Supplementary information for

Mediastinal Large B-Cell Lymphoma:
A unique subset of diffuse large B-cell lymphoma with an expression signature resembling that of Hodgkin's Reed-Sternberg Cells

Kerry J. Savage, Stefano Monti, Jeffery L. Kutok, Giorgio Cattoretti, Donna Neuberg, Laurence de Laval, Paul Kurtin, Paola Dal Cin, Christine Ladd, Friedrich Feuerhake, Ricardo Aguiar, Sigui Li, Gilles Salles, Francoise Berger, Wen Jing, Geraldine Pinkus, Thomas Habermann, Riccardo Dalla-Favera, Nancy Harris, Jon C. Aster, Todd Golub, and Margaret Shipp.

Abstract

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 at our web site www.broad.mit.edu/cancer/pub/mediastinal.

Case selection and Histologic classification

Frozen diagnostic tumor specimens from 34 MLBCL patients and 176 DLBCL patients were analyzed according to an Institution Review Board-approved protocol. MLBCL tumor specimens were derived from mediastinal masses or contiguous nodal biopsies and DLBCLs were all nodal tumor specimens. Primary MLBCLs were identified using clinical criteria (predominant mediastinal mass with or without local extension and no extrathoracic disease) and pathologic features. The histopathology and immunophenotype of each primary MBLCL was reviewed and expert hematopathologist to confirm diagnosis.

Target cRNAs of oligonucleotide microarrays

Total RNA was extracted from each frozen tumor specimen and biotinylated cRNAs were generated as previously described [1]. Samples were hybridized overnight to Affymetrix U133A and U133B oligonucleotide microarrays (Affymetrix, Santa Clara, CA) which include probe sets

from over 44,000 genes. Arrays were subsequently developed with phycoerythrin-conjugated streptavidin (SAPE) and biotinylated antibody against streptavidin, and scanned to obtain quantitative gene expression levels [1].

Preprocessing and Re-scaling

The raw expression data consists of the Affymetrix's scanner "signal" units as obtained from Affymetrix's GeneChip MAS5. This raw data is re-scaled to account for different chip intensities. Each sample (column) in the data set is multiplied by the factor *constant/sample_intensity*, where *sample_intensity* denotes the sample's average intensity (i.e., the sample's expression level averaged across all probe sets in chips A and B combined); and *constant* is the same quantity for all samples (chosen to be the average intensity of the median sample).

After this preprocessing, genes were ranked according to a variation filter so as to give higher priority to genes with expression values showing maximal variation across the samples being analyzed. For this purpose, we used the *median absolute deviation* (MAD) as our ranking score. The MAD for gene *g*, denoted as MAD(*g*), was computed as follows:

$$\mathrm{MAD}(g) = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} \left| g_i - med(g) \right|,$$

where N is the number of samples, g_i denotes g's expression level in sample i, and $med(g)$ denotes the median expression level of gene g.

Gene expression differential analysis and permutation test

The top 15,000 genes from the U133A and U133B Affymetrix chips were selected based on their ranking as measured by MAD. From within this 15,000-gene pool, genes correlated with the class distinction of interest (1="mediastinal" *vs.* 0="non-mediastinal") were identified by ranking them according to their signal-to-noise ratio (SNR) [1, 2],

$$\mathrm{SNR} = \frac{\mu_1 - \mu_0}{\sigma_1 + \sigma_0},$$

where μ_i and σ_i denote, respectively, the sample mean and sample standard deviation within class $i=1,0$. Similar rankings were obtained by using the median in place of the mean, or by using the t-statistic in place of the SNR. A Monte Carlo simulation of the permutation distribution of the SNR's was performed by permuting the sample labels indicating class membership (n=1000); thereafter, the observed values in the data were compared to the 99th percentile of the permutation. Notice that the permuted SNRs are associated to a gene's rank, not its identity. That is, the p-value for the k^{th} ranked gene is computed with respect to the empirical null distribution obtained by repeating the following steps multiple times: i) shuffle the class labels; ii) compute the associated SNR's for all genes; and iii) select the k^{th} SNR (irrespective of the identity of the gene achieving this score). In so doing, the resulting p-values account for the effect of multiple testing. Furthermore, by shuffling the class labels, the expression correlation among the genes is preserved, thus making the permutation test more stringent (i.e., making it harder to achieve statistical significance).

Classification

The discriminatory power of the gene expression signature was also evaluated by building classifiers for the MLBCL *vs.* DLBCL distinction. To this end, we used both the naïve-Bayes classifier and the weighted-voting classifier, and evaluated their error rates.

The naïve-Bayes classifier

The naïve-Bayes (NB) classifier [3] provides an estimate of the probability that a given sample belongs to one of the given classes. This is the *class posterior probability* $P(C|g_1,g_2,\ldots,g_M)$, where C denotes the binary class variable taking values 1 (MLBCL) and 0 (DLBCL), and where $g_1,\ldots,g_M$ denote the M gene markers used for prediction. The NB classifier derives its name from the independence assumption on which it is built. This assumption asserts that, within each class, the expression level of a gene is independent of the expression level of other genes. The assumption is often unrealistic (naïve), but it allows for the parsimonious factorization of the probability of interest, and for its computation by a straightforward application of the Bayes theorem:

$$1.\quad P(C \mid g_1, g_2, ..., g_M) = \frac{P(g_1, g_2, ..., g_M \mid C)P(C)}{P(g_1, g_2, ..., g_M)} = \frac{P(C)\prod_{i=1}^{M} P(g_i \mid C)}{\sum_{C=1,2} P(C)\prod_{i=1}^{M} P(g_i \mid C)},$$

where $P(C)$ denotes the *prior probability* of class membership, and $P(g_i|C)$ denotes the *conditional probability* distribution of a gene within a given class. Notice that to compute the posterior probability of interest all that is needed is an estimate of the prior probability of class membership $P(C)$, and of the conditional probabilities $P(g_i|C)$ for each gene marker g_i. For example, if we use our 210-sample dataset as our training set (with 34 MLBCLs and 176 DLBCLs), the prior $P(C)$ can be estimated as

$$P(C = 1) = \frac{34}{210} = 0.16$$

$$P(C = 0) = 1 - P(C = 1) = \frac{176}{210} = 0.84$$

The conditional probability $P(g_i|C=1)$ for any gene g_i is estimated by fitting a Normal distribution to the set of 34 observations of g_i in the MLBCL class. Similarly, the conditional probability $P(g_i|C=0)$ is estimated by fitting a Normal distribution to the set of 176 observations of g_i in the DLBCL class.

Once these quantities are estimated from data, the classification of a new sample can be performed by applying Equation (1) to the new sample, and assigning the sample to class 1 if $P(C=1|g_1, g_2,\ldots,g_M) > P(C=0|g_1, g_2,\ldots,g_M)$, and to class 0 otherwise.

The weighted-voting classifier

The weighted-voting (WV) classifier [4] is very similar to the NB classifier (see the appendix in [4]). The WV classifier makes a weighted linear combination of relevant marker genes obtained in the training set to provide a classification scheme for new samples. The selection of the classifier input features (marker genes) is accomplished by computing a signal-to-noise ratio S_g =

$(\mu_1-u_0)/(\sigma_1+\sigma_0)$. The class predictor is uniquely defined by the initial set of samples and marker genes. In addition to computing S_g, the algorithm also finds the decision boundaries (half way) between the class means: $B_g = (\mu_1+u_0)/2$ for each gene g. To predict the class of a test sample y, each gene g in the feature set casts a vote: $V_g(y)= S_g (g_y - B_g)$, where g_y denotes the expression level of gene g in sample y. The final vote for class 0 or 1 is computed as follows:

$$\text{vote}(y) = \text{sign}\left(\sum_{i=1}^{M} V_{g_i}(y)\right).$$

The sample y is assigned to class 1 if vote(y)>0, and to class 0 otherwise. The strength or confidence in the prediction of the winning class is $(V_{win} \_ V_{lose})/(V_{win} + V_{lose})$ (i.e., the relative margin of victory for the vote).

Training and testing procedure

With both classifiers, the following supervised learning procedure was followed:

1. Build a classifier by leave-one-out cross validation (LOOCV). That is, for each sample repeat the following steps: i) hold out the given sample; ii) use the remaining samples as a training set to build a classifier; and iii) test the resulting classifier on the held-out sample.
2. Each classifier is built by selecting the "marker" genes with the highest correlation with the target class as measured by SNR. Markers are selected in a balanced fashion. That is, when selecting *M* markers, *M/2* markers will be selected as the most up-regulated for class 1, and the other *M/2* as the most up-regulated for class 0. The marker selection is carried out within the CV loop. That is, when holding out a given sample, that sample is excluded from the computation of the SNR.
3. Several models are built using different numbers *M* of marker genes and the final chosen model is the one that minimizes the total error in cross-validation.

Since the two classes are heavily unbalanced (with the DLBCL class including 176 samples, and the MLBLC class including 34 samples), the error rate is computed as the average of the error rates within the two classes, referred to as *balanced* error rate.[1] Figure 1.a reports the balanced 210-sample LOOCV error rates for classifiers based on different numbers of genes/markers. Classifiers with numbers of genes ranging from 10 to 1000 were built and evaluated. The NB classifier with 100 genes achieved the best error rate of 11.33%, with 5 MLBCLs wrongly classified and 14 DLBCLs wrongly classified.

[1] This approach yields an error rate that is in general higher than the simple proportion of samples wrongly classified. On the other hand, it is a more accurate way of reporting error rates when handling unbalanced classes. For example, if we were to report the simple error rate of the majority classifier (i.e., the classifier assigning all samples to the majority class) on our data, this would yield an error rate of 34/210=16%. The corresponding balanced error rate, on the other hand, would be 50%.

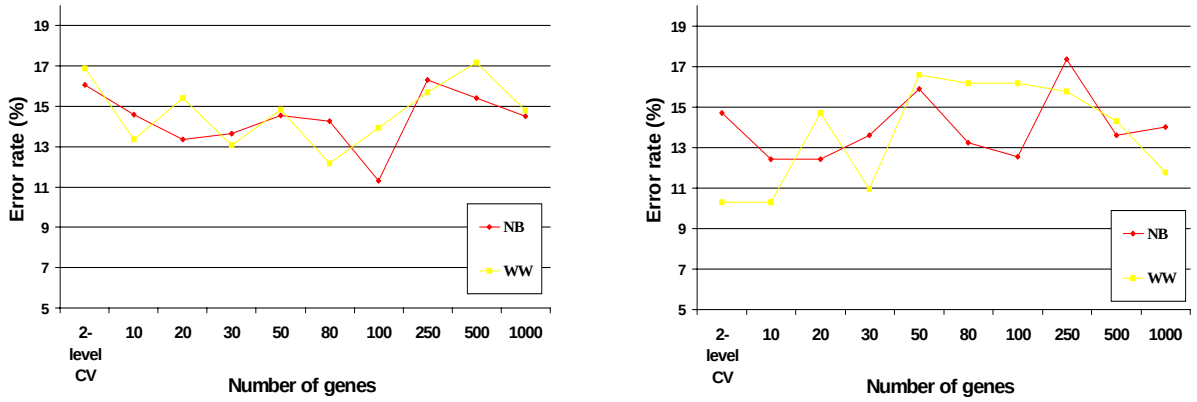

Figure 1: Balanced error rates computed by LOO-CV. a) Error rates based on 210 samples. b) Error rates based on the 159 samples remaining after the removal of the 51 samples with "mediastinum involvement" (see text).

Notice that building classifiers based on different numbers of genes, and choosing the one that achieves the lowest error rate is not an entirely unbiased procedure, since we are using the "test set" multiple times. To compute a more rigorous estimate of the error rate we can expect to attain on new data, we also used a 2-level cross-validation procedure. This procedure uses the CV outer-loop to estimate the error rate reported (as described above). Within each outer-loop iteration, another cross-validation loop (inner-loop CV) is carried out to choose the best number of features to use. This means that classifiers based on different numbers of features can be used in each of the outer-loop CV iterations. Based on this procedure, the estimated balanced error rate for the NB (WV) classifier was 16.07% (16.88%), with 9 MBLCL errors and 10 DLBCL errors (8 DLBCL and 18 MLBCL). The number of genes selected ranged from a minimum of 10 to a maximum of 250, with a median number of 10 (min: 10, max: 50, median: 20).

Within the set of 210 samples, a subset of 51 samples was clinically diagnosed as DLBCL while having some level of mediastinum involvement. We reasoned that these samples would render the distinction between the two classes less sharp. Consequently, we built classifiers based on the remaining 159 samples (125 DLBCLs, 34 MLBCLs), and we tested them by LOOCV. Figure 1.b reports the balanced error rates for classifiers based on different numbers of genes/markers. As shown, this procedure yielded slightly lower error rates. The best classifier (a WV classifier based on 10 genes) achieved a balanced error rate of 10.28%, with 4 MLBCL errors and 11 DLBCL errors. The 2-level CV error rate for the WV classifier is also 10.28% (4 MLBCL errors, 10 DLBCL errors), with the number of genes selected ranging from a minimum of 10 to a maximum of 50, and with a median number of 10.

Enrichment test

We carried out an enrichment test [5] in order to assess the significance of the similarity between the Hodgkin Lymphoma (HL) signature and the MLBCL signature. The HL signature was defined using a set of genes independently identified by Kuppers et al. as differentially expressed in Hodgkin's Reed-Sternberg (HRS) cell lines when compared with normal B-cells using Affymetrix U95 oligonucleotide arrays [6]. The U95 probes were mapped to U133 probes

according to the mapping provided by Affymetrix. The similarity between the two signatures was assessed based on the following procedure:

1. 15,000 genes were selected from the U133A/B chips according to a MAD-based variation filter, and ranked according to their SNR with respect to the "MLBCL vs. DLBCL" class membership (i.e. ranked from the most under-expressed to the most over-expressed genes in MLBCL).
2. The set of 294 genes identified in [6] as under-expressed in HL were located within the ranked list of 15K genes, and their proximity to the under-expressed end of the list measured by a Kolmogorov-Smirnoff (KS) score (with a higher score corresponding to a higher proximity).
3. Permutation of the "MLBCL *vs.* DLBCL" sample labels, associated re-ranking of the 15K genes, and computation of the corresponding KS score were performed multiple times (n=1000), so as to compare the observed KS score with the KS score that could be expected by chance under a random class labeling. An empirical p-value was thus computed to quantify the significance of the similarity between the HL signature and the MLBCL signature.
4. An alternative, less stringent, method to compute an empirical p-value is based on the computation of the KS score for a random set of 294 genes (since 294 is the number of genes comprising the HL signature) multiple times. This allows for the assessment of how likely it would be to observe a KS score equal to or greater than the one actually observed, if a random set of genes were to be selected. The drawback of this method is that it discounts possible correlations among the genes comprising the actual HL signature.

A similar procedure can be adopted to compare the gene signatures with respect to the over-expressed genes. To this end, we can use the set of 195 genes identified in [6] as significantly over-expressed in HL cell lines, and test for their proximity to the over-expressed end of the list.

A graphical rendition of the computation of the KS score for the set of over-expressed and under-expressed genes is shown in Figure 2.a and Figure 2.b, respectively. In Figure 2.a, the bottom panel indicates with vertical bars the location of the over-expressed HL genes within the ranked set of 15K genes, with the genes over-expressed in MLBCL to the left, and the under-expressed genes to the right. The top panel shows the assignment of "rewards" and "penalties" to the overall KS score as the list of 15K ranked genes is scanned from the over-expressed end (left) to the under-expressed end (right) of the ranking. Every "hit" (i.e., the encounter of a Hodgkin gene during the scan) increments the KS score, and every "miss" (i.e., the encounter of a non-Hodgkin gene) decrements the score, resulting in the step-wise curve showed. The final score corresponds to the highest value in the plot (in the y-axis). High enrichment would correspond to a steep climb upward to the left. Lack of enrichment would correspond to a lack of clear upward trend in the curve. We denote with p_{min} the less stringent p-value described in item 4 above, and with p_{max} the more stringent empirical p-value obtained by permuting the class labels, ad described in item 3 above. All p-values were computed based on 1000 permutation iterations. To make sure that the results of the analysis were not overly dependent on the gene pool used (the 15K genes selected by MAD), we repeated the analysis based on the entire set of unfiltered 44K genes, and we obtained similar p-values.

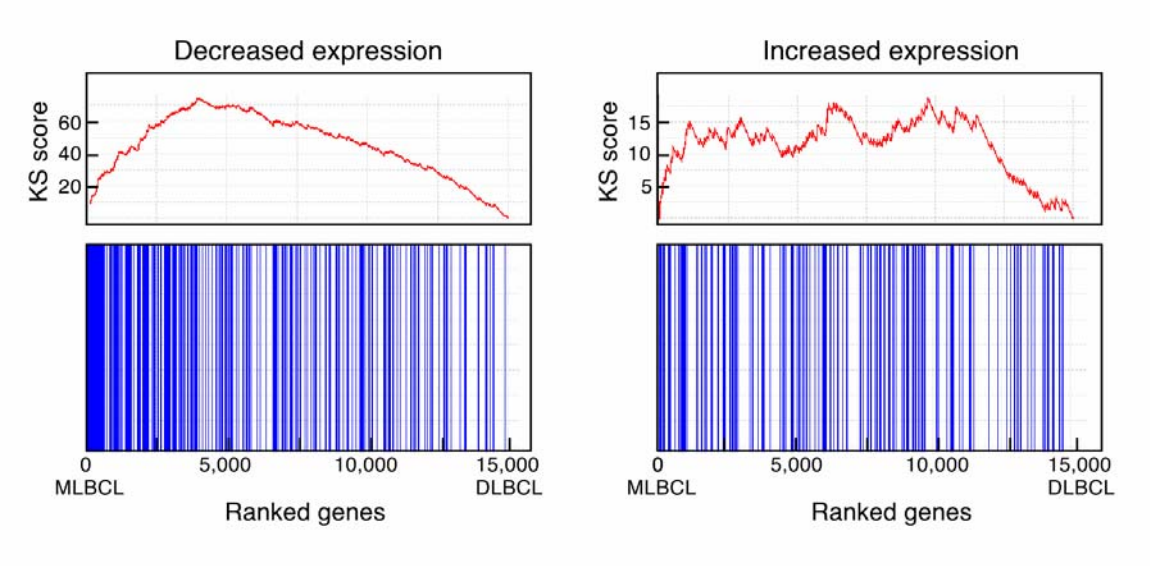

Figure 2: Kolmogorov-Smirnof-based enrichment test. a) Evaluation of the similarity between the HL signature and the mediastinal signature with respect to the set of up-regulated HL genes. b) Evaluation of the similarity between the HL signature and the mediastinal signature with respect to the set of down-regulated HL genes.

As shown in Figure 2, the enrichment analysis indicates that the observed relationship between the under-expressed genes in MLBCL and HL is highly significant (p_{max}=0.012, p_{min}<0.001). There is a less significant similarity between the over-expressed genes in HL RS cell lines and primary MLBCLs (p_{max}=0.213, p_{min}=0.007). This likely reflects the contribution of tumor microenvironment to primary MLBCL and HL signatures and the absence of these features in the signatures of isolated HL RS cell lines [6].

1. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression.* Science, October 15 1999. **286**(5439): p. 531-537.
2. Slonim, D.K., et al., *Class Prediction and Discovery Using Gene Expression Data*, in *RECOMB 2000: The Fourth Annual International Conference on Research in Computational Molecular Biology*. 2000: Tokyo, Japan. p. 263--272.
3. Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis.* 1973.
4. Slonim, D.K., et al., *Class Prediction and Discovery Using Gene Expression Data*, in *RECOMB 2000: The Fourth Annual International Conference on Research in Computational Molecular Biology*. 2000: Tokyo, Japan. p. 263-272.
5. Mootha, V.K., et al., *PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nature Genetics, 2003. **34**(3): p. 267-273.
6. Kuppers, R., et al., *Biology of Hodgkin's lymphoma.* Ann Oncol, 2002. **13**(90001): p. 11-18.