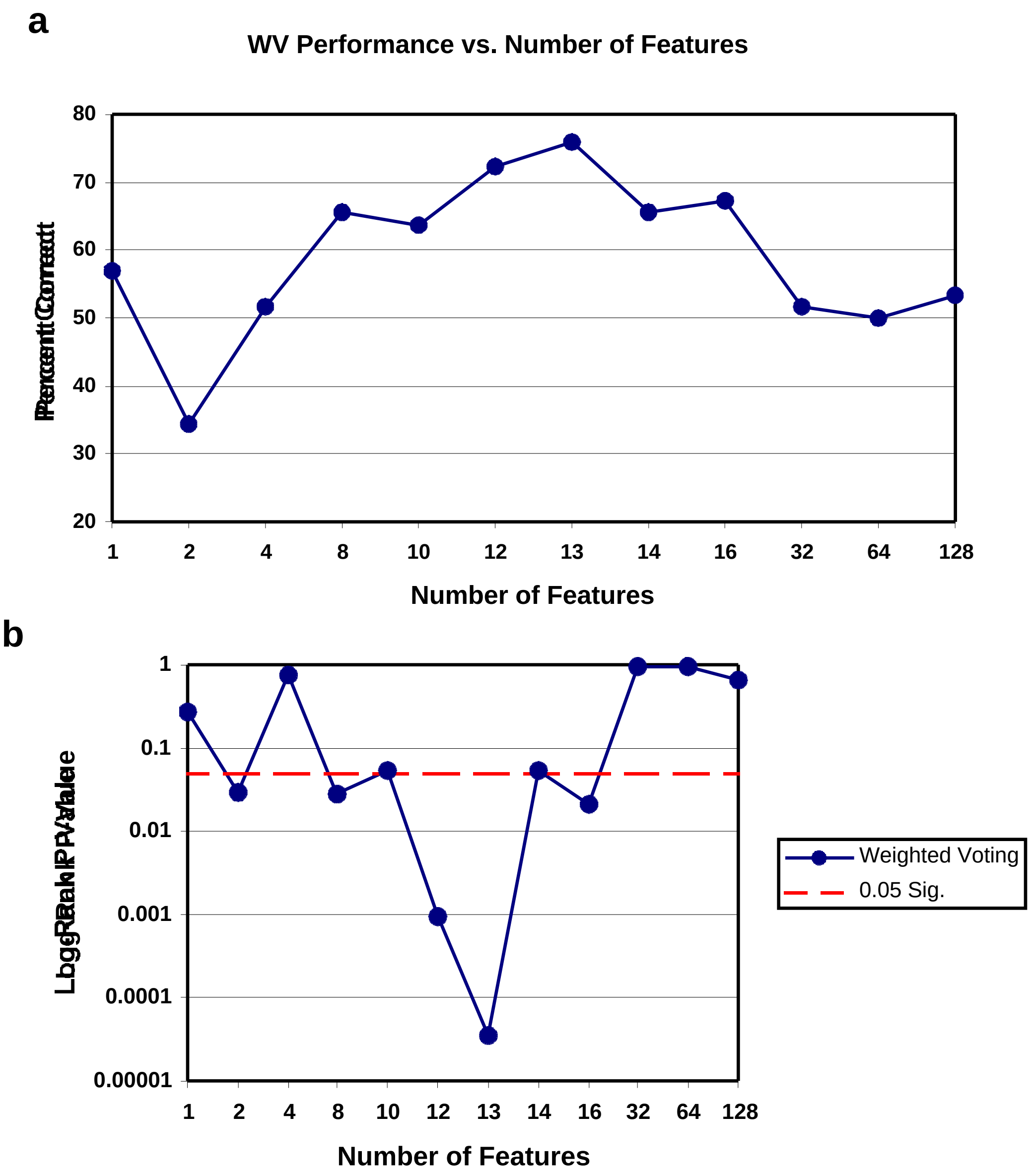
The following figure shows the Kaplan-Meier plot for the 5-year overall survival for the IPI HI-risk patients for the predicted “cured” and “fatal/refractory” risk groups where the predicted groups were defined by the 13-gene model described above. The group of HI patients in the predicted “cured” group had 57% 5-year overall survival versus only 0% of patients in the predicted “fatal/ref.” group had 5-year overall survival (nominal log-rank p-value = 0.02).

E



We decided to focus our attention on a thirteen gene weighted voting (WV) model based upon experiments with the “cured” versus “fatal/refractory” predictor with respect to the number of features used in the model. Below we show a table summarizing results and plots for performance with respect to the number of features for WV predictors that used mean in the signal-to-noise ratio. Within the figure below the table, the first plot (a) shows the percent correct as a function of the number of features and the second plot (b) shows the log-rank p-value as a function of the number of features.

Number of Features	Percent Correct	P-value
1	56.89655	0.271
2	34.48276	0.0298
4	51.72414	0.754
8	65.51724	0.0275
10	63.7931	0.0528
12	72.41379	0.000941
13	75.86207	3.55E-05
14	65.51724	0.0544
16	67.24138	0.021
32	51.72414	0.973
64	50	0.942
128	53.44828	0.648



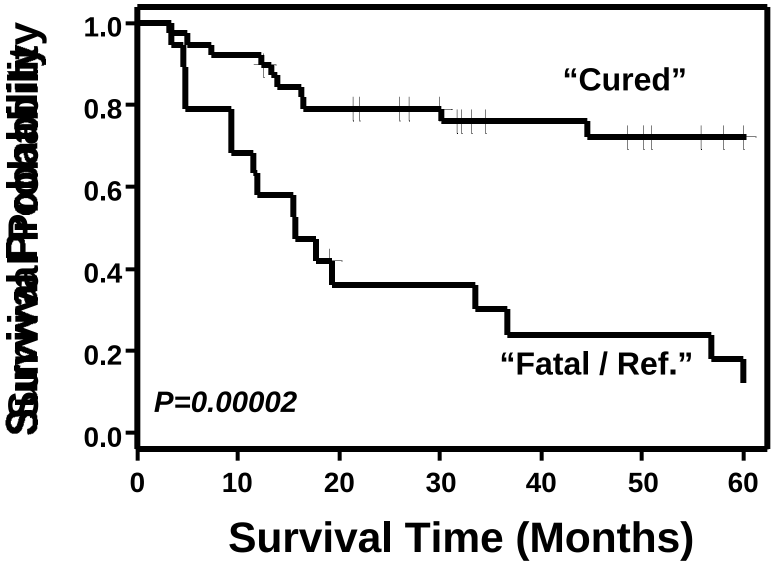
We looked at other methods for predicting outcome besides weighted voting (WV) including k-nearest neighbors (KNN) and support vector machines (SVM). These algorithms are described in detail in the background section. The KNN predictor was created using the GeneCluster software while the SVM predictor was created with custom software described in the background section. The SVM used its own feature selection method as described in the background section. The KNN outcome predictions used standard settings for preprocessing (expression values were thresholded at a minimum of 20 and a maximum of 16,000 and genes with maximum/minimum<3 (3-fold variation) or maximum-minimum<100 were filtered). Mean was used in the signal-to-noise calculation for KNN feature selection. The KNN outcome predictor was a nine-feature predictor that used the 7 nearest neighbors and distance weighting. The prediction results are shown in the table below.

Sample	True Class	SVM Predicted Class	KNN Predicted Class	Survival (months)	Truncated Survival

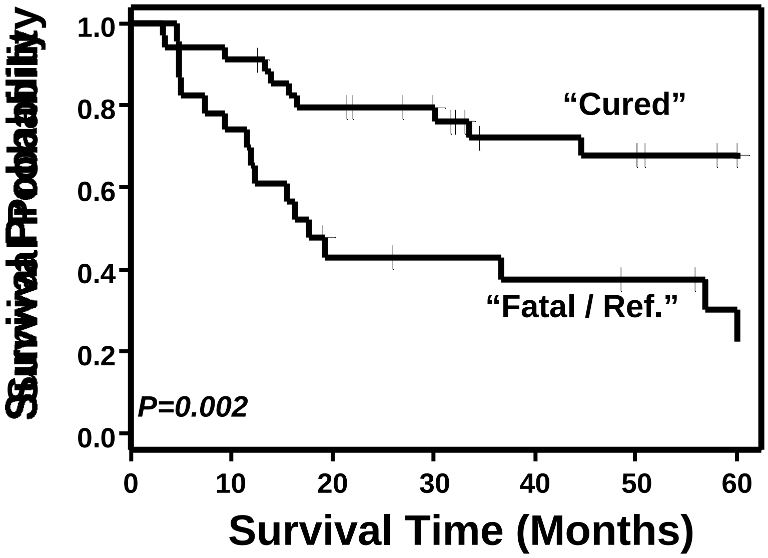
DLBC1	0	1	1	72.9	60
DLBC2	0	0	0	143.1	60
DLBC3	0	0	0	144.2	60
DLBC4	0	0	0	61	60
DLBC5	0	0	0	86.5	60
DLBC6	0	0	1	84.2	60
DLBC7	0	0	0	112.5	60
DLBC8	0	0	0	133.2	60
DLBC9	0	0	0	22.1	22.1
DLBC10	0	0	0	182.4	60
DLBC11	0	0	0	66.4	60
DLBC12	0	0	0	146.8	60
DLBC13	0	1	1	62.9	60
DLBC14	0	0	0	50.9	50.9
DLBC15	0	0	0	78.5	60
DLBC16	0	0	1	48.6	48.6
DLBC17	0	0	1	55.9	55.9
DLBC18	0	0	0	12.6	12.6
DLBC19	0	0	0	50.2	50.2
DLBC20	0	0	0	58	58
DLBC21	0	0	0	66.4	60
DLBC22	0	0	0	65.7	60
DLBC23	0	0	0	50.2	50.2
DLBC24	0	0	0	26.9	26.9
DLBC25	0	0	0	34.4	34.4
DLBC26	0	0	1	26	26
DLBC27	0	0	0	30	30
DLBC28	0	0	0	31.7	31.7
DLBC29	0	0	0	32.2	32.2
DLBC30	0	1	1	19.2	19.2
DLBC31	0	0	0	33.1	33.1
DLBC32	0	0	0	21.4	21.4
DLBC33	1	1	0	15.7	15.7
DLBC34	1	1	1	11.6	11.6
DLBC35	1	1	0	3.4	3.4
DLBC36	1	1	1	36.6	36.6
DLBC37	1	0	1	5	5
DLBC38	1	1	0	9.5	9.5
DLBC39	1	0	0	3.2	3.2
DLBC40	1	1	1	4.9	4.9
DLBC41	1	1	1	12	12
DLBC42	1	1	1	4.9	4.9
DLBC43	1	1	1	60.4	60
DLBC44	1	0	1	16.3	16.3
DLBC45	1	0	0	16.4	16.4
DLBC46	1	1	1	9.5	9.5

DLBC47	1	1	1	15.6	15.6
DLBC48	1	1	1	17.8	17.8
DLBC49	1	1	1	56.9	56.9
DLBC50	1	0	0	13.3	13.3
DLBC51	1	0	1	12.3	12.3
DLBC52	1	0	0	44.6	44.6
DLBC53	1	1	1	4.6	4.6
DLBC54	1	0	1	7.5	7.5
DLBC55	1	1	1	19.3	19.3
DLBC56	1	0	0	30.1	30.1
DLBC57	1	1	0	33.6	33.6
DLBC58	1	0	0	13.9	13.9

The following figure shows a Kaplan-Meier plot for the outcome predictor created by the SVM using the survival times that were truncated to 60 months. The group of patients in the predicted “cured” group had 72% 5-year overall survival versus only 12% of patients in the predicted “fatal/ref.” group had 5-year overall survival (nominal log-rank p-value = 0.00002).



The following figure shows a Kaplan-Meier plot for the outcome predictor created by the KNN method. The group of patients in the predicted “cured” group had 68% 5-year overall survival versus only 23% of patients in the predicted “fatal/ref.” group had 5-year overall survival (nominal log-rank p-value = 0.002).



In Silico ***Model Validation***

This section discusses the methods used to perform ***In Silico*** validation to discover if there was any connection between the lymphoma outcome prediction models presented in this paper and the cell-of-origin classification described by Alizadeh et al.[13] This ***In Silico*** validation used the lymphochip data from the Alizadeh ***et al*** paper. This comparison of results was difficult because i) different genes were measured by the arrays, ii) the microarray technology was different (cDNA versus oligonucleotide arrays, iii) different computational methods were used, and iv) different patient samples were studied.

Discovery of Genes Common to the Oligonucleotide and Lymphochip Data

First we set out to identify genes common to the cell-of-origin signature (Figure 3c of Alizadeh et al.) and the Affymetrix HU6800 oligonucleotide arrays. For the lymphochip data, we mapped the clone IMAGE numbers to GenBank accession numbers (using the list http://llmpp.nih.gov/lymphoma/data/clones.txt) and then mapped the accession numbers to Unigene cluster numbers. Similarly, we mapped accession numbers for our oligonucleotide array data to Unigene cluster numbers. Using this method, we identified 90 Unigene clusters that were common to both the Alizadeh et al. cell-of-origin signature and the oligonucleotide arrays. These 90 Unigene clusters are represented by 139 clones in the data of Figure 3c of Alizadeh et al.[13] and by 100 genes on the oligonucleotide arrays (after our array data has been passed through our standard thresholding and variational filtering).

The following table shows the list of 90 Unigene cluster ids and descriptions for genes common to both data sets (this table is available from our supplemental information website).

Unigene Tag	**Description**

Hs.108327	Damage-specific DNA binding protein 1 (127 kD)
Hs.115617	CRF-BP=corticotropin-releasing factor binding protein
Hs.115907	Diacylglycerol kinase delta
Hs.118021	ABR=guanine nucleotide regulatory protein
Hs.129695	WIP/HS PRPL-2=WASP interacting protein
Hs.1298	CD10=CALLA=Neprilysin=enkepalinase
Hs.129914	core binding factor alpha1b subunit=CBF alpha1=PEBP2aA1 transcription factor =AML1 Proto-oncogene=translocated in acute myeloid leukemia
Hs.146355	abl tyrosine-protein kinase
Hs.147097	histone H2A.X
Hs.151051	JNK3=Stress-activated protein kinase
Hs.151988	MAPKKK5=ASK1=mitogen-activated kinase kinase 5
Hs.154365	ELF-1=ets family transcription factor
Hs.155024	BCL-6
Hs.155342	PKC delta=Protein kinase C, delta
Hs.155530	IFI16=interferon-gamma-inducible myeloid differentiation transcriptional activator
Hs.155894	PTP-1B=phosphotyrosyl-protein phosphatase
Hs.157441	spi-1=PU.1=ets family transcription factor
Hs.167246	Cytochrome P450 reductase
Hs.169081	Tel=ets family transcription factor translocated in acute leukemias
Hs.169610	CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor
Hs.169832	zinc finger protein 42 MZF-1
Hs.169948	Potassium voltage-gated channel, shaker-related subfamily, member 3
Hs.170195	OP-1=osteogenic protein in the TGF-beta family
Hs.171763	CD22
Hs.172195	Unknown UG Hs.172195 ESTs, Weakly similar to KIAA0226 [H.sapiens]
Hs.173936	Cytokine receptor family II, member 4
Hs.180677	ZFM1=signal transduction and activator of RNA (STAR) transcription factor=splicing factor SF1
Hs.180841	CD27
Hs.180919	Id2=Id2H=Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
Hs.181390	casein kinase I gamma 2
Hs.184402	cam kinase I
Hs.184585	TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif
Hs.188	3' 5'-cyclic AMP phosphodiesterase=rolipram-sensitive cAMP-specific phosphodiesterase (PDE4B)
Hs.195175	FLICE-like inhibitory protein long form=I-FLICE=FLAME-1=Casper=MRIT=CASH=cFLIP=CLARP
Hs.197540	HIF-1 alpha=hypoxia-inducible factor 1 alpha
Hs.203420	tyrosine kinase (Tnk1)
Hs.211563	BCL-7A
Hs.211588	RDC-1=POU domain transcription factor
Hs.211773	checkpoint suppressor 1
Hs.239138	PBEF=pre-B cell enhancing factor
Hs.241510	HNPP=nuclear phosphoprotein
Hs.24340	Unknown
Hs.250505	RAR-alpha-1=Retinoic acid receptor
Hs.252280	Unknown UG Hs.83583 actin related protein 2/3 complex, subunit 2 (34 kD)
Hs.2537	myb-related gene A=A-myb

Hs.256278	TNFR2=TNF alpha Receptor II=p80
Hs.271980	erk3=extracellular signal-regulated kinase 3
Hs.278597	BDP1=protein-tyrosine-phosphatase
Hs.278674	thymosin beta-4
Hs.32942	Phosphatidylinositol 3-kinase p110 catalytic, gamma isoform
Hs.3446	MEK1=MAP kinase kinase 1
Hs.40202	JAW1=lymphoid-restricted membrane protein
Hs.44566	Unknown UG Hs.97530 ESTs
Hs.47007	NIK=serine/threonine protein kinase
Hs.54472	FMR2=Fragile X mental retardation 2=putative transcription factor=LAF-4 and AF-4 homologue
Hs.66052	CD38
Hs.72927	IL-7
Hs.73792	CD21=B-lymphocyte CR2-receptor (for complement factor C3d and Epstein-Barr virus)
Hs.75339	51C protein=Similar to signaling inositol polyphosphate 5 phosphatase SIP-110
Hs.75367	SLAP=src-like adapter protein
Hs.75545	IL-4 receptor alpha chain
Hs.75586	Cyclin D2/KIAK0002=3Õ end of KIAK0002 cDNA
Hs.75596	IL-2 receptor beta chain
Hs.75859	Similar to (Z72511) F55A11.3
Hs.76894	Deoxycytidylate deaminase
Hs.77617	SP100=Nuclear body protein
Hs.78353	SRPK2 serine kinase
Hs.784	EBI2=Epstein-Barr virus induced G-protein coupled receptor=Putative chemokine receptor
Hs.78436	NET PTK=tyrosine kinase
Hs.79070	c-myc
Hs.79241	BCL-2
Hs.79933	Cyclin I
Hs.81221	Immunoglobulin heavy chain V(H)5 pseudogene L2-9 transcript
Hs.82127	IL-16=Lymphocyte chemoattractant factor (LCF)
Hs.82132	IRF-4=LSIRF=Mum1=homologue of Pip=Lymphoid-specific interferon regulatory factor =Multiple myeloma oncogene 1
Hs.82251	myosin-IC
Hs.82829	T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2
Hs.82845	clone 23815 mRNA
Hs.82911	Unknown UG Hs.82911 protein tyrosine phosphatase type IVA, member 2
Hs.82979	Germinal center kinase=BL44=B lymphocyte serine/threonine protein kinase
Hs.85283	LYSP100=SP140
Hs.86958	Interferon alpha/beta receptor-2=Interferon-alpha/beta receptor beta chain precursor=IFN-alpha-rec=Type I interferon receptor=IFN-R
Hs.89230	KCNN3=SKCA3=AAD14=calcium-activated potassium channel
Hs.89499	Arachidonate 5-lipoxygenase=5-lipoxygenase=5-LO
Hs.9235	nucleoside-diphosphate kinase
Hs.9408	KIAA0151=serine/threonine kinase
Hs.95821	osteoclast stimulating factor=contains SH3 domain and ankyrin repeat
Hs.96	APR=immediate-early-response gene=ATL-derived PMA-responsive peptide
Hs.96063	IRS-1=Insulin receptor substrate-1
Hs.96398	OGG1=8-oxoguanine DNA glycosylase=DNA alkylation repair protein

The following table shows the list of 129 (plus 10 duplicates) cell-of-origin clones that are on the Lymphochip and have corresponding features on the HU6800 oligonucleotide array (this table is available on the supplemental information website).

Unigene ID	Clone ID	Accession	Num. Copies	Description
Hs.108327	279482	N48804	1	Unknown UG Hs.108327 damage-specific DNA binding protein 1 (127kD)
Hs.115617	193828	H51657	1	CRF-BP=corticotropin-releasing factor binding protein
Hs.115907	705274	AA280692	2	Diacylglycerol kinase delta
Hs.118021	52408	H23143	1	ABR=guanine nucleotide regulatory protein
Hs.129695	1319062	AA811088	1	WIP/HS PRPL-2=WASP interacting protein
Hs.129695	1337701	AA811758	1	WIP/HS PRPL-2=WASP interacting protein
Hs.1298	200814	R98936	1	CD10=CALLA=Neprilysin=enkepalinase
Hs.1298	701606	AA287043	1	CD10=CALLA=Neprilysin=enkepalinase
Hs.1298	1286850	AA741127	1	CD10=CALLA=Neprilysin=enkepalinase
Hs.129914	157828	R72866	1	core binding factor alpha1b subunit=CBF alpha1=PEBP2aA1 transcription factor =AML1 Proto-oncogene=translocated in acute myeloid leukemia
Hs.146355	1306105	AA765967	1	abl tyrosine-protein kinase
Hs.147097	687166	AA258156	1	fos39554_1 predicted protein from fosmid 39554
Hs.147097	713213	AA283631	1	fos39554_1 predicted protein from fosmid 39554
Hs.147097	824366	AA489684	1	Unknown UG Hs.19399 Homo sapiens chromosome 19, fosmid 39554
Hs.151051	23173	R39221	1	JNK3=Stress-activated protein kinase
Hs.151988	28450	R40676	1	MAPKKK5=ASK1=mitogen-activated kinase kinase 5
Hs.154365	201976	R99515	1	ELF-1=ets family transcription factor
Hs.155024	712395	AA281781	1	BCL-6
Hs.155342	428733	AA005215	2	PKC delta=Protein kinase C, delta
Hs.155342	1289165	AA761831	1	PKC delta=Protein kinase C, delta
Hs.155530	824602	AA490996	1	IFI16=interferon-gamma-inducible myeloid differentiation transcriptional activator
Hs.155894	472182	AA057376	1	PTP-1B=phosphotyrosyl-protein phosphatase
Hs.155894	685177	AA252649	1	PTP-1B=phosphotyrosyl-protein phosphatase
Hs.157441	278808	N66572	1	spi-1=PU.1=ets family transcription factor
Hs.167246	234180	H70626	1	Cytochrome P450 reductase
Hs.169081	35356	R45543	1	Neurotrophic tyrosine kinase, receptor, type 3 (TrkC)
Hs.169081	1355435	AA831368	1	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene)
Hs.169610	703824	AA279047	1	CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor
Hs.169610	713145	AA283090	2	CD44=Pgp-1=extracellular matrix receptor-III=Hyaluronate receptor
Hs.169832	490387	AA120779	1	zinc finger protein 42 MZF-1
Hs.169948	1337856	AA811374	1	Potassium voltage-gated channel, shaker-related subfamily, member 3
Hs.170195	344430	W73473	2	OP-1=osteogenic protein in the TGF-beta family
Hs.171763	1234404	AA687354	1	CD22
Hs.172195	1352465	AA828553	1	Unknown UG Hs.172195 ESTs, Weakly similar to KIAA0226 [H.sapiens]
Hs.173936	202498	H53121	1	Cytokine receptor family II, member 4
Hs.180677	701059	AA287877	1	ZFM1=signal transduction and activator of RNA (STAR) transcription factor=splicing factor SF1
Hs.180841	34637	R45026	2	CD27
Hs.180841	1288550	AA761422	1	CD27
Hs.180841	1353030	AA830238	1	CD27
Hs.180919	324873	W49655	1	Id2=Id2H=Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein

Hs.180919	704815	AA282782	1	Id2=Id2H=Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
Hs.180919	814824	AA465647	1	Id2=Id2H=Inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
Hs.181390	346031	W72092	1	casein kinase I gamma 2
Hs.181390	510002	AA052932	1	casein kinase I gamma 2
Hs.181390	687112	AA258926	1	Unknown UG Hs.181390 casein kinase 1, gamma 2
Hs.181390	1234475	AA687130	1	casein kinase I gamma 2
Hs.184402	52629	H29322	1	cam kinase I
Hs.184402	1357636	AA831996	1	cam kinase I
Hs.184585	685456	AA261902	1	TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif
Hs.184585	712829	AA280651	1	TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif
Hs.188	377708	AA056219	1	3' 5'-cyclic AMP phosphodiesterase=rolipram-sensitive cAMP-specific phosphodiesterase (PDE4B)
Hs.195175	427786	AA002262	1	FLICE-like inhibitory protein long form=I-FLICE=FLAME-1=Casper=MRIT=CASH=cFLIP=CLARP
Hs.197540	325117	W47003	1	HIF-1 alpha=hypoxia-inducible factor 1 alpha
Hs.203420	1317098	AA767135	1	tyrosine kinase (Tnk1)
Hs.211563	306139	N91028	1	BCL-7A
Hs.211563	1337241	AA812170	1	BCL-7A
Hs.211588	773568	AA428196	1	RDC-1=POU domain transcription factor
Hs.211773	814651	AA481039	1	checkpoint suppressor 1
Hs.239138	1270880	AA748507	1	PBEF=pre-B cell enhancing factor
Hs.241510	154493	R54613	1	HNPP=nuclear phosphoprotein
Hs.24340	1351593	AA806970	1	Unknown
Hs.250505	159381	H15011	1	RAR-alpha-1=Retinoic acid receptor
Hs.252280	1305130	AA872531	1	p115-RhoGEF=guanine nucleotide exchange factor similar to Lsc oncogene (Mus)=Actin related protein 2/3 complex, subunit 2 (34kD)
Hs.2537	825476	AA504350	1	myb-related gene A=A-myb
Hs.256278	71046	T47383	1	TNFR2=TNF alpha Receptor II=p80
Hs.271980	684169	AA251095	1	erk3=extracellular signal-regulated kinase 3
Hs.278597	953383	AA527826	1	BDP1=protein-tyrosine-phosphatase
Hs.278674	150804	H02553	1	thymosin beta-4
Hs.32942	290151	N63285	1	Phosphatidylinositol 3-kinase p110 catalytic, gamma isoform
Hs.32942	1251617	AA810310	1	Phosphatidylinositol 3-kinase p110 catalytic, gamma isoform
Hs.32942	1358163	AA826284	1	Phosphatidylinositol 3-kinase p110 catalytic, gamma isoform
Hs.3446	309258	N98340	1	MEK1=MAP kinase 1
Hs.40202	417502	W88799	1	JAW1=lymphoid-restricted membrane protein
Hs.40202	815539	AA457051	2	JAW1=lymphoid-restricted membrane protein
Hs.44566	1286300	AA740710	1	Unknown UG Hs.97530 ESTs
Hs.47007	342349	W61116	1	NIK=serine/threonine protein kinase
Hs.47007	1333762	AA812002	1	NIK=serine/threonine protein kinase
Hs.54472	1352112	AA808138	1	FMR2=Fragile X mental retardation 2=putative transcription factor=LAF-4 and AF-4 homologue
Hs.66052	123264	R00276	1	CD38
Hs.72927	701422	AA287945	1	IL-7
Hs.73792	814917	AA465705	1	CD21=B-lymphocyte CR2-receptor (for complement factor C3d and Epstein-Barr virus)
Hs.73792	824695	AA482292	1	CD21=B-lymphocyte CR2-receptor (for complement factor C3d and Epstein-

				Barr virus)
Hs.75339	686441	AA252822	1	Unknown UG Hs.75339 inositol polyphosphate phosphatase-like 1
Hs.75367	52564	H29540	1	SLAP=src-like adapter protein
Hs.75367	701768	AA284417	1	SLAP=src-like adapter protein
Hs.75367	815774	AA485141	1	SLAP=src-like adapter protein
Hs.75545	701796	AA292716	1	IL-4 receptor alpha chain
Hs.75545	714453	AA293306	1	IL-4 receptor alpha chain
Hs.75586	366412	AA026336	1	Cyclin D2/KIAK0002=3Õ end of KIAK0002 cDNA
Hs.75586	1357360	AA831970	1	Cyclin D2/KIAK0002=3Õ end of KIAK0002 cDNA
Hs.75596	489079	AA057204	1	IL-2 receptor beta chain
Hs.75859	1270305	AA748184	1	Unknown UG Hs.75859 chromosome 11 open reading frame 4
Hs.76894	489681	AA099570	2	Deoxycytidylate deaminase
Hs.76894	1302032	AA731918	1	Deoxycytidylate deaminase
Hs.77617	1367790	AA810108	1	SP100=Nuclear body protein
Hs.78353	21899	T65211	1	SRPK2 serine kinase
Hs.78353	1283120	AA744494	1	SRPK2 serine kinase
Hs.784	182764	H45306	1	EBI2=Epstein-Barr virus induced G-protein coupled receptor=Putative chemokine receptor
Hs.78436	153760	R48171	2	NET PTK=tyrosine kinase
Hs.79070	417226	W87741	1	c-myc
Hs.79241	232714	H74208	1	BCL-2
Hs.79241	342181	W63749	2	BCL-2
Hs.79933	248295	N58511	1	Cyclin I
Hs.81221	825648	AA505045	1	Immunoglobulin heavy chain V(H)5 pseudogene L2-9 transcript
Hs.82127	701504	AA286934	1	IL-16=Lymphocyte chemoattractant factor (LCF)
Hs.82127	714394	AA293249	1	IL-16=Lymphocyte chemoattractant factor (LCF)
Hs.82132	270770	N33454	1	IRF-4=LSIRF=Mum1=homologue of Pip=Lymphoid-specific interferon regulatory factor =Multiple myeloma oncogene 1
Hs.82132	1272196	AA743459	1	IRF-4=LSIRF=Mum1=homologue of Pip=Lymphoid-specific interferon regulatory factor =Multiple myeloma oncogene 1
Hs.82251	488373	AA044832	1	myosin-IC
Hs.82251	1184506	AA648536	1	myosin-IC
Hs.82829	665903	AA193262	1	T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2
Hs.82829	740402	AA477822	1	T-cell protein-tyrosine phosphatase=Protein tyrosine phosphatase, non-receptor type 2
Hs.82845	294506	N71007	1	clone 23815 mRNA
Hs.82911	1671743	AI056571	1	Unknown UG Hs.82911 protein tyrosine phosphatase type IVA, member 2
Hs.82979	37234	R50953	2	Germinal center kinase=BL44=B lymphocyte serine/threonine protein kinase
Hs.85283	229723	H66484	1	LYSP100=SP140
Hs.86958	366884	AA029587	1	Interferon alpha/beta receptor-2=Interferon-alpha/beta receptor beta chain precursor=IFN-alpha-rec=Type I interferon receptor=IFN-R
Hs.89230	824081	AA491238	1	KCNN3=SKCA3=AAD14=calcium-activated potassium channel
Hs.89499	826541	AA521143	1	Arachidonate 5-lipoxygenase=5-lipoxygenase=5-LO
Hs.89499	1013324	AA552491	1	Arachidonate 5-lipoxygenase=5-lipoxygenase=5-LO
Hs.9235	203003	H54417	1	nucleoside-diphosphate kinase
Hs.9408	364448	AA022666	1	KIAA0151=serine/threonine kinase
Hs.9408	1271889	AA743048	1	KIAA0151=serine/threonine kinase

Hs.95821	1351622	AA806978	1	osteoclast stimulating factor=contains SH3 domain and ankyrin repeat
Hs.96	328550	W40261	1	APR=immediate-early-response gene=ATL-derived PMA-responsive peptide
Hs.96	685398	AA262439	1	APR=immediate-early-response gene=ATL-derived PMA-responsive peptide
Hs.96063	796284	AA460841	1	IRS-1=Insulin receptor substrate-1
Hs.96398	1288192	AA761117	1	OGG1=8-oxoguanine DNA glycosylase=DNA alkylation repair protein
Hs.96398	1351027	AA806527	1	OGG1=8-oxoguanine DNA glycosylase=DNA alkylation repair protein

The following table lists the 100 probes on the Affymetrix oligonucleotide arrays that have a corresponding clone appearing on the Lymphochip (this table is also available on the supplemental information website).

Unigene ID	**Probe ID**	**Accession**	**Description**
Hs.108327	U32986_s_at	U32986	Human xeroderma pigmentosum group E UV-damaged DNA binding factor
Hs.115907	D63479_s_at	D63479	Human mRNA for KIAA0145 gene, complete cds
Hs.118021	U01147_at	U01147	Human guanine nucleotide regulatory protein (ABR) mRNA, complete cds
Hs.129695	X86019_at	X86019	H.sapiens mRNA for PRPL-2 protein
Hs.1298	J03779_at	J03779	Human common acute lymphoblastic leukemia antigen (CALLA)
Hs.129914	D43968_at	D43968	Human AML1 mRNA for AML1b protein (alternatively spliced product),
Hs.129914	X90978_at	X90978	H.sapiens mRNA for an acute myeloid leukaemia protein (1793bp)
Hs.146355	U07563_cds1_at	U07563	Human ABL gene, exon 1b and intron 1b, and putative M8604 Met protein (M8604 Met) gene
Hs.146355	X16416_at	X16416	Human c-abl mRNA encoding p150 protein
Hs.147097	X14850_at	X14850	Human H2A.X mRNA encoding histone H2A.X
Hs.151051	U07620_at	U07620	Human MAP kinase mRNA, complete cds
Hs.151988	U67156_at	U67156	Human mitogen-activated kinase 5 (MAPKKK5) mRNA
Hs.154365	M82882_at	M82882	Human cis-acting sequence
Hs.155024	U00115_at	U00115	Human zinc-finger protein (bcl-6) mRNA, complete cds
Hs.155342	D10495_at	D10495	Human mRNA for protein kinase C delta-type
Hs.155530	M63838_s_at	M63838	Human interferon-gamma induced protein (IFI 16) gene
Hs.155894	M31724_at	M31724	Human phosphotyrosyl-protein phosphatase (PTP-1B) mRNA
Hs.155894	M33684_s_at	M33684	Human (clone lambda-10-2) non-receptor tyrosine phosphatase 1 (PTPN1) gene
Hs.157441	X52056_at	X52056	Human mRNA for spi-1 proto-oncogene
Hs.167246	S90469_at	S90469	cytochrome P450 reductase [human, placenta, mRNA Partial, 2403 nt]
Hs.169081	U11732_at	U11732	Human ets-like gene (tel) mRNA, complete cds
Hs.169610	L05424_cds2_at	L05424	CD44 gene (cell surface glycoprotein CD44) extracted from Human hyaluronate receptor (CD44) gene
Hs.169832	M58297_at	M58297	Human zinc finger protein 42 (MZF-1) mRNA, complete cds
Hs.170195	X51801_at	X51801	Human OP-1 mRNA for osteogenic protein
Hs.171763	X59350_at	X59350	H.sapiens mRNA for B cell membrane protein CD22
Hs.172195	U15128_at	U15128	Human beta-1,2-N-acetylglucosaminyltransferase II (MGAT2) gene
Hs.173936	Z17227_at	Z17227	H.sapiens mRNA for transmenbrane receptor protein
Hs.180677	L49380_at	L49380	Homo sapiens clone B4 transcription factor ZFM1 mRNA
Hs.180677	Y08765_s_at	Y08765	H.sapiens mRNA for splicing factor, SF1-HL1 isoform
Hs.180841	M63928_at	M63928	Homo sapiens T cell activation antigen (CD27) mRNA, complete cds
Hs.180919	M96843_at	M96843	Human striated muscle contraction regulatory protein (Id2B) mRNA, complete cds
Hs.180919	M97796_s_at	M97796	Human helix-loop-helix protein (Id-2) mRNA, complete cds
Hs.181390	U89896_at	U89896	Human casein kinase I gamma 2 mRNA, complete cds
Hs.184402	L41816_at	L41816	Homo sapiens cam kinase I mRNA, complete cds

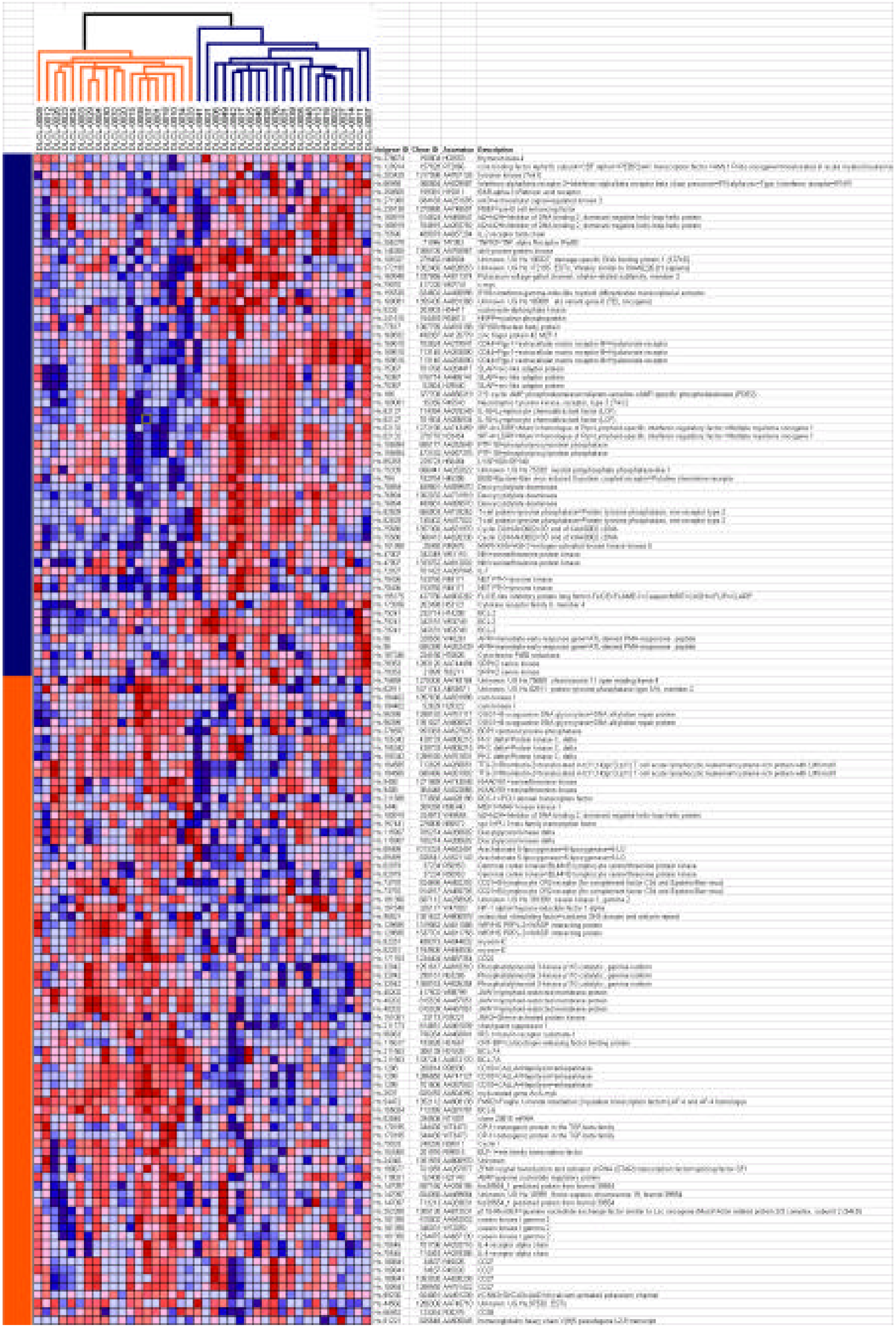
Hs.184585	X61118_rna1_at	X61118	TTG-2a gene extracted from Human TTG-2 mRNA for a cysteine rich protein with LIM motif
Hs.188	L20971_at	L20971	Human phosphodiesterase mRNA, complete cds
Hs.195175	AF005775_at	AF005775	Homo sapiens caspase-like apoptosis regulatory protein 2 (clarp) mRNA, alternatively spliced, complete cds.
Hs.197540	U22431_s_at	U22431	Human hypoxia-inducible factor 1 alpha (HIF-1 alpha) mRNA, complete cds
Hs.203420	U43408_at	U43408	Human tyrosine kinase (Tnk1) mRNA, complete cds
Hs.211563	X89984_at	X89984	H.sapiens mRNA for BCL7A protein
Hs.211588	X64624_s_at	X64624	H.sapiens mRNA for RDC-1 POU domain containing protein
Hs.211773	U68723_at	U68723	Human checkpoint suppressor 1 mRNA, complete cds.
Hs.211973	U07563_cds1_at	U07563	Human ABL gene, exon 1b and intron 1b, and putative M8604 Met protein (M8604 Met) gene
Hs.239138	U02020_at	U02020	Human pre-B cell enhancing factor (PBEF) mRNA, complete cds+C91
Hs.241510	L22342_at	L22342	Human nuclear phosphoprotein mRNA, complete cds
Hs.24340	D26069_at	D26069	Human mRNA for KIAA0041 gene, partial cds
Hs.250505	X06614_at	X06614	Human mRNA for receptor of retinoic acid
Hs.252280	U64105_at	U64105	Human guanine nucleotide exchange factor p115-RhoGEF mRNA, partial cds
Hs.2537	S75881_s_at	S75881	A-myb=DNA-binding transactivator {3 region} [human, CCRF-CEM T-leukemia line, mRNA Partial, 831 nt]
Hs.2537	X66087_at	X66087	H.sapiens a-myb mRNA
Hs.256278	M32315_at	M32315	Human tumor necrosis factor receptor mRNA, complete cds
Hs.271980	X80692_at	X80692	H.sapiens ERK3 mRNA
Hs.278597	X79568_at	X79568	H.sapiens BDP1 mRNA for protein-tyrosine-phosphatase
Hs.278674	D85181_at	D85181	Human mRNA for fungal sterol-C5-desaturase homolog, complete cds
Hs.32942	X83368_at	X83368	H.sapiens mRNA for phosphatidylinositol 3 kinase gamma
Hs.3446	L05624_s_at	L05624	Homo sapiens MAP kinase mRNA, complete cds
Hs.3446	L11284_at	L11284	Homosapiens ERK activator kinase (MEK1) mRNA
Hs.40202	U10485_at	U10485	Human lymphoid-restricted membrane protein (Jaw1) mRNA, complete cds
Hs.44566	U28831_at	U28831	Human protein immuno-reactive with anti-PTH polyclonal antibodies
Hs.47007	Y10256_at	Y10256	H.sapiens mRNA for serine/threonine protein kinase, NIK
Hs.54472	U48436_s_at	U48436	Human fragile X mental retardation protein FMR2p (FMR2) mRNA
Hs.66052	D84276_at	D84276	Human mRNA for CD38, complete cds
Hs.72927	J04156_at	J04156	Human interleukin 7 (IL-7) mRNA, complete cds
Hs.73792	M26004_s_at	M26004	Human CR2/CD21/C3d/Epstein-Barr virus receptor mRNA, complete cds
Hs.73792	S62696_s_at	S62696	EBV/C3d receptor {alternatively spliced, exons 8a,9,10} [human, Jurkat T cells, mRNA Partial, 151 nt]
Hs.75339	L36818_at	L36818	Human (clone 51C-3) 51C protein mRNA, complete cds
Hs.75367	D89077_at	D89077	Human mRNA for Src-like adapter protein, complete cds
Hs.75545	X52425_at	X52425	Human IL-4-R mRNA for the interleukin 4 receptor
Hs.75586	D13639_at	D13639	Human mRNA for KIAK0002 gene, complete cds
Hs.75596	M26062_at	M26062	Human interleukin 2 receptor beta chain (p70-75) mRNA, complete cds
Hs.75859	U39400_at	U39400	Human NOF1 mRNA, complete cds
Hs.76894	L39874_at	L39874	Homo sapiens deoxycytidylate deaminase gene, complete cds
Hs.77617	U36501_at	U36501	Human SP100-B (SP100-B) mRNA, complete cds
Hs.78353	U88666_at	U88666	Human serine kinase SRPK2 mRNA, complete cds
Hs.784	L08177_at	L08177	Human EBV induced G-protein coupled receptor (EBI2) mRNA, complete cds
Hs.78436	L40636_at	L40636	Homo sapiens (clone FBK III 16) protein tyrosine kinase (NET PTK)
Hs.79070	L00058_at	L00058	Human (GH) germline c-myc proto-oncogene, 5 flank

Hs.79070	M13929_s_at	M13929	Human c-myc-P64 mRNA, initiating from promoter P0, (HLmyc2.5) partial cds
Hs.79241	M13994_s_at	M13994	Human B-cell leukemia/lymphoma 2 (bcl-2) proto-oncogene mRNA encoding bcl-2-alpha protein, complete cds
Hs.79241	M14745_at	M14745	Human bcl-2 mRNA
Hs.79933	D50310_at	D50310	Human mRNA for cyclin I, complete cds
Hs.81221	X58399_at	X58399	Human L2-9 transcript of unrearranged immunoglobulin V(H)5 pseudogene.
Hs.82127	M90391_s_at	M90391	Human putative IL-16 protein precursor, mRNA, complete cds
Hs.82132	U52682_at	U52682	Human lymphocyte specific interferon regulatory factor/interferon regulatory factor 4 (LSIRF/IRF4), complete cds
Hs.82251	U14391_at	U14391	Human myosin-IC mRNA, complete cds
Hs.82829	M25393_at	M25393	Human protein tyrosine phosphatase (PTPase) mRNA, complete cds
Hs.82845	U90916_at	U90916	Human clone 23815 mRNA sequence
Hs.82911	U14603_at	U14603	Human protein-tyrosine phosphatase (HU-PP-1) mRNA, partial sequence
Hs.82979	U07349_at	U07349	Human B lymphocyte serine/threonine protein kinase mRNA, complete cds
Hs.85283	U36500_at	U36500	Human lymphoid-specific SP100 homolog (LYSP100-B) mRNA, complete cds
Hs.86958	L42243_cds1_at	L42243	IFNAR2 gene (interferon receptor) extracted from Homo sapiens (clone Q-2OD3) interferon receptor (IFNAR2) gene
Hs.89230	Y08263_at	Y08263	H.sapiens mRNA for AAD14 protein, partial
Hs.89499	J03600_at	J03600	Human lipoxygenase mRNA, complete cds
Hs.9235	Y07604_at	Y07604	H.sapiens mRNA for nucleoside-diphosphate kinase
Hs.9408	D63485_at	D63485	Human mRNA for KIAA0151 gene, complete cds
Hs.95821	U63717_at	U63717	Human osteoclast stimulating factor mRNA, complete cds
Hs.96	D90070_s_at	D90070	Human ATL-derived PMA-responsive (APR) peptide mRNA
Hs.96063	S62539_at	S62539	insulin receptor substrate-1 [human, skeletal muscle, mRNA, 5828 nt]
Hs.96063	S85963_at	S85963	hIRS-1=rat insulin receptor substrate-1 homolog [human, cell line FOCUS, Genomic, 6152 nt]
Hs.96398	AB000410_s_at	AB000410	Human hOGG1 mRNA, complete cds

Clustering Based upon Putative Cell-of-Origin

Separate data files were created containing expression data for only the cell-of-origin data that was common to both the Alizadeh et al.[13] lymphochip and our oligonucleotide arrays. Lymphochip cell-of-origin data was obtained from the public website (http://llmpp.nih.gov/lymphoma) as contained in the file figure3c.cdt and a data subset was formed by selecting the genes common to both data sets. The Alizadeh et al.[13] DLBCL series and our DLBCL series were separately clustered using the clones or oligonucleotide probes representing these common cell-of-origin signature genes and a hierarchical clustering program[4]. Results were visualized using TreeView software[4]. Average linked-clustering was used, which organizes all of the data elements into a single tree with the highest levels of the tree representing the discovered classes.

Below are shown the results from hierarchical clustering of the cell-of-origin genes common to both sets using the data from the Alizadeh et al. data. The samples were represented by the expression levels of the 139 clones corresponding to the 90 common cell-of-origin Unigene clusters and were clustered using average linkage clustering with an uncentered correlation similarity metric and no additional preprocessing. This figure is a version of figure 5A of our paper that includes the details of the hierarchical clustering (an expanded version of this figure can be downloaded from our website http://www-genome.wi.mit.edu/MPR/lymphoma).


Unigene ID
Clone ID
Accession
Description

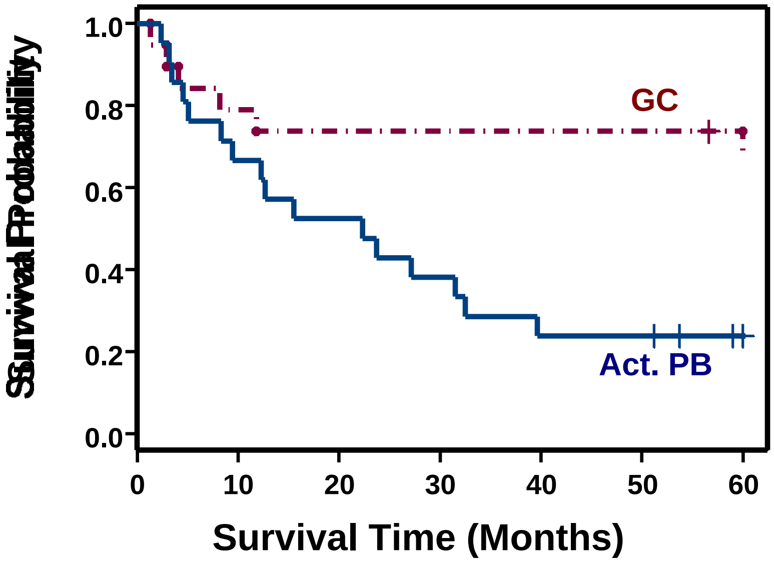
The results of the hierarchical clustering of the Alizadeh et al. data are summarized in the following table.

Sample Identifier	Cluster	Survival Category	Overall Survival	Alizadeh Cluster	Truncated Survival
DLCL-0001	0	0	77.4	0	60
DLCL-0004	0	0	69.6	0	60
DLCL-0005	1	0	51.2	1	51.2
DLCL-0008	0	0	102.4	0	60
DLCL-0009	0	0	89.8	0	60
DLCL-0010	0	0	88.1	0	60
DLCL-0014	1	0	59	1	59
DLCL-0015	0	0	56.6	0	56.6
DLCL-0020	0	0	80.4	0	60
DLCL-0024	0	0	129.9	0	60
DLCL-0028	1	0	90.2	1	60
DLCL-0029	0	0	83.8	0	60
DLCL-0030	0	0	71.3	0	60
DLCL-0032	0	0	69.1	0	60
DLCL-0033	0	0	68.8	0	60
DLCL-0037	0	0	72.03	0	60
DLCL-0039	1	0	91.33	1	60
DLCL-0040	1	0	53.73	1	53.73
DLCL-0002	1	1	3.4	1	3.4
DLCL-0003	0	1	71.3	0	60
DLCL-0006	1	1	3.2	1	3.2
DLCL-0007	1	1	8.3	1	8.3
DLCL-0011	1	1	27.1	1	27.1
DLCL-0012	0	1	4.1	0	4.1
DLCL-0013	1	1	23.7	1	23.7
DLCL-0016	1	1	15.5	1	15.5
DLCL-0017	1	1	2.4	1	2.4
DLCL-0018	0	1	2.9	0	2.9
DLCL-0021	1	1	4.6	1	4.6
DLCL-0023	0	1	8.2	0	8.2
DLCL-0025	1	1	32.5	1	32.5
DLCL-0026	0	1	11.8	0	11.8
DLCL-0027	1	1	5.1	1	5.1
DLCL-0031	1	1	12.3	1	12.3
DLCL-0034	0	1	1.3	0	1.3
DLCL-0036	1	1	12.67	1	12.67
DLCL-0041	1	1	31.47	1	31.47
DLCL-0042	1	1	39.6	1	39.6
DLCL-0048	1	1	9.45	1	9.45
DLCL-0049	1	1	22.3	1	22.3

The cluster number in the above table is defined by the two main branches on the hierarchical clustering dendogram. A confusion matrix, shown below, gives information about the confusion between the cluster discovered here using the common cell-of-origin clones and the putative cell-of-origin cluster defined in Alizadeh et al.

		Alizadeh Cluster	
		C0	C1
Cluster	C0	19	0
	C1	0	21

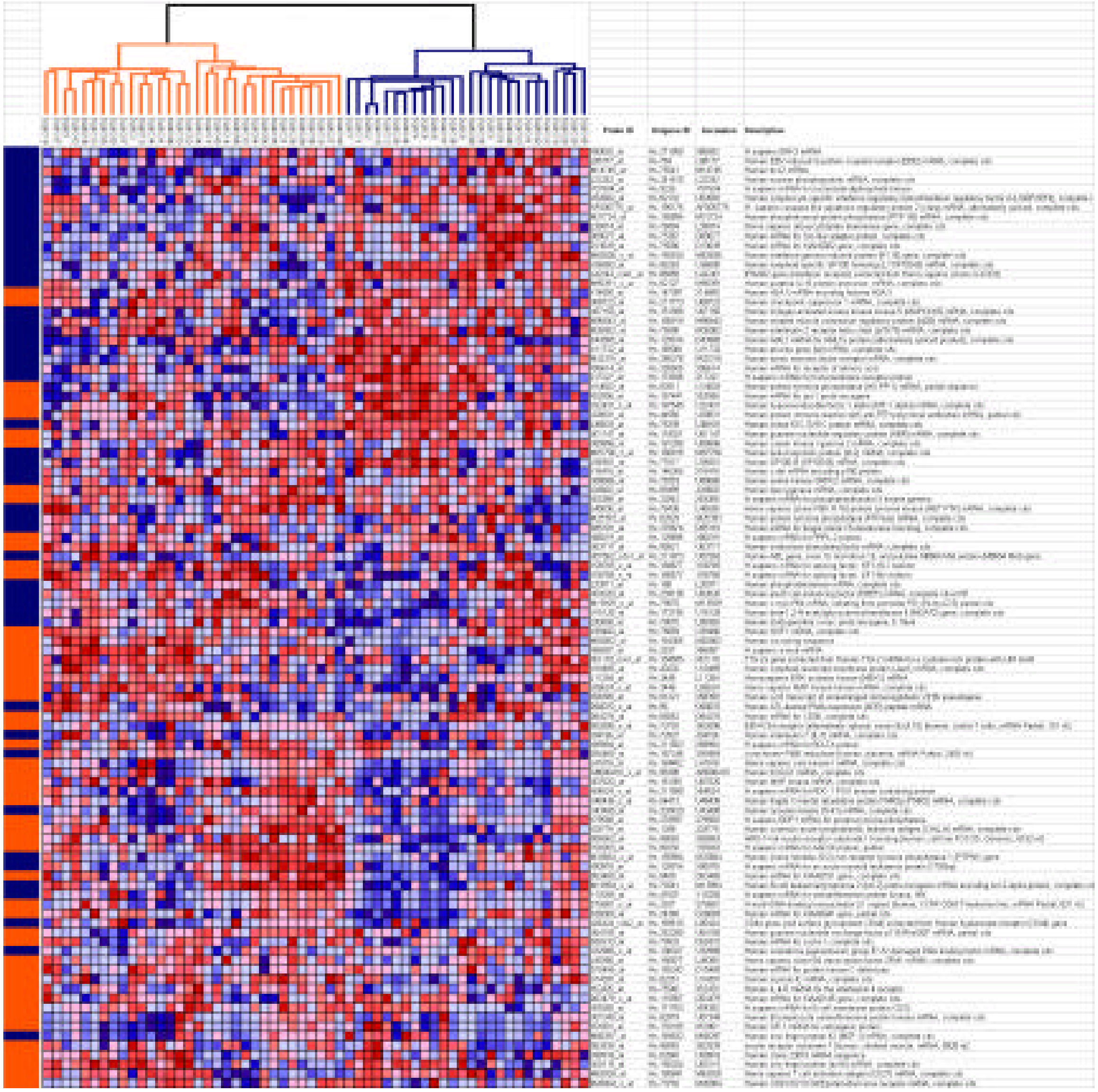
As can be seen by the above confusion matrix, the clustering on the cell-of-origin genes common to both data sets reproduced the original cell-of-origin clusters. The survival curve for the cell-of-origin clusters, using survival times that have been truncated to 60 months, is shown below.



The confusion matrix between the clusters and the observed survival is as follows:

		Observed Survival	
		Alive	Dead
Cluster	C0	13	6
	C1	5	16

Below are shown the results from hierarchical clustering of the cell-of-origin genes common to both sets using the data from our oligonucleotide arrays. The samples were represented by the expression levels of the 100 genes corresponding to the 90 common cell-of-origin Unigene clusters and were clustered using average linkage clustering with an uncentered correlation similarity metric. This data was first preprocessed using our standard preprocessing (thresholded to 10 minimum, 16000 maximum and filtered by max/min>3 and max-min>100) and then adjusted within the clustering software using the following sequence of actions: log transform, mean center genes, mean center arrays, mean center genes mean center arrays, normalize genes and normalize arrays. This figure is a version of figure 5B of our paper that includes the details of the hierarchical clustering (an expanded version of this figure can be downloaded from our website http://www-genome.wi.mit.edu/MPR/lymphoma).



The following table shows the list of genes with their accession numbers Unigene cluster numbers, the cluster number in the gene clustering, and whether the gene was in the Alizadeh et al activated B-like DLBCLs (0) versus GC-like DLBCLs (1) high expression level class.

Probe ID	Our Clustered Class	Alizadeh Clustered Class	Unigene ID	Accession	Description
X80692_at	0	0	Hs.271980	X80692	H.sapiens ERK3 mRNA
L08177_at	0	0	Hs.784	L08177	Human EBV induced G-protein coupled receptor (EBI2) mRNA, complete cds
M14745_at	0	0	Hs.79241	M14745	Human bcl-2 mRNA
L22342_at	0	0	Hs.241510	L22342	Human nuclear phosphoprotein mRNA, complete cds
Y07604_at	0	0	Hs.9235	Y07604	H.sapiens mRNA for nucleoside-diphosphate kinase
U52682_at	0	0	Hs.82132	U52682	Human lymphocyte specific interferon regulatory factor/interferon regulatory factor 4 (LSIRF/IRF4), complete cds
AF005775_at	0	0	Hs.195175	AF005775	H. sapiens caspase-like apoptosis regulatory protein 2

					(clarp) mRNA, alternatively spliced, complete cds.
M31724_at	0	0	Hs.155894	M31724	Human phosphotyrosyl-protein phosphatase (PTP-1B) mRNA, complete cds
L39874_at	0	0	Hs.76894	L39874	Homo sapiens deoxycytidylate deaminase gene, complete cds
D89077_at	0	0	Hs.75367	D89077	Human mRNA for Src-like adapter protein, complete cds
D13639_at	0	0	Hs.75586	D13639	Human mRNA for KIAK0002 gene, complete cds
M63838_s_at	0	0	Hs.155530	M63838	Human interferon-gamma induced protein (IFI 16) gene, complete cds
U36500_at	0	0	Hs.85283	U36500	Human lymphoid-specific SP100 homolog (LYSP100-B) mRNA, complete cds
L42243_cds1_at	0	0	Hs.86958	L42243	IFNAR2 gene (interferon receptor) extracted from Homo sapiens (clone Q-2OD3)
M90391_s_at	0	0	Hs.82127	M90391	Human putative IL-16 protein precursor, mRNA, complete cds
X14850_at	0	1	Hs.147097	X14850	Human H2A.X mRNA encoding histone H2A.X
U68723_at	0	1	Hs.211773	U68723	Human checkpoint suppressor 1 mRNA, complete cds.
U67156_at	0	0	Hs.151988	U67156	Human mitogen-activated kinase kinase kinase 5 (MAPKKK5) mRNA, complete cds
M96843_at	0	0	Hs.180919	M96843	Human striated muscle contraction regulatory protein (Id2B) mRNA, complete cds
M26062_at	0	0	Hs.75596	M26062	Human interleukin 2 receptor beta chain (p70-75) mRNA, complete cds
D43968_at	0	0	Hs.129914	D43968	Human AML1 mRNA for AML1b protein (alternatively spliced product), complete cds
U11732_at	0	0	Hs.169081	U11732	Human ets-like gene (tel) mRNA, complete cds
M32315_at	0	0	Hs.256278	M32315	Human tumor necrosis factor receptor mRNA, complete cds
X06614_at	0	0	Hs.250505	X06614	Human mRNA for receptor of retinoic acid
Z17227_at	0	0	Hs.173936	Z17227	H.sapiens mRNA for transmenbrane receptor protein
U14603_at	0	1	Hs.82911	U14603	Human protein-tyrosine phosphatase (HU-PP-1) mRNA, partial sequence
X52056_at	0	1	Hs.157441	X52056	Human mRNA for spi-1 proto-oncogene
U22431_s_at	0	1	Hs.197540	U22431	Human hypoxia-inducible factor 1 alpha (HIF-1 alpha) mRNA, complete cds
U28831_at	0	1	Hs.44566	U28831	Human protein immuno-reactive with anti-PTH polyclonal antibodies mRNA, partial cds
L36818_at	0	0	Hs.75339	L36818	Human (clone 51C-3) 51C protein mRNA, complete cds
U01147_at	0	1	Hs.118021	U01147	Human guanine nucleotide regulatory protein (ABR) mRNA, complete cds
U89896_at	0	1	Hs.181390	U89896	Human casein kinase I gamma 2 mRNA, complete cds
M97796_s_at	0	0	Hs.180919	M97796	Human helix-loop-helix protein (Id-2) mRNA, complete cds
U36501_at	0	0	Hs.77617	U36501	Human SP100-B (SP100-B) mRNA, complete cds
X16416_at	0	0	Hs.146355	X16416	Human c-abl mRNA encoding p150 protein
U88666_at	0	0	Hs.78353	U88666	Human serine kinase SRPK2 mRNA, complete cds
J03600_at	0	1	Hs.89499	J03600	Human lipoxygenase mRNA, complete cds
X83368_at	0	1	Hs.32942	X83368	H.sapiens mRNA for phosphatidylinositol 3 kinase gamma
L40636_at	0	0	Hs.78436	L40636	Homo sapiens (clone FBK III 16) protein tyrosine kinase (NET PTK) mRNA, complete cds
M25393_at	0	0	Hs.82829	M25393	Human protein tyrosine phosphatase (PTPase) mRNA,

					complete cds
D85181_at	0	0	Hs.278674	D85181	Human mRNA for fungal sterol-C5-desaturase homolog, complete cds
X86019_at	0	1	Hs.129695	X86019	H.sapiens mRNA for PRPL-2 protein
U63717_at	0	1	Hs.95821	U63717	Human osteoclast stimulating factor mRNA, complete cds
U07563_cds1_at	0	0	Hs.211973	U07563	Human ABL gene, exon 1b and intron 1b, and putative M8604 Met protein (M8604 Met) gene
Y08765_s_at	1	1	Hs.180677	Y08765	H.sapiens mRNA for splicing factor, SF1-HL1 isoform
Y08766_s_at	1	1	Hs.180677	Y08766	H.sapiens mRNA for splicing factor, SF1-Bo isoform
L20971_at	1	0	Hs.188	L20971	Human phosphodiesterase mRNA, complete cds
U02020_at	1	0	Hs.239138	U02020	Human pre-B cell enhancing factor (PBEF) mRNA, complete cds+C91
M13929_s_at	1	0	Hs.79070	M13929	Human c-myc-P64 mRNA, initiating from promoter P0, (HLmyc2.5) partial cds
U15128_at	1	0	Hs.172195	U15128	Human beta-1,2-N-acetylglucosaminyltransferase II (MGAT2) gene, complete cds
L00058_at	1	0	Hs.79070	L00058	Human (GH) germline c-myc proto-oncogene, 5 flank
U39400_at	1	1	Hs.75859	U39400	Human NOF1 mRNA, complete cds
M82882_at	1	1	Hs.154365	M82882	Human cis-acting sequence
X66087_at	1	1	Hs.2537	X66087	H.sapiens a-myb mRNA
X61118_rna1_at	1	1	Hs.184585	X61118	TTG-2a gene extracted from Human TTG-2 mRNA for a cysteine rich protein with LIM motif
U10485_at	1	1	Hs.40202	U10485	Human lymphoid-restricted membrane protein (Jaw1) mRNA, complete cds
L11284_at	1	1	Hs.3446	L11284	Homosapiens ERK activator kinase (MEK1) mRNA
L05624_s_at	1	1	Hs.3446	L05624	Homo sapiens MAP kinase kinase mRNA, complete cds
X58399_at	1	1	Hs.81221	X58399	Human L2-9 transcript of unrearranged immunoglobulin V(H)5 pseudogene.
D90070_s_at	1	0	Hs.96	D90070	Human ATL-derived PMA-responsive (APR) peptide mRNA
D84276_at	1	1	Hs.66052	D84276	Human mRNA for CD38, complete cds
S62696_s_at	1	1	Hs.73792	S62696	EBV/C3d receptor {alternatively spliced, exons 8a,9,10} [human, Jurkat T cells, mRNA Partial, 151 nt]
J04156_at	1	0	Hs.72927	J04156	Human interleukin 7 (IL-7) mRNA, complete cds
X89984_at	1	1	Hs.211563	X89984	H.sapiens mRNA for BCL7A protein
S90469_at	1	0	Hs.167246	S90469	cytochrome P450 reductase [human, placenta, mRNA Partial, 2403 nt]
L41816_at	1	1	Hs.184402	L41816	Homo sapiens cam kinase I mRNA, complete cds
AB000410_s_at	1	1	Hs.96398	AB000410	Human hOGG1 mRNA, complete cds
U07620_at	1	1	Hs.151051	U07620	Human MAP kinase mRNA, complete cds
X64624_s_at	1	1	Hs.211588	X64624	H.sapiens mRNA for RDC-1 POU domain containing protein
U48436_s_at	1	1	Hs.54472	U48436	Human fragile X mental retardation protein FMR2p (FMR2) mRNA, complete cds
U43408_at	1	0	Hs.203420	U43408	Human tyrosine kinase (Tnk1) mRNA, complete cds
X79568_at	1	1	Hs.278597	X79568	H.sapiens BDP1 mRNA for protein-tyrosine-phosphatase
J03779_at	1	1	Hs.1298	J03779	Human common acute lymphoblastic leukemia antigen (CALLA) mRNA, complete cds
S85963_at	1	1	Hs.96063	S85963	hIRS-1=rat insulin receptor substrate-1 homolog [human, cell line FOCUS, Genomic, 6152 nt]

Y08263_at	1	1	Hs.89230	Y08263	H.sapiens mRNA for AAD14 protein, partial
M33684_s_at	1	0	Hs.155894	M33684	Human (clone lambda-10-2) non-receptor tyrosine phosphatase 1 (PTPN1) gene
X90978_at	1	0	Hs.129914	X90978	H.sapiens mRNA for an acute myeloid leukaemia protein (1793bp)
D63485_at	1	1	Hs.9408	D63485	Human mRNA for KIAA0151 gene, complete cds
M13994_s_at	1	0	Hs.79241	M13994	Human B-cell leukemia/lymphoma 2 (bcl-2) proto-oncogene mRNA encoding bcl-2-alpha protein, complete cds
Y10256_at	1	0	Hs.47007	Y10256	H.sapiens mRNA for serine/threonine protein kinase, NIK
S75881_s_at	1	1	Hs.2537	S75881	A-myb=DNA-binding transactivator {3 region} [human, CCRF-CEM T-leukemia line, mRNA Partial, 831 nt]
D26069_at	1	1	Hs.24340	D26069	Human mRNA for KIAA0041 gene, partial cds
L05424_cds2_at	1	0	Hs.169610	L05424	CD44 gene (cell surface glycoprotein CD44) extracted from Human hyaluronate receptor (CD44) gene
U64105_at	1	1	Hs.252280	U64105	Human guanine nucleotide exchange factor p115-RhoGEF mRNA, partial cds
D50310_at	1	1	Hs.79933	D50310	Human mRNA for cyclin I, complete cds
U32986_s_at	1	0	Hs.108327	U32986	Human xeroderma pigmentosum group E UV-damaged DNA binding factor mRNA, complete cds
L49380_at	1	1	Hs.180677	L49380	Homo sapiens clone B4 transcription factor ZFM1 mRNA, complete cds
D10495_at	1	1	Hs.155342	D10495	Human mRNA for protein kinase C delta-type
U14391_at	1	1	Hs.82251	U14391	Human myosin-IC mRNA, complete cds
X52425_at	1	1	Hs.75545	X52425	Human IL-4-R mRNA for the interleukin 4 receptor
D63479_s_at	1	1	Hs.115907	D63479	Human mRNA for KIAA0145 gene, complete cds
X59350_at	1	1	Hs.171763	X59350	H.sapiens mRNA for B cell membrane protein CD22
U07349_at	1	1	Hs.82979	U07349	Human B lymphocyte serine/threonine protein kinase mRNA, complete cds
X51801_at	1	1	Hs.170195	X51801	Human OP-1 mRNA for osteogenic protein
M58297_at	1	0	Hs.169832	M58297	Human zinc finger protein 42 (MZF-1) mRNA, complete cds
S62539_at	1	1	Hs.96063	S62539	insulin receptor substrate-1 [human, skeletal muscle, mRNA, 5828 nt]
U90916_at	1	1	Hs.82845	U90916	Human clone 23815 mRNA sequence
U00115_at	1	1	Hs.155024	U00115	Human zinc-finger protein (bcl-6) mRNA, complete cds
M63928_at	1	1	Hs.180841	M63928	Homo sapiens T cell activation antigen (CD27) mRNA, complete cds
M26004_s_at	1	1	Hs.73792	M26004	Human CR2/CD21/C3d/Epstein-Barr virus receptor mRNA, complete cds

In our series, the top branch of the hierarchical tree includes 72.7% (32/44) of the Alizadeh et al.-defined GC marker genes whereas the bottom branch includes 71.4% (40/56) of the similarly defined PB marker genes (p = .00001, Chi-squared test). The fact that the GC versus PB marker distinction is replicated fairly well in our data set suggests that the sample clustering should also replicate the GC B-like versus activated B-like DLBCL distinction. The results of clustering our samples using the cell-of-origin genes, when using the two main branches of the dendogram to define the clusters, is as follows:

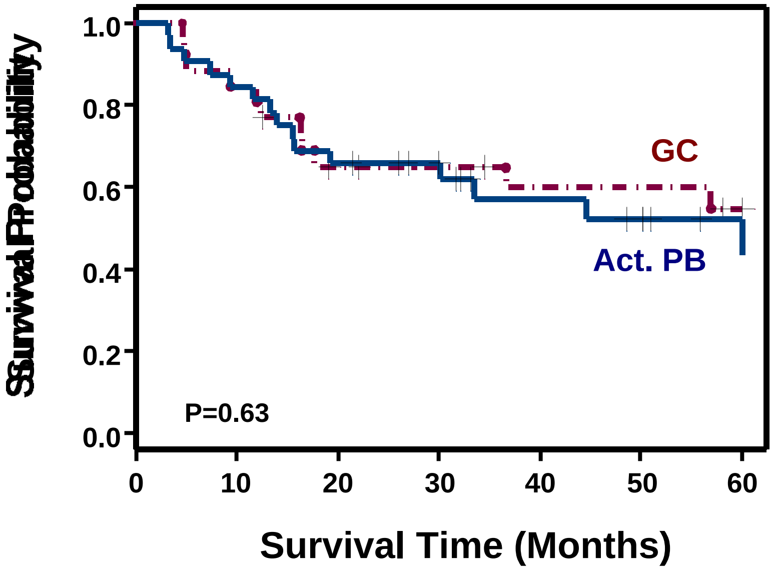
Sample	Predicted Class (cell-of-origin)	Outcome	Survival (months)	Truncated Survival
DLBC1	1	0	72.9	60
DLBC2	0	0	143.1	60
DLBC3	1	0	144.2	60
DLBC4	0	0	61	60
DLBC5	0	0	86.5	60
DLBC6	0	0	84.2	60
DLBC7	0	0	112.5	60
DLBC8	0	0	133.2	60
DLBC9	0	0	22.1	22.1
DLBC10	0	0	182.4	60
DLBC11	0	0	66.4	60
DLBC12	0	0	146.8	60
DLBC13	1	0	62.9	60
DLBC14	1	0	50.9	50.9
DLBC15	1	0	78.5	60
DLBC16	1	0	48.6	48.6
DLBC17	1	0	55.9	55.9
DLBC18	0	0	12.6	12.6
DLBC19	1	0	50.2	50.2
DLBC20	0	0	58	58
DLBC21	1	0	66.4	60
DLBC22	0	0	65.7	60
DLBC23	1	0	50.2	50.2
DLBC24	1	0	26.9	26.9
DLBC25	0	0	34.4	34.4
DLBC26	1	0	26	26
DLBC27	1	0	30	30
DLBC28	1	0	31.7	31.7
DLBC29	1	0	32.2	32.2
DLBC30	0	0	19.2	19.2
DLBC31	1	0	33.1	33.1
DLBC32	1	0	21.4	21.4
DLBC33	1	1	15.7	15.7
DLBC34	1	1	11.6	11.6
DLBC35	1	1	3.4	3.4
DLBC36	0	1	36.6	36.6
DLBC37	0	1	5	5
DLBC38	0	1	9.5	9.5
DLBC39	1	1	3.2	3.2
DLBC40	1	1	4.9	4.9
DLBC41	0	1	12	12
DLBC42	0	1	4.9	4.9
DLBC43	1	1	60.4	60
DLBC44	0	1	16.3	16.3

DLBC45	0	1	16.4	16.4
DLBC46	1	1	9.5	9.5
DLBC47	1	1	15.6	15.6
DLBC48	0	1	17.8	17.8
DLBC49	0	1	56.9	56.9
DLBC50	1	1	13.3	13.3
DLBC51	0	1	12.3	12.3
DLBC52	1	1	44.6	44.6
DLBC53	0	1	4.6	4.6
DLBC54	1	1	7.5	7.5
DLBC55	1	1	19.3	19.3
DLBC56	1	1	30.1	30.1
DLBC57	1	1	33.6	33.6
DLBC58	1	1	13.9	13.9

The cluster number in the above table is defined by the two main branches on the hierarchical clustering dendogram. The confusion matrix between the clusters and the observed survival is as follows:

		Observed Survival	
		Alive	Dead
Cluster	C0	15	11
	C1	16	15

The corresponding survival curve for the cell-of-origin clusters, using survival times that have been truncated to 60 months, is shown below.



The GC B-like versus Act. PB-like distinction was not significantly correlated with patient outcome in our DLBCL series (Chisq= 0.2 on 1 degrees of freedom, p= 0.631). This observation suggests that although the signature genes may reflect cell of origin, they do not explain a significant portion of the clinical variability seen in this DLBCL data set. One possible explanation may be the additional heterogeneity within each of the two major subgroups defined by the cell-of-origin signature in our larger series of samples as shown in the pink-o-gram above.

Validation of Our Outcome Predictor

We also asked whether we could find support for ***our*** outcome predictor in the expression data of Alizadeh et al[13]. We mapped the oligonucleotide array accession numbers for the thirteen genes of our outcome predictor to the Unigene cluster numbers. The mapping is as follows:

Affymetrix Identifier	Unigene ID	Description
U43519_at	Hs.159291	DRP2 Dystrophin related protein 2
Y09836_at	Hs.82503	3'UTR of unknown protein
HG2314-HT2410_at		Uncharacterized
Z15114_at	Hs.2890	PRKCG Protein kinase C, gamma
U12767_at	Hs.80561	Mitogen induced nuclear orphan receptor (MINOR)
X77307_at	Hs.2507	5-HYDROXYTRYPTAMINE 2B RECEPTOR
U83908_at	Hs.100407	Nuclear antigen H731 mRNA
M99435_at	Hs.28935	TRANSDUCIN-LIKE ENHANCER PROTEIN 1
L20971_at	Hs.188	PDE4B Phosphodiesterase 4B, cAMP-specific
AC002450_at		Uncharacterized
M18255_cds2_s_at	Hs.77202	PRKACB gene (protein kinase C-beta-1)
U09550_at	Hs.1154	Oviductal glycoprotein mRNA
U38864_at	Hs.108139	Zinc-finger protein C2H2-150

In Silico model validation was then performed by identifying genes from the thirteen-gene microarray-based outcome predictor (listed above) that were represented on the lymphochip. We mapped the lymphochip clone IMAGE numbers to GenBank accession numbers (using the list http://llmpp.nih.gov/lymphoma/data/clones.txt) and then mapped the accession numbers to Unigene cluster numbers. Three of the eleven Unigene cluster numbers representing our 13-gene model were represented on the lymphochip. These three Unigene cluster numbers (for the genes MINOR / NOR-1, PDE4B and PKCβ) were represented by ten clones as shown in the table below.

Stanford Clone Number	Copies	Well Expressed Copies	Unigene ID	Accession Number	Description
1184411	1	1	Hs.80561	AA648528	MINOR=mitogen induced nuclear orphan receptor=NOR-1=Nur77 orphan nuclear receptor family member
323151	1	0	Hs.80561	W42606	MINOR=mitogen induced nuclear orphan receptor=NOR-1=Nur77 orphan nuclear receptor family member
190468	1	0	Hs.80561	H37761	MINOR=mitogen induced nuclear orphan receptor=NOR-1=Nur77 orphan nuclear receptor family member
377708	2	1	Hs.188	AA056219	3' 5'-cyclic AMP phosphodiesterase=rolipram-sensitive cAMP-specific phosphodiesterase (PDE4B)
685194	6	6	Hs.77202	AA243358	Protein kinase C, beta 2
1308435	1	1	Hs.77202	AA737573	Protein kinase C, beta 2
1368281	1	1	Hs.77202	AA837054	Protein kinase C, beta 1
1371673	3	3	Hs.77202	AA826104	Protein kinase C, beta 2
284459	1	0	Hs.77202	N52338	Protein kinase C, beta 2

753923	1	1	Hs.77202	AA479102	Protein kinase C, beta 1

The raw lymphochip data from the 40 DLBCL specimens and the associated outcome information was obtained from the public website (http://llmpp.nih.gov/lymphoma). Some of these clones had multiple copies (as shown in the “copies” column in the above table) on the lymphochip where only a fraction might be “well-expressed”. Genes were considered “well-expressed” using the Alizadeh et al. metric where all non-flagged array elements that had fluorescent intensity in each channel that was greater than 1.4 times the local background. Predictors using single genes (PKCβ, PDE4B, MINOR/NOR-1) were constructed by finding the boundary halfway between the classes ($\boldsymbol{b}_{\boldsymbol{x}}$ = ($\mu_{\boldsymbol{class0}}$ + $\mu_{\boldsymbol{class1}}$)/2) in the data set and predicting the unknown sample according to its gene expression value with respect to that boundary. This method is equivalent to performing weighted voting with only 1 gene. When there were multiple copies of clones, we used the average of the expression values for the clones.

The genes NOR-1 and PDE4B were represented by single well-expressed markers in the Alizadeh et al. data set so we evaluated these using normalized values obtained from the file figure1.cdt on the public website. Building a one gene predictor using the well-expressed MINOR / NOR-1 clone (clone number 1184411) produced the results shown in the table below.

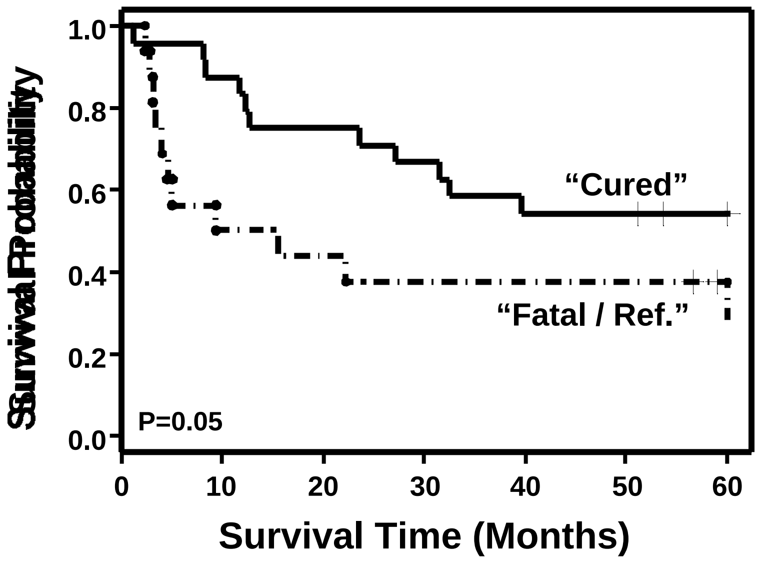
Sample Identifier	Predicted Class	True Class	Error?	Overall Survival	Truncated Survival	Stanford Class
DLCL-0001	0	0		77.40	60	0
DLCL-0004	0	0		69.60	60	0
DLCL-0005	0	0		51.20	51.2	1
DLCL-0008	0	0		102.40	60	0
DLCL-0009	0	0		89.80	60	0
DLCL-0010	1	0	*	88.10	60	0
DLCL-0014	1	0	*	59.00	59	1
DLCL-0015	1	0	*	56.60	56.6	0
DLCL-0020	0	0		80.40	60	0
DLCL-0024	0	0		129.90	60	0
DLCL-0028	0	0		90.20	60	1
DLCL-0029	0	0		83.80	60	0
DLCL-0030	1	0	*	71.30	60	0
DLCL-0032	0	0		69.10	60	0
DLCL-0033	1	0	*	68.80	60	0
DLCL-0037	0	0		72.03	60	0
DLCL-0039	0	0		91.33	60	1
DLCL-0040	0	0		53.73	53.73	1
DLCL-0002	1	1		3.40	3.4	1
DLCL-0003	1	1		71.30	60	1
DLCL-0006	1	1		3.20	3.2	1
DLCL-0007	0	1	*	8.30	8.3	1
DLCL-0011	0	1	*	27.10	27.1	1
DLCL-0012	1	1		4.10	4.1	0
DLCL-0013	0	1	*	23.70	23.7	1
DLCL-0016	1	1		15.50	15.5	1

DLCL-0017	1	1		2.40	2.4	1
DLCL-0018	1	1		2.90	2.9	0
DLCL-0021	1	1		4.60	4.6	1
DLCL-0023	0	1	*	8.20	8.2	0
DLCL-0025	0	1	*	32.50	32.5	1
DLCL-0026	0	1	*	11.80	11.8	0
DLCL-0027	1	1		5.10	5.1	1
DLCL-0031	0	1	*	12.30	12.3	1
DLCL-0034	0	1	*	1.30	1.3	0
DLCL-0036;OCT	0	1	*	12.67	12.67	1
DLCL-0041	0	1	*	31.47	31.47	1
DLCL-0042	0	1	*	39.60	39.6	1
DLCL-0048	1	1		9.45	9.45	1
DLCL-0049	1	1		22.30	22.3	1

The outcome prediction results using MINOR shown in the above table are summarized in the confusion matrix below.

		True	
		Alive	Dead
Predicted	Alive	13	11
	Dead	5	11

The Kaplan-Meier survival plot for the gene MINOR/NOR-1 in the lymphochip data is shown below.



The p-value for whether there is a difference between the two survival curves resulting from predicting outcome in the Alizadeh et al. data using MINOR is equal to 0.05.

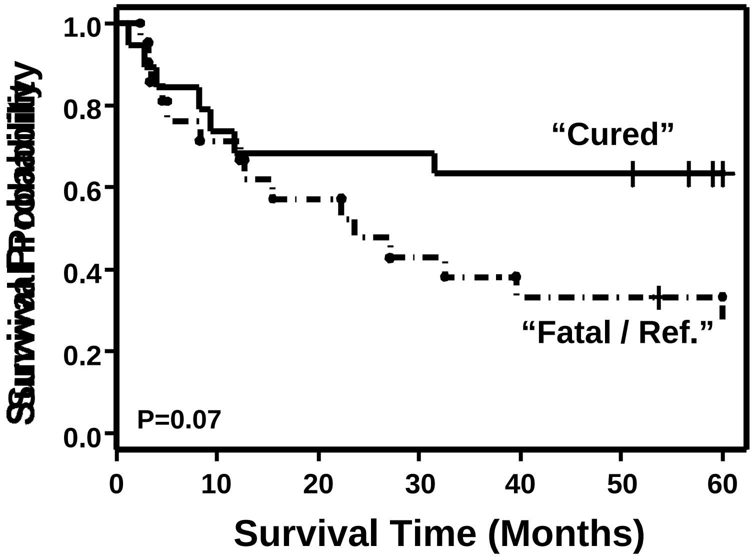
Similarly, we built a single gene predictor using PDE4B (clone 377708) from the Alizadeh et al data set, which produced the following set of results.

Sample Identifier	Predicted Class	True Class	Error?	Overall Survival	Truncated Survival	Stanford Class
DLCL-0001	0	0		77.4	60	0
DLCL-0004	1	0	*	69.6	60	0
DLCL-0005	0	0		51.2	51.2	1
DLCL-0008	0	0		102.4	60	0
DLCL-0009	0	0		89.8	60	0
DLCL-0010	0	0		88.1	60	0
DLCL-0014	0	0		59	59	1
DLCL-0015	0	0		56.6	56.6	0
DLCL-0020	1	0	*	80.4	60	0
DLCL-0024	0	0		129.9	60	0
DLCL-0028	1	0	*	90.2	60	1
DLCL-0029	0	0		83.8	60	0
DLCL-0030	1	0	*	71.3	60	0
DLCL-0032	0	0		69.1	60	0
DLCL-0033	0	0		68.8	60	0
DLCL-0037	0	0		72.03	60	0
DLCL-0039	1	0	*	91.33	60	1
DLCL-0040	1	0	*	53.73	53.73	1
DLCL-0002	1	1		3.4	3.4	1
DLCL-0003	1	1		71.3	60	1
DLCL-0006	1	1		3.2	3.2	1
DLCL-0007	1	1		8.3	8.3	1
DLCL-0011	1	1		27.1	27.1	1
DLCL-0012	0	1	*	4.1	4.1	0
DLCL-0013	1	1		23.7	23.7	1
DLCL-0016	1	1		15.5	15.5	1
DLCL-0017	1	1		2.4	2.4	1
DLCL-0018	0	1	*	2.9	2.9	0
DLCL-0021	1	1		4.6	4.6	1
DLCL-0023	0	1	*	8.2	8.2	0
DLCL-0025	1	1		32.5	32.5	1
DLCL-0026	0	1	*	11.8	11.8	0
DLCL-0027	1	1		5.1	5.1	1
DLCL-0031	1	1		12.3	12.3	1
DLCL-0034	0	1	*	1.3	1.3	0
DLCL-0036;OCT	1	1		12.67	12.67	1
DLCL-0041	0	1	*	31.47	31.47	1
DLCL-0042	1	1		39.6	39.6	1
DLCL-0048	0	1	*	9.45	9.45	1
DLCL-0049	1	1		22.3	22.3	1

The outcome prediction results using PDE4B shown in the above table are summarized in the confusion matrix below.

		Observed	
		Alive	Dead
Predicted	Alive	12	7
	Dead	6	15

The Kaplan-Meier survival plot of PDE4B for predicting outcome in the lymphochip data is shown below.



The log-rank p-value for whether there is a difference between the two survival curves resulting from predicting outcome in the Alizadeh et al. data using PDE4B is equal to 0.07.

Multiple PKCβ cDNAs are included on the lymphochip. RAT2 values for all the PKCβ clones were obtained from the raw data files. The data were pre-processed by setting minimum values to 0 and normalizing arrays to a mean value of 0 and variance of 1. In our 13-gene model, PKCβ was specifically associated with outcome, but the clones in the Alizadeh et al. dataset gave discordant expression results in the DLBCL patients, perhaps reflecting varying degrees of specificity for the β isoforms of PKC. Therefore we analyzed the PKCβ clones individually. We considered only clones that were determined to be "well-expressed" by the Alizadeh et al. metric (which eliminated clone number 284459). We also eliminated from consideration two PKC beta clones (1368281 and 1371673) that had partial sequence matches with PKC gamma (340/430 and 78/99 respectively) and therefore may be cross hybridizing with PKC gamma. The remaining three clones show varying degrees of ability to predict outcome. Two (1308435 and 685194) have Kaplan-Meier log-rank p-values less than 0.05 and one (753923) has a p-value greater than 0.05. We show below the single gene prediction results for PKC beta clones 1308435, 685194, and 753923.

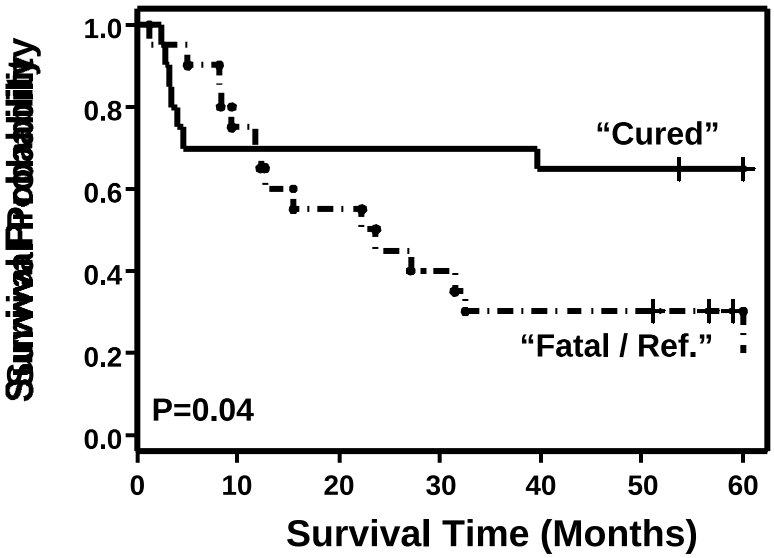
One of the PKCβ clones on the lymphochip was clone number 1308435 (accession number AA737573) which had a single copy. Building a single gene predictor using this version of PKCβ from the Alizadeh et al. data set produced the following set of results.

Sample Identifier	Predicted Class	True Class	Error?	Overall Survival	Truncated Survival	Alizadeh et al. Class
DLCL-0001	0	0		77.4	60	0
DLCL-0004	0	0		69.6	60	0
DLCL-0005	1	0	*	51.2	51.2	1
DLCL-0008	0	0		102.4	60	0
DLCL-0009	0	0		89.8	60	0
DLCL-0010	0	0		88.1	60	0
DLCL-0014	1	0	*	59	59	1
DLCL-0015	1	0	*	56.6	56.6	0
DLCL-0020	0	0		80.4	60	0
DLCL-0024	1	0	*	129.9	60	0
DLCL-0029	0	0		83.8	60	0
DLCL-0030	0	0		71.3	60	0
DLCL-0032	1	0	*	69.1	60	0
DLCL-0033	0	0		68.8	60	0
DLCL-0002	0	1	*	3.4	3.4	1
DLCL-0003	1	1		71.3	60	0
DLCL-0006	0	1	*	3.2	3.2	1
DLCL-0007	1	1		8.3	8.3	1
DLCL-0011	1	1		27.1	27.1	1
DLCL-0012	0	1	*	4.1	4.1	0
DLCL-0013	1	1		23.7	23.7	1
DLCL-0016	1	1		15.5	15.5	1
DLCL-0018	0	1	*	2.9	2.9	0
DLCL-0021	0	1	*	4.6	4.6	1
DLCL-0023	1	1		8.2	8.2	0
DLCL-0025	1	1		32.5	32.5	1
DLCL-0026	1	1		11.8	11.8	0
DLCL-0027	1	1		5.1	5.1	1
DLCL-0031	1	1		12.3	12.3	1
DLCL-0034	1	1		1.3	1.3	0
DLCL-0042	0	1	*	39.6	39.6	1
DLCL-0048	1	1		9.45	9.45	1
DLCL-0049	1	1		22.3	22.3	1
DLCL-0017	0	1	*	2.4	2.4	1
DLCL-0036;OCT	1	1		12.67	12.67	1
DLCL-0037	0	0		72.03	60	0
DLCL-0039	0	0		91.33	60	1
DLCL-0040	0	0		53.73	53.73	1
DLCL-0041	1	1		31.47	31.47	1
DLCL-0028	0	0		90.2	60	1

The outcome prediction results using PKCβ clone number 1308435 that are shown in the above table are summarized in the confusion matrix below.

		True	
		Alive	Dead
Predicted	Alive	13	7
	Dead	5	15

The Kaplan-Meier survival plot for PKCβ clone number 1308435 is shown below.



The p-value for whether there is a difference between the two survival curves resulting from predicting outcome in the Alizadeh et al. data using PKCβ clone number 1308435 is equal to 0.04.

Another one of the PKCβ clones on the lymphochip was clone number 685194 (accession number AA243358) which had six copies on the lymphochip. We took the mean of the expression values for these six copies of the clone as the expression value for this clone and built a single predictor for it. The results from building a single gene predictor of outcome using the mean PKCβ clone 685194 are shown in the following table.

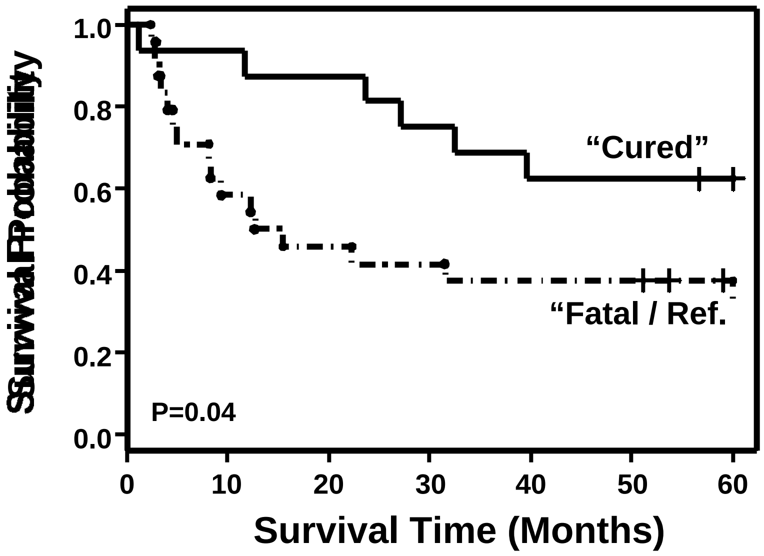
Sample Identifier	Predicted Class	True Class	Error?	Overall Survival	Truncated Survival	Alizadeh et al. Class
DLCL-0001	1	0	*	77.4	60	0
DLCL-0004	0	0		69.6	60	0
DLCL-0005	1	0	*	51.2	51.2	1
DLCL-0008	0	0		102.4	60	0
DLCL-0009	0	0		89.8	60	0
DLCL-0010	1	0	*	88.1	60	0
DLCL-0014	1	0	*	59	59	1
DLCL-0015	0	0		56.6	56.6	0
DLCL-0020	1	0	*	80.4	60	0
DLCL-0024	1	0	*	129.9	60	0
DLCL-0028	1	0	*	90.2	60	1
DLCL-0029	0	0		83.8	60	0
DLCL-0030	0	0		71.3	60	0
DLCL-0032	0	0		69.1	60	0
DLCL-0033	0	0		68.8	60	0

DLCL-0037	0	0		72.03	60	0
DLCL-0039	0	0		91.33	60	1
DLCL-0040	1	0	*	53.73	53.73	1
DLCL-0002	1	1		3.4	3.4	1
DLCL-0003	1	1		71.3	60	0
DLCL-0006	1	1		3.2	3.2	1
DLCL-0007	1	1		8.3	8.3	1
DLCL-0011	0	1	*	27.1	27.1	1
DLCL-0012	1	1		4.1	4.1	0
DLCL-0013	0	1	*	23.7	23.7	1
DLCL-0016	1	1		15.5	15.5	1
DLCL-0017	1	1		2.4	2.4	1
DLCL-0018	1	1		2.9	2.9	0
DLCL-0021	1	1		4.6	4.6	1
DLCL-0023	1	1		8.2	8.2	0
DLCL-0025	0	1	*	32.5	32.5	1
DLCL-0026	0	1	*	11.8	11.8	0
DLCL-0027	1	1		5.1	5.1	1
DLCL-0031	1	1		12.3	12.3	1
DLCL-0034	0	1	*	1.3	1.3	0
DLCL-0036;OCT	1	1		12.67	12.67	1
DLCL-0041	1	1		31.47	31.47	1
DLCL-0042	0	1	*	39.6	39.6	1
DLCL-0048	1	1		9.45	9.45	1
DLCL-0049	1	1		22.3	22.3	1

The outcome prediction results using PKCβ clone number 685194 that are shown in the above table are summarized in the confusion matrix below.

		True	
		Alive	Dead
Predicted	Alive	10	6
	Dead	8	16

The Kaplan-Meier survival plot for PKCβ clone number 685194 is shown below.



The p-value for whether there is a difference between the two survival curves resulting from predicting outcome in the Alizadeh et al. data using PKCβ clone number 685194 is equal to 0.04.

The last of the PKCβ clones on the lymphochip was clone number 753923 (accession number AA479102) which had a single copy. Building a single gene predictor using this version of PKCβ from the Alizadeh et al. data set produced the following set of results.

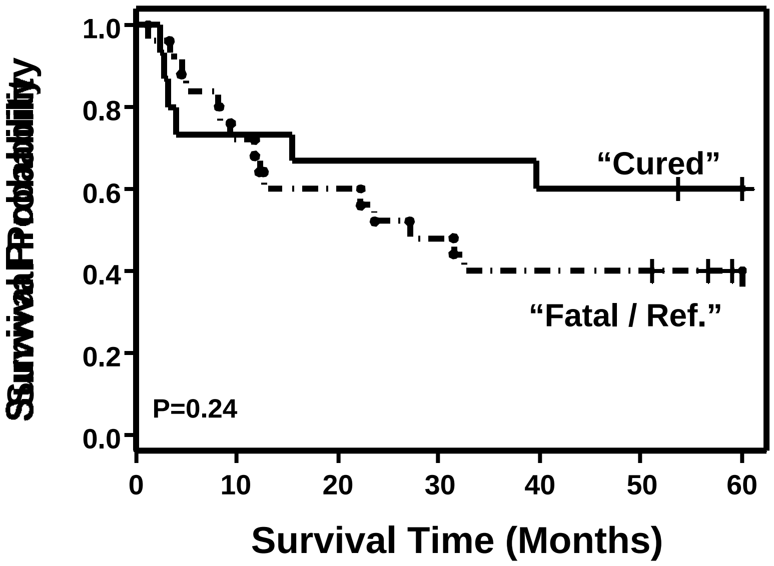
Sample Identifier	Predicted Class	True Class	Error?	Overall Survival	Truncated Survival	Alizadeh et al. Class
DLCL-0001	0	0		77.4	60	0
DLCL-0004	0	0		69.6	60	0
DLCL-0005	1	0	*	51.2	51.2	1
DLCL-0008	1	0	*	102.4	60	0
DLCL-0009	0	0		89.8	60	0
DLCL-0010	1	0	*	88.1	60	0
DLCL-0014	1	0	*	59	59	1
DLCL-0015	1	0	*	56.6	56.6	0
DLCL-0020	0	0		80.4	60	0
DLCL-0024	1	0	*	129.9	60	0
DLCL-0028	1	0	*	90.2	60	1
DLCL-0029	0	0		83.8	60	0
DLCL-0030	1	0	*	71.3	60	0

DLCL-0032	0	0		69.1	60	0
DLCL-0033	1	0	*	68.8	60	0
DLCL-0037	0	0		72.03	60	0
DLCL-0039	0	0		91.33	60	1
DLCL-0040	0	0		53.73	53.73	1
DLCL-0002	1	1		3.4	3.4	1
DLCL-0003	1	1		71.3	60	0
DLCL-0006	0	1	*	3.2	3.2	1
DLCL-0007	1	1		8.3	8.3	1
DLCL-0011	1	1		27.1	27.1	1
DLCL-0012	0	1	*	4.1	4.1	0
DLCL-0013	1	1		23.7	23.7	1
DLCL-0016	0	1	*	15.5	15.5	1
DLCL-0017	0	1	*	2.4	2.4	1
DLCL-0018	0	1	*	2.9	2.9	0
DLCL-0021	1	1		4.6	4.6	1
DLCL-0023	1	1		8.2	8.2	0
DLCL-0025	1	1		32.5	32.5	1
DLCL-0026	1	1		11.8	11.8	0
DLCL-0027	1	1		5.1	5.1	1
DLCL-0031	1	1		12.3	12.3	1
DLCL-0034	1	1		1.3	1.3	0
DLCL-0036;OCT	1	1		12.67	12.67	1
DLCL-0041	1	1		31.47	31.47	1
DLCL-0042	0	1	*	39.6	39.6	1
DLCL-0048	1	1		9.45	9.45	1
DLCL-0049	1	1		22.3	22.3	1

The outcome prediction results using PKCβ clone number 753923 that are shown in the above table are summarized in the confusion matrix below.

		Observed	
		Alive	Dead
Predicted	Alive	9	6
	Dead	9	16

The Kaplan-Meier survival plot for PKCβ clone number 753923 is shown below.



The p-value for whether there is a difference between the two survival curves resulting from predicting outcome in the Alizadeh et al. data using PKCβ clone number 753923 is equal to 0.24.

Immunohistochemical Staining for PKC Beta

The potential extension of this outcome prediction approach to the clinical setting was further explored using immunohistochemical detection methods. For this purpose, we created a tissue array containing the study DLBCLs for which formalin-fixed paraffin-embedded tumor tissue was available (n=21). PKCβ protein expression was pursued because of the commercial availability of a PKCβ monoclonal antibody known to function in immunohistochemistry assays (see section Immunohistochemical Staining for a detailed discussion of the procedure used). The intensity of staining on each core was graded from 0 (no staining) to 3 (maximal staining), and an average staining intensity (the mean of all five cores) was generated for each tumor. The p-value for the association between PKCβ immunostaining intensities and the array-based transcript levels was determined by using median to divide measured intensities into two levels and applying the Fisher exact test to evaluate the degree of association between the quantized measurements. The following table shows the data for the immunostaining.

Sample	True Class	Survival (months)	M18255_cds2 (PKC beta I)	M18255 class	PKCβ IHX	PKCβ IHC Class
DLBC13	0	62.9	261	0	0	0
DLBC14	0	50.9	424	1	0.2	0
DLBC15	0	78.5	208	0	0	0

DLBC16	0	48.6	178	0	0.6	1
DLBC19	0	50.2	359	1	0.4	1
DLBC21	0	66.4	228	0	0	0
DLBC23	0	50.2	320	1	0.2	0
DLBC25	0	34.4	469	1	0	0
DLBC29	0	32.2	214	0	0	0
DLBC30	0	19.2	94	0	0	0
DLBC31	0	33.1	202	0	0	0
DLBC40	1	4.9	539	1	1.5	1
DLBC43	1	60.4	20	0	0	0
DLBC49	1	56.9	318	1	0.2	0
DLBC51	1	12.3	761	1	2	1
DLBC52	1	44.6	964	1	0.6	1
DLBC53	1	4.6	296	0	0	0
DLBC54	1	7.5	358	1	2	1
DLBC55	1	19.3	1394	1	0.6	1
DLBC56	1	30.1	232	0	0.4	1
DLBC57	1	33.6	159	0	0	0
		Median	296		0.2	

The following is a confusion matrix between the two classes defined by the PKCβ immunostaining and the expression levels for the microarray PKCβ probe M18255_cds2.

		PKCB Microarray	
		0	1
PKCB IHC	0	9	4
	1	2	6

The following is a confusion matrix between the two classes defined by the PKCβ immunostaining and the observed “cured” versus “fatal/refractory” classes for the outcome data.

		Observed Outcome	
		0	1
PKCB IHC	0	9	4
	1	2	6

The correlation between outcome and the PKCβ immunostaining as measured by a standard two-sample t-test is 0.03. The Fisher’s exact test correlation between the median defined PKCβ immunostaining class and the PKCβ probe M18255_cds2 median defined class is equal to 0.08. The Fisher’s exact test correlation between the median defined PKCβ immunostaining class and the outcome is also equal to 0.08.

References

1. M. Shipp, N. Harris and P. Mauch. The non-Hodgkin's lymphomas, in: DeVita, V.T., Hellman, S., and Rosenberg, S.A. (eds.) Cancer Principles & Practices of Oncology, Philadelphia, PA: J.B. Lippincott Company: 2165-2220 (1997).
2. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmirovsky, E. Lander, T. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. (USA) **96**, 2907-2912 (1999).
3. T. R. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286**, 531-537 (1999).
4. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. (USA) **95** 14863-14868 (1998).
5. D. K. Slonim, P. Tamayo, J. P. Mesirov, T. R. Golub, and E. S. Lander. Class prediction and discovery using gene expression data. Procs. of the Fourth Annual International Conference on Computational Molecular Biology, Tokyo, Japan April 8 - 11, p263-272, 2000. http://www.genome.wi.mit.edu/MPR/publications/cancer_class_preprint_version.rtf
6. C. J. Huberty, ***Applied Discriminant Analysis***, John Wiley and Sons Inc. (1994).
7. M. J. Kearns and U. V. Vazirani, "An Introduction to Computational Learning Theory", MIT Press. 1997.
8. Dasarathy V.B. (ed), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE computer society press, Los Alamitos, Calif., December 1991. ISBN: 0818689307.
9. Mukherjee S. et al. Support vector machine classification of microarray data. CBCL Paper #182/AI Memo #1676, Massachusetts Institute of Technology, Cambridge, MA, December 1999. http://www.ai.mit.edu/projects/cbcl/publications/ps/cancer.ps
10. Brown M.P.S., et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. (USA) **97**, 262-267 (2000).
11. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, Feature Selection for SVMs. Advances in Neural Information Processing Systems, vol. 13, 2000.
12. Kaplan, E.L. and Meier, P. Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457-481 (1958).
13. A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, J. L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Welsenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, and L.

Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. ***Nature* 4051**, 503-511 (2000).

14. M. Shipp, D. Harrington, Chairpersons, J. Anderson, J. Armiage, G. Bonadonna, G. Brittinger, F. Cabanillas, G. Canellos, B. Coiffier, J. Connors, R. Cowan, D. Crowther, M. Engelhard, R. Fisher, C. Gisselbrecht, S. Horning, E. Lepage, A. Lister, J. Neerwaldt, E. Montserrat, N. Nissen, M. Oken, B. Peterson, C. Tondini, W. Velasquez, and B. Yeap. A predictive model for aggressive non-Hodgkin's lymphoma: The International NHL Prognostic Factors Project. N Engl J Med 329:987-994 (1993).