Supplementary information for

Molecular profiling of diffuse large B-cell lymphoma reveals discrete clusters including one characterized by host inflammatory response

Stefano Monti[1*], Kerry J. Savage[2*], Jeffery L. Kutok[3], Friedrich Feuerhake[2], Paul Kurtin[4], Martin Mihm[5], Bingyan Wu[6], Laura Pasqualucci[7], Donna Neuberg[6], Ricardo C.T. Aguiar[2], Paola Dal Cin[3], Christine Ladd[1], Geraldine S. Pinkus[3], Gilles Salles[8], Nancy L. Harris[6], Riccardo Dalla-Favera[7], Thomas Habermann[9], Jon C. Aster[3], Todd R. Golub[10**], Margaret A. Shipp[2**]

[1]Whitehead Institute/Massachusetts Institute of Technology (MIT) Center for Genome Research, Cambridge, MA; [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA; [3]Department of Pathology, Brigham and Women's Hospital, Boston, MA; [4]Department of Pathology, Mayo Clinic, Rochester, MN; [5]Department of Pathology, Massachusetts General Hospital, Boston, MA; [6]Department of Biostatistics, Dana-Farber Cancer Institute, Boston, MA; [7]Institute for Cancer Genetics, Columbia University, New York, NY; [8]Hematology Department, Centre Hospitalier Lyon-Sud, Lyon, France; [9]Division of Hematology and Department of Medicine, Mayo Clinic, Rochester, MN; [10]Department of Pediatrics, Dana-Farber Cancer Institute and Howard Hughes Medical Institute, Boston, MA

* These authors contributed equally to the work.

** These authors contributed equally to the work.

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 http://www.broad.mit.edu/cancer/pub/dlbcl.

Table of contents

Case selection and Histological classification

Frozen diagnostic tumor specimens and retrospective clinical data from 176 DLBCL patients were analyzed according to an Institution Review Board-approved protocol. DLBCL tumor specimens were derived from nodal biopsies of newly diagnosed, previously untreated patients (Table 1).

Clinical feature	**# of patients**[a]
Age[b]	
< = 60	64
> 60	112
Sex	
Male	92
Female	84
Performance Status	
0,1	129
2 or more	32
Not available	15
Stage	
I,II	53
III,IV	115
Not available	8
LDH	
Normal	64
Abnormal	81
Not available	31

# Extranodal sites	
<2	153
2 or more	21
Not available	2
IPI risk group	
Low – 0-1	50
Low Int - 2	32
High Int - 3	38
High – 4,5	24
Not available	32
	57% (95%CI 48 – 66%)
Overall survival (5 yr. Median)	

[a]Total 176 patients

[b]Median age 64 years (range 20-92 years)

Table 1: clinical characteristics of DLBCL study patients.

The histopathology and immunophenotype of each DLBCL tumor was reviewed by expert hematopathologists to confirm the diagnosis. Clinical variables that constitute the International Prognostic Index (IPI) (age, stage, number of extranodal sites, LDH, PS) were obtained and a full IPI score was available for 144 patients. Overall survival (OS) and freedom from progression (FFP) were determined by the Kaplan-Meier method in 128 study patients who received full-dose CHOP-based (cyclophosphamide, adriamycin, vincristine, prednisone) therapy (e.g. 3-4 cycles +XRT for localized disease or minimum of 6 cycles for advanced disease) and had long-term clinical follow-up or disease progression during induction therapy.

Preprocessing and Re-scaling

The raw expression data for the 176 samples 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 chip's values in the dataset are multiplied by the factor *constant/chip_intensity*, where *chip_intensity* denotes the chip's average intensity (i.e., the expression level averaged across all probe sets in the chip); and *constant* is the same quantity for all chips (chosen to be the average intensity of the median chip). This process is repeated for chip A and chip B separately (in general, the average intensity on chip A is three to four times the average intensity of chip B. The chip-specific rescaling maintains this ratio).

Gene selection based on duplicates

When carrying out unsupervised analysis (clustering), the selection of the genes to include in the dataset is particularly critical, since a known phenotype is not available to discriminate between informative and un-informative genes. Inclusion of too large a proportion of "noise" genes may drown the expression signature of interest. In order to try to identify a set of maximally informative genes, we carried out the selection based on duplicate samples. In particular, 17 samples were randomly chosen from the overall population of 176 de-novo DLBCL samples. These were processed in duplicates (that is, a total of four chips, two U133A chips and two U133B chips, were processed for each of the 17 samples) and a statistical analysis of these samples was performed to identify

genes with high reproducibility within pairs/duplicates, and high variation across pairs/patients. Within-duplicate reproducibility guarantees that the expression measurement is reliable. Across-patients variation is evidence that the gene captures the variation in the sample population. Genes were ranked according to a modification of the F statistic, and the top 5 percentile was selected for inclusion in the final gene set (with higher values of the F statistic at the top). Figure 1 visually summarizes the rationale behind the selection criteria and the statistic used to rank the genes. The standard F statistic is routinely used in the analysis of variance to test for differences among means, and it is computed as follows:

$$F = \frac{B}{W}$$

with

$$B = \sum_{i \in \text{num.pairs}} \sum_{j \in \text{pair}_i} (\bar{g}_{i.} - \bar{g}_{..})^2 / (\text{num.pairs} - 1)$$

$$W = \sum_{i \in \text{num.pairs}} \sum_{j \in \text{pair}_i} (g_{ij} - \bar{g}_{i.})^2 / \text{num.pairs}$$

Where $\bar{g}_{i.}$ denotes the average intensity of gene *j* within pair *i*, and $\bar{g}_{..}$ denotes the average intensity across all samples. In other words, *W* measures the average *within*-duplicate variation, which we should aim to minimize, and *B* measures the *between*-patients variation, which we should aim to maximize. Therefore, the larger the value of F, the better.

The problem with the F statistic thus defined is its sensitivity to outliers. The modification of the F statistic we use for ranking genes, which we refer to as *robust* F statistic, is obtained by replacing the mean with the median, and by replacing the power 2 with the absolute value.

A total of 2118 genes were identified by selecting the top 5-percentile of the distribution of genes ranked according to their robust F statistic. To evaluate the sensitivity of the unsupervised analysis to the set of genes included, a larger gene pool was also selected by taking the top 10-percentile of the distribution, which yielded a set of 4224 genes. Figure 2 shows the duplicates' plots for some representative genes. Notice that even the "worst" gene within the top 5-percentile (plot in the middle) still manifests a considerable reproducibility and variation.

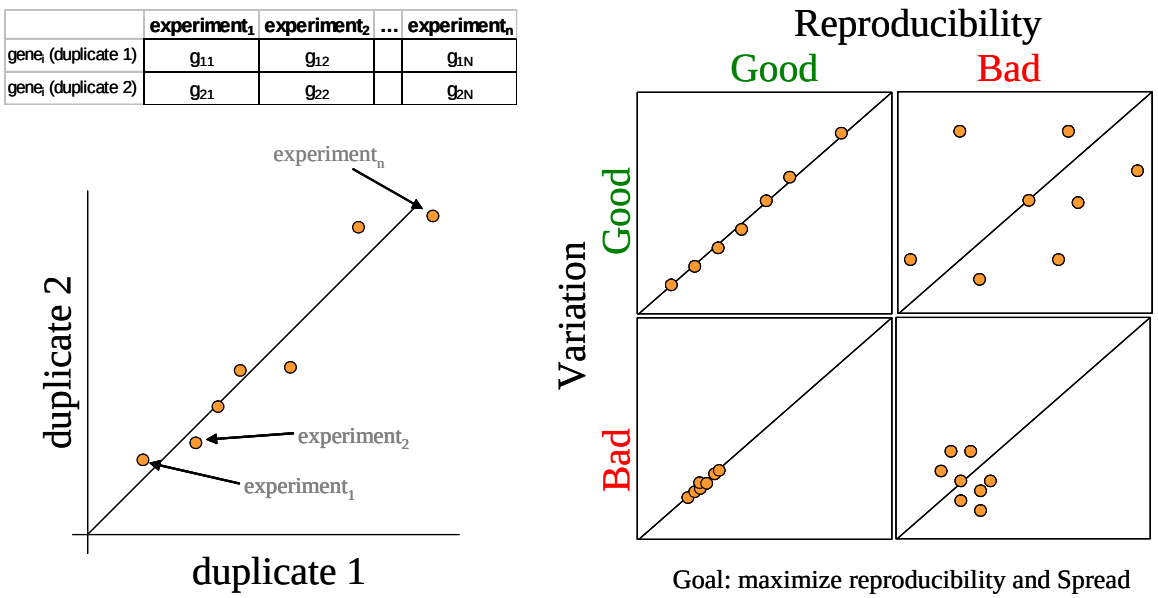

Figure 1: visual rendition of the criteria of gene selection. Given a gene (*gene*$_i$), and its expression across N=20 duplicate pairs, each pair defines the coordinate of a point in a 2-dimensional plot. Perfect reproducibility, would translate into all 20 points falling on the *y=x* diagonal line. High variation would translate into the 20 points being widely spread along this same diagonal line. The robust F statistic employed is a summary statistic accounting for both these measures (see text).

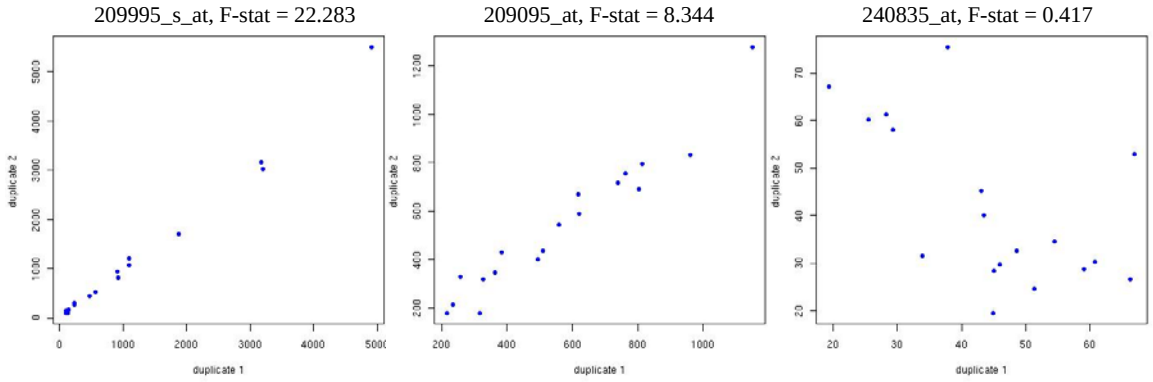

Figure 2: Duplicates' expression levels for some representative genes. Each point in a plot corresponds to a distinct duplicate pair. The left-most plot corresponds to the gene with the largest value of the robust F statistic. The plot in the middle corresponds to the gene with the smallest F statistic within the top 5-percentile. The right-most plot corresponds to the gene with smallest F-statistic.

Class discovery

Given the recognized heterogeneity of the broadly defined class of DLBCLs, clustering analysis was carried out in an attempt to identify biologically meaningful subsets of

DLBCL samples sharing a similar gene expression signature. To this end, *consensus clustering* [1] was applied to the set of 176 de-novo samples projected on the space of 2118 genes identified by duplicate analysis. Gene markers for the putative clusters were identified, and enrichment analysis carried out to elucidate the biological significance of the putative clusters.

Consensus Clustering

Consensus clustering (CC) provides for a method to represent the consensus across multiple runs of a clustering algorithm and to assess the stability of the discovered clusters (i.e., to assess the sensitivity of the cluster boundaries to sampling variability). Perturbations of the original dataset are simulated by resampling techniques; the clustering algorithm of choice is applied to each of the perturbed datasets; and the agreement, or *consensus*, among the multiple runs is assessed and summarized in a *consensus matrix.* Each matrix entry is indexed by a sample pair, and it measures the proportion of times the pair's samples are clustered together across the resampling iterations. A consensus matrix corresponding to perfect consensus would contain 1's and 0's only, reflecting the fact that any two samples are always or never clustered together. A distinct consensus matrix is generated for each of the number of clusters considered. We should select the number of clusters whose corresponding matrix most closely approaches perfect consensus. The consensus matrix can also be visualized. If the items in the matrix are arranged so that samples belonging to the same cluster are adjacent to each other, perfect consensus translates into a block-diagonal matrix with non-overlapping blocks of 1's along the diagonal – each block corresponding to a different cluster – surrounded by 0's. If we associate a color gradient to the 0-1 range of real numbers, so that white corresponds to 0, and red corresponds to 1, a matrix corresponding to perfect consensus will be displayed as a color-coded *heat map* characterized by red blocks along the diagonal, on a white background. Visual inspection of the sorted consensus matrices, and of the corresponding summary statistics can be used to determine the best number of clusters. The summary statistics used are the Lorenz curve and the corresponding Gini index for the distribution of consensus matrix entries. These quantities are measures of inequality routinely used in economics. They measure the deviation from the scenario where the all values in a population are the same (perfect equality).

In our analysis, the data-set perturbations were obtained by sub-sampling, whereby 80% of the original samples are randomly selected without replacement from the total set of 176 samples, yielding a dataset of 141 samples. We ran 200 sub-sampling iterations for each of the clustering algorithms considered. Since different clustering algorithms yield different results, we explored the application of hierarchical clustering [2], the self-organizing map [3, 4], and model-based probabilistic clustering [5, 6] as implemented in AutoClass [7].

Hierarchical clustering (HC) and the self-organizing map (SOM) both require the number of clusters to be specified. Consequently, we built consensus matrices for a number *K* of clusters ranging between 2 and 9. Figure 3 and Figure 4 show the results of consensus clustering based on HC and SOM respectively. In both cases, inspection of the consensus matrices and the corresponding summary statistics (Lorenz curve and change in Gini index) suggests not to go past 3 clusters.

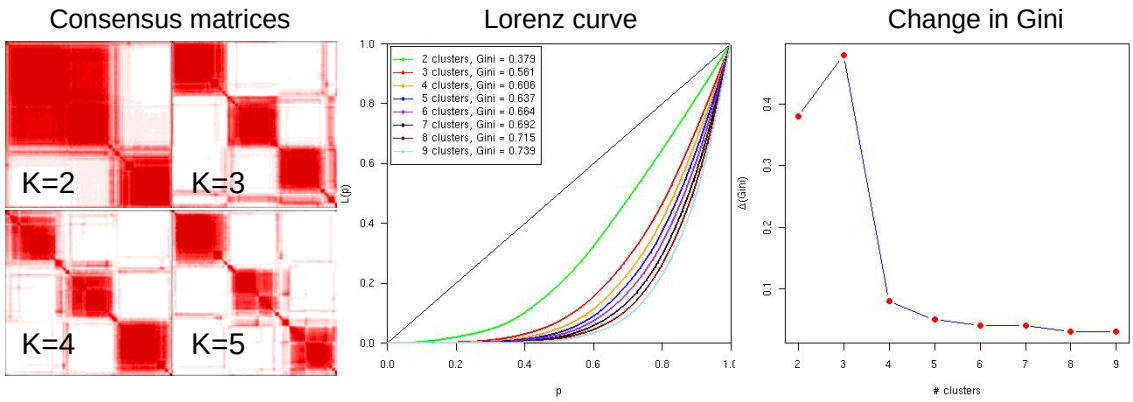

Figure 3: Results of consensus clustering with HC. a) Consensus matrices corresponding to a number *K* of clusters ranging between 2 and 5. b) Plotting the Lorenz curve for *K=2,…, 9*. c) Corresponding change in the Gini index (area of the Lorenz curve).

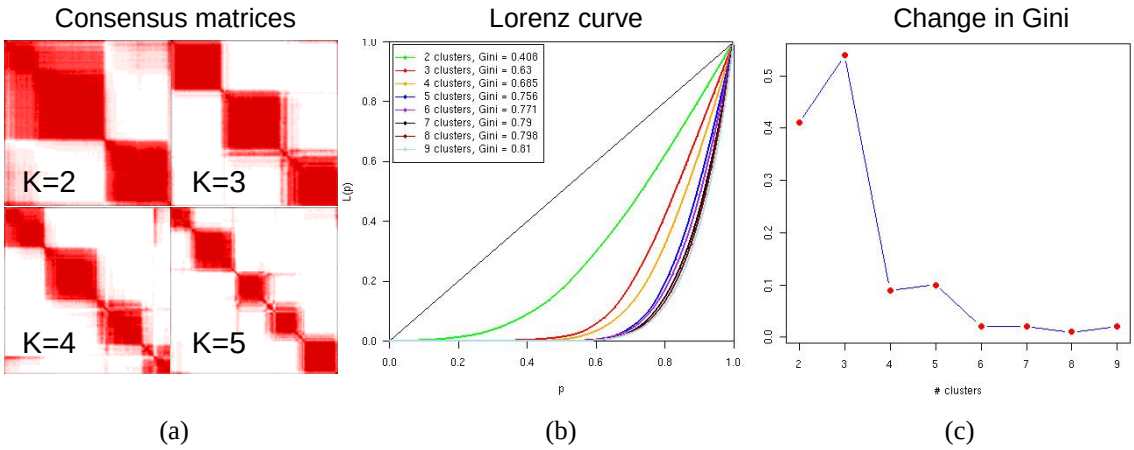

Figure 4: Results of consensus clustering with SOM. a) Consensus matrices corresponding to a number *K* of clusters ranging between 2 and 5. b) Plotting the Lorenz curve for *K=2,…, 9*. c) Corresponding change in the Gini index (area of the Lorenz curve).

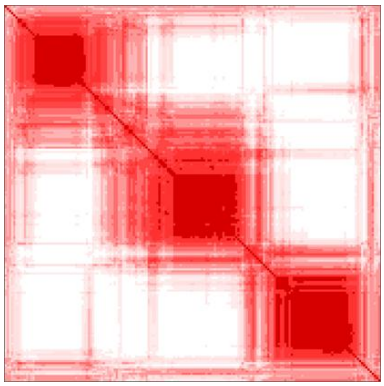

Figure 5: Results of consensus clustering with probabilistic clustering (PC). At each iteration, PC automatically selects the best number of clusters (see text). Notice the 3-cluster structure that emerges.

Model-based probabilistic clustering (PC) is different from HC and SOM in that it uses a probabilistic score to automatically determine the best number of clusters. Consequently, when running consensus clustering with PC, a single consensus matrix is built, representing the consensus across multiple clustering iterations, with each iteration possibly yielding a different number of clusters. Inspection of the consensus matrix, shown in Figure 5, and of the distribution of number of clusters selected by PC across the 200 iterations can still be used to choose the optimal *K*. The color-coded consensus matrix for PC clearly manifests a 3-cluster structure. This is in agreement with the observation that in most of the 200 iterations PC selected 3 clusters.

In most cases, application of different clustering yields clusters of different composition. Figure 6 shows the confusion matrices measuring the agreement between the clusters produced by the different clustering algorithms considered. For example, the confusion matrix for the PC *vs.* SOM comparison (left-most matrix) shows that: the 1st PC cluster has 55 items in common with the 3rd SOM cluster; the 2nd PC cluster has 50 items in common with the 2nd SOM cluster; and the 3rd PC cluster has 43 items in common with the 1st SOM cluster. This means that 148 of the 176 samples are assigned to the same clusters by the two algorithms. The confusion matrices for the other comparisons (PC *vs.* HC and HC *vs.* SOM) can be similarly interpreted. Figure 6 also reports the value of the *adjusted Rand index* (RI) for each comparison [8, 9]. The adjusted Rand index is a measure of the agreement between two data partitions. It ranges between 0 and 1, with 1 corresponding to perfect agreement, and 0 corresponding to the expected value of the index for two independent random partitions (for examples and a detailed explanation of its derivation, see, e.g., [6]). The high values observed (0.58, 0.69. 0.65) indicate that the clusters produced by the three clustering algorithms manifest a high level of agreement.

		SOM			
		C3	**C2**	**C1**	RTOT
	C1	**55**	5	10	70
PC	**C2**	5	**50**	0	55
	C3	5	3	**43**	51
	CTOT	65	58	53	176

RI = 0.576

		HC			
		C2	**C1**	**C3**	**RTOT**
	C1	**57**	1	12	70
PC	**C2**	4	**51**	0	55
	C3	2	1	**48**	51
	CTOT	63	53	60	176

RI = 0.689

		SOM			
		C3	**C1**	**C2**	**RTOT**
	C1	**50**	1	2	53
HC	**C2**	3	**49**	6	58
	C3	7	3	**55**	65
	CTOT	60	53	63	176

RI = 0.655

Figure 6: measuring the agreement between the output of two clustering algorithms. Each confusion matrix measures the number of items in common between any two clusters in the two partitions considered. The left-most matrix compares the clusters produced by consensus clustering based on PC with the clusters produced by consensus clustering based on SOM, and it shows that 148 out of the 176 items are assigned to the same clusters. The center matrix compares PC to SOM, and the right matrix compares HC to SOM. Below each matrix is reported the value of the adjusted Rand Index (see text).

PC *vs.* HC	PC *vs.* SOM				
	C1	**C2**	**C3**	**DK**	**RTOT**
C1	**50**	0	0	5	55
C2	0	**49**	0	1	50
C3	0	0	**42**	1	43
DK	7	2	6	**13**	28
CTOT	57	51	48	20	176

Meta-consensus

Figure 7: Representing the meta-consensus of the three clustering methodologies considered.

Since there are no objective criteria for selecting the output of one clustering algorithm over the other two, we considered the "meta-consensus" obtained by selecting those cluster members that are in agreement among the three. This approach is equivalent to assigning a "don't know" (DK) label to those items whose cluster membership is not consistent across the three clustering partitions. By doing so, we focus only on those samples whose cluster membership we are more confident about. The downside is that we reduce the number of samples considered for analysis. Figure 7 shows the process of reconciling the partitions induced by the three clustering algorithms, which yields a meta-consensus partition consisting of three clusters containing 50, 49, and 42 samples respectively (with 35 samples left unassigned).

In an attempt to recover the "DK" samples, we also built a naïve-Bayes (NB) classifier [10] trained on the 141-samples and the corresponding cluster labels, and used it to predict the cluster membership of the 35 DK samples. We refer to the 141-sample dataset partitioned according to meta-consensus as DB_{141}, and to the 176-sample dataset partitioned according to the NB predictions as DB_{176}. Most of the subsequent analysis is carried out on DB_{141}. However, since several clinical and cytogenetic annotations are available on some of the DK samples, some of the analysis involving these annotations is also carried out on DB_{176}.

Differential analysis for the selection of clusters' markers

From the 2218-gene pool, genes associated with each of the three clusters in DB_{141} were identified by ranking them according to the signal-to-noise ratio (SNR) of the "one *vs.* all" binary labeling, a statistic for marker selection that has been used in several studies in place of the standard t statistic [11, 12]. In particular, for each of the three clusters *C=1,2,3*, we considered the binary distinction "cluster C *vs*. NOT cluster C", and ranked the genes according to the SNR thus defined:

$$\mathrm{SNR} = \frac{\mu_C - \mu_{\neg C}}{\sigma_C + \sigma_{\neg C}},$$

where μ_i and σ_i denote, respectively, the sample mean and sample standard deviation within class $i=C,\neg C$ (where $\neg C$ is short for "NOT cluster C"). Similar rankings were obtained by using the median in place of the mean, or by using the t-statistic in place of the SNR.

Gene-set enrichment analysis

In an effort to elucidate the biological meaning of the clusters identified, we carried out enrichment tests with respect to a wide set of functionally related gene-sets. The purpose of these tests is to establish whether the gene signature for a given cluster is enriched in genes that share some biologically relevant annotation.

A total of 281 gene sets from 4 independent sources were utilized: 1) Biocarta, an internet resource (www.biocarta.com) that includes 169 biological pathways involved in adhesion, apoptosis, cell activation, cell cycle regulation, cell signaling, cytokines/chemokines, developmental biology, hematopoesis, immunology, metabolism, and neuroscience; 2) GenMAPP (Gene MicroArray Pathway Profiler, www.GenMAPP.org), a set of web-accessible pathways and gene families including 45 gene sets involved in metabolic and cell signaling processes; 3) 64 recently described manually curated pathways involved in mitochondrial function as well as fatty-acid, amino acid, lipid and glucose metabolism and an expression compendium of gene sets that are co-regulated in normal murine tissues [13]; and 4) three recently described co-regulated gene sets in DLBCL [14] (Table 2). Each of these gene sets was tested for enrichment against the sets of up- and down-regulated markers for each of the three clusters. To this end, the non-parametric Kolmogorov-Smirnoff (KS) rank test was used as previously described in [13, 15] and as implemented in the GSEA (Gene Set Enrichment Analysis) software.

In particular, given a gene expression dataset (e.g. DB_{141} defined over 2118 genes), a cluster of interest (e.g., cluster 1), and a gene set of interest (e.g., the cell cycle pathway, defined in the GenMAPP repository as containing 150 genes, of which 30 are included in the 2118-gene set), enrichment is assessed as follows: 1) the 2118 genes in DB_{141} are ranked according to their differential expression with respect to the phenotype "cluster 1 vs. not cluster 1"; 2) the genes in the set (30 genes, in the example) are located within the ranked DB_{141} genes, and their proximity to the up-regulated (or down-regulated) end of the list is measured by computing a Kolmogorov-Smirnoff (KS) score (with higher values of the score corresponding to higher proximity); 3) a permutation test is performed where cluster labels are scrambled multiple times and a distribution of "permutated" KS scores is obtained to test for the significance of the observed KS score. This procedure is repeated for each of the gene sets and clusters considered. Since multiple gene sets are tested for enrichment, reported p values for each observation are corrected for multiple hypothesis testing (MHT-p) by comparing the observed KS score to the distribution of permutated KS scores for *all* of the gene sets. That is, the p-values were obtained by pooling the permuted KS scores for all the gene sets tested, and by locating the observed KS scores within the resulting permutation distribution. A $p \leq 0.005$ was used to identify highly significant associations between specific gene sets and identified DLBCL clusters. Table 1 shows only those gene sets with $p \leq 0.005$. All p-values were computed based on 1000 permutation iterations. The stringent threshold of 0.005 was used so as to focus only on the most striking gene sets annotations.[1]

[1] In the original analysis, we used the method to correct for MHT available in the GSEA software [16]. More recently, additional methods of correcting for MHT have gained wide acceptability, such as the false discovery rate (FDR) [17]. Computation of individual permutation-based p-values and multiple-hypothesis correction by FDR yield largely concordant results, with similar cluster-defining gene sets.

	OxPhos		BCR/Proliferation		Host Response	
A.	**KS**	**MHT p**	**KS**	**MHT p**	**KS**	**MHT p**
Mitochondrial pathways						
PGC	130.9	0.004	13.2	0.763	3.8	0.931
VOXPHOS	156.1	0.001	13.3	0.760	2.9	0.948
Human mito DB	152.6	0.002	11.8	0.790	0.6	0.987
Mitochondrial	157.5	0.001	16.2	0.703	0.4	0.991
OXPHOS	141.0	0.003	13.7	0.753	2.9	0.946
Gen MAPP						
Electron transport	148.1	0.000	14.4,	0.641	3.1	0.839
Cell cycle	33.4	0.298	104.8	0.004	0.8	0.873
Complement activation - classical	11.9	0.685	7.4	0.766	105.9	0.004
BioCarta						
Complement	12.0	0.745	6.4	0.846	91.7	0.004
T cytotoxic	8.5	0.809	7.0	0.835	113.5	0.000
T helper	8.5	0.809	7.0	0.835	113.5	0.000
T ob1	27.0	0.440	2.8	0.902	103.2	0.002
Co-regulated gene sets						
C7	26.2	0.494	7.4	0.870	164.6	0.001
C10	130.5	0.004	1.4	0.98	19.7	0.62
B.						
Additional DLBCL gene sets						
Proliferation	120.4	0.103	142.0	0.064	3.1	0.854
Lymph node	24.3	0.577	1.7	0.902	279.7	0.000

Table 2: enrichment test results.

Clusters' validation on duplicate samples

In order to assess the sensitivity of the discovered clusters to noise and possible bias in the protocol followed to process the sample tissues, we used the set of 17 samples for which replicates were available. As previously illustrated, the dataset contains 17 samples for which three replicates were collected and processed (the replicate included in the clustering analysis, plus the two replicates used for the duplicate-based gene selection previously described). We therefore built a predictive model (a naïve-Bayes classifier) based on the 176 samples including one copy of the 17 replicates, and we tested its predictions on the other 2 copies of these 17 samples. All 17 samples were correctly classified, as shown in Figure 8.

		Predicted		
		IR	**BCR**	**Oxphos**
True	**IR**	4	0	0
	BCR	0	26	0
	Oxphos	0	0	4

34 predictions (17 x 2)

Figure 8: confusion matrix for the prediction of the 17 duplicate samples.

Clusters' validation on an independent dataset

To further assess the sensitivity of the clustering results to sampling variability, we replicated our clustering analysis on an independent group of 221 newly diagnosed DLBCLs with available cDNA microarray ("lymphochip") profiles [14, 18]. For this analysis, we first mapped the list of 2118 genes in the lymphochip platform. The mapping, based on unique accession numbers, resulted in the identification of 703 Affymetrix probe sets with a corresponding lymphochip probe set; the mapping was one-to-many, yielding a final set of 1784 lymphochip probes. Table 3 shows the number of matching markers assigned to each cluster, with 208, 229, and 266 markers for the OxPhos, BCR, and HR clusters respectively. The set of gene markers assigned to a given cluster is defined as the set of genes whose SNR for the corresponding one-vs-all distinction is the highest (e.g., for a three-cluster partition, the set of markers for cluster 1 is the set of genes whose "cluster 1 *vs.* NOT cluster 1" SNR is higher than the "cluster 2 *vs.* NOT cluster 2" SNR and the "cluster 3 *vs.* NOT cluster 3" SNR). We replaced the multiple probes corresponding to the same Affymetrix gene with a meta-probe whose expression value corresponds to the average of the component probes' values. Thereafter, we utilized the 703 lymphochip probes, and combined by meta-consensus the output of HC, SOM, and PC algorithms to identify the dominant structure in the independent DLBCL dataset. Ninety percent of tumors were assigned to the same groups by all 3 algorithms. In the remainder of this section, we refer to the Affymetrix data and the resulting DB_{141} clustering as the "Affymetrix clusters", and to the Lymphochip data and the corresponding clusters as the "Lymphochip clusters."

	Affy			
	Count	**Avg. Rank**	**Avg. Score**	**Fold Chg**
OxPhos	**209**	333	0.31	1.39
BCR	228	304	0.38	1.59
Immune	266	290	0.49	1.84
Total	703			

Table 3: Apportionment to the three DLBCL clusters of the 703 Affymetrix markers with a valid mapping in the Lymphochip. The average rank, SNR score, and fold change are all computed over the top 209 markers per clusters, so as to be able to compare the relative representation of the three clusters.

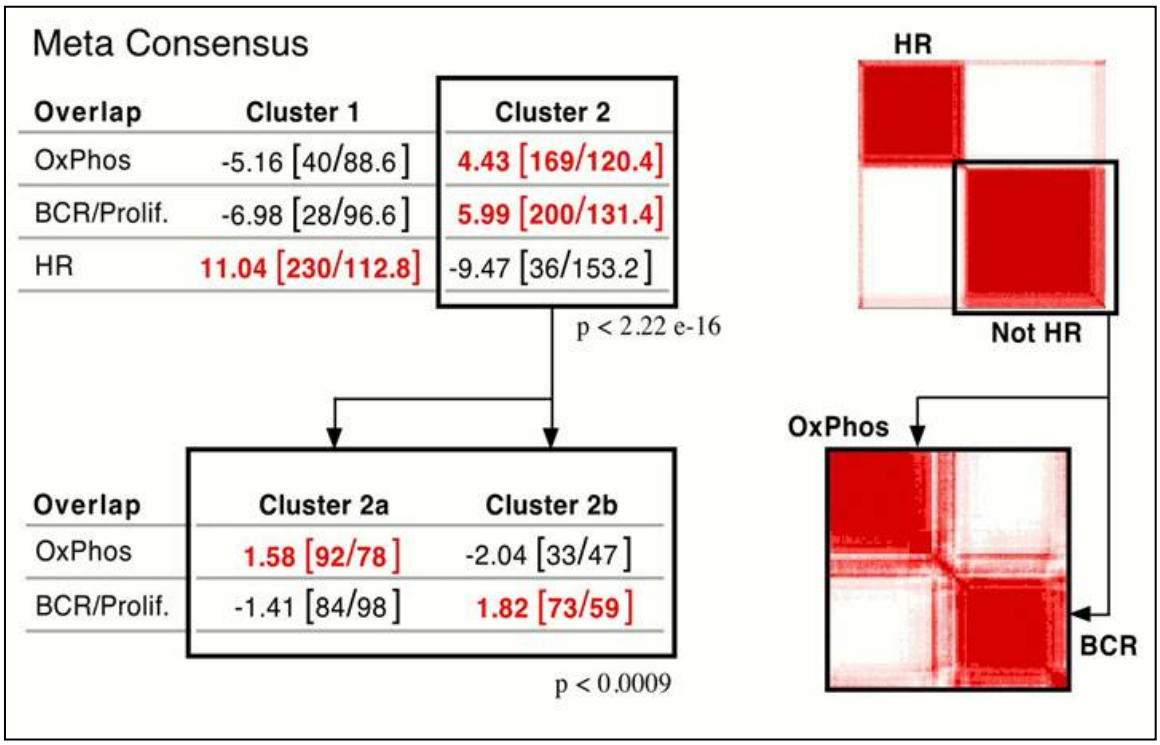

Figure 9: Clustering of Lymphochip data. Top: "meta-consensus" matrix corresponding to two clusters (right), and corresponding contingency table (left) comparing the Affymetrix markers (rows) to the Lympochip markers (columns). Each contingency table entry has the form *residual [observed/expected]*, where *observed* is the number of observed markers, *expected* is the number of markers expected under the null hypothesis, and *residual* is the statistical score measuring the difference between the observed and the expected counts. Bottom: "meta-consensus" matrix (right) corresponding to clustering the "Not HR" samples in the space of "Not HR" markers; and corresponding contingency table (left).

To compare the clusters obtained on the two independent datasets, we compare the gene markers of all Affymetrix-Lymphochip cluster pairs in terms of their *overlap*. In particular, the entire set of genes (in our case, 703 common genes), is partitioned into marker subsets for the clusters considered based on their one-vs-all SNRs (e.g., for the three Affymetrix clusters, the gene set will be partitioned into three subsets). The overlap within all cluster pairs can then be measured and organized into a 2-dimensional contingency table. A Fisher test is performed to assess the deviation of the overlap from what would be expected by chance assuming the absence of any signature similarity between the two clustering partitions.

Our clustering procedure subdivided the independent DLBCL series into two major groups (Fig. 9, top right panel) rather than the three clusters found in our data. Ninety percent of tumors were assigned to the same groups by all 3 algorithms. The signature for one of the independent clusters was highly enriched for HR markers (overlap P-value $<2.2 \times 10^{-16}$). The second independent cluster signature was enriched for BCR/proliferation transcripts, but it also included OxPhos markers (Fig. 9, top left panel). However, as summarized in Table 1, a number of the most discriminating OxPhos markers are not represented on the cDNA microarray platform, a potential reason for the

less precise definition of the OxPhos subgroup. For this reason, we further analyzed the "non-HR" tumors by clustering this group in the space of non-HR markers (i.e., by excluding the samples in the "HR" cluster and the corresponding markers). As shown (Figure 9, bottom right panel), the "non-HR" tumors separated into 2 discrete clusters with significant enrichment for either BCR/proliferation or OxPhos markers (overlap P-value < 0.0009, bottom left panel).

In summary, when validating our 3-cluster structure in the Lymphochip data, we find a very strong correlation with the HR cluster, and the BCR cluster, and a less strong, although still significant, correlation with the OxPhos cluster.

Cell of origin signature

Recent studies suggest that subsets of newly diagnosed DLBCLs share elements of the transcriptional profile of normal purified germinal center B-cells (GCB), *in vitro*-activated peripheral blood B-cells (ABC) or other less well-defined cells (Other) [14, 18]. We used our dataset to further validate these results.

In particular, we considered the 27 Lymphochip probes described in [18] as defining the COO signature, we identified the corresponding Affymetrix probes, and projected our 176-sample dataset on the space of these probes. Of the 27 Lymphochip probes, only 23 have a (one-to-many) mapping on the U133 chip-pair, yielding a total of 59 Affymetrix probes. From this pool, we excluded 9 probes whose average expression level was lower than 10. Furthermore, since some of the genes has a one-to-many mapping, we replaced the set of probes corresponding to the same gene with a new *meta-probe* whose expression value is given by the average expression values of the component probes (averaging). In some cases, the multiple probes corresponding to the same gene manifest very poor correlation, and taking their average does not seem warranted. In these cases, we take the expression value of the meta-probe to be equal to the expression value of the component probe with highest average expression (maximize). We used a correlation of 0.2 as the threshold to determine whether to average (when $\rho>0.2$) or to maximize (when $\rho\leq 0.2$) the expression value of multiple probes.

To establish a partition of the data based on the COO signature, we replicated the procedure described in [18]. This procedure defines a *linear predictive score* (LPS) for each sample:

$$\mathrm{LPS}(\mathrm{sample}) = \sum_{i=1}^{27} \beta_i g_i ,$$

where β_i is the "GCB vs ABC" t-statistic for the *i*-th gene, and g_i its measured expression in the given sample. Under the assumption that the LPSs follow a Normal distribution, their within-class means and standard deviations are estimated and combined in a Bayes predictor as follows:

$$P(\mathrm{sample} \in \mathrm{GCB}) = \frac{\Phi(\mathrm{LPS}(\mathrm{sample}) \mid \mu_{GCB}, \sigma_{GCB})}{\Phi(\mathrm{LPS}(\mathrm{sample}) \mid \mu_{GCB}, \sigma_{GCB}) + \Phi(\mathrm{LPS}(\mathrm{sample}) \mid \mu_{ABC}, \sigma_{ABC})},$$

where $\Phi(x \mid \mu, \sigma)$ denotes the Normal probability density function of a random variable *x* with mean μ and standard deviation σ. Given the probability of membership in the GCB class, the probability of membership in the ABC class is P(sample $\in$ ABC) = 1 –

P(sample $\in$ GCB). Samples are then assigned to the GCB class (or the ABC class) if their probability of membership in that class is greater than 0.9.

Replication of this procedure on our data involved the following steps: i) linear predictive scores (LPS) based on the 23 common genes were defined (that is, we used the LPS coefficients for the 23 mapped genes as estimated on the Lymphochip data); ii) within-class distributions of the 23-gene LPSs were estimated based on the Lymphochip data; iii) the Affymetrix data was normalized so as to have the same mean and standard deviation as the Lymphochip data; and iv) a Bayesian classification rule based on the 23-gene LPSs was applied to the normalized data to assign each sample to either the GCB class or the ABC class, using a probability of 0.9 as the threshold for assignment (that is, samples not assigned to either class with probability>0.9 were assigned to a Type-3 ("Other") class. Application of this classification procedure yielded a partition of the data into 24 ABC samples, 64 GCB samples, and 42 Type-3 samples. Figure 10 shows the class-specific Kaplan-Meier curves, and the highly significant correlation of the COO classes with survival ($p<0.003$). Of note, tumors identified as GCB were associated with significantly longer overall survivals.

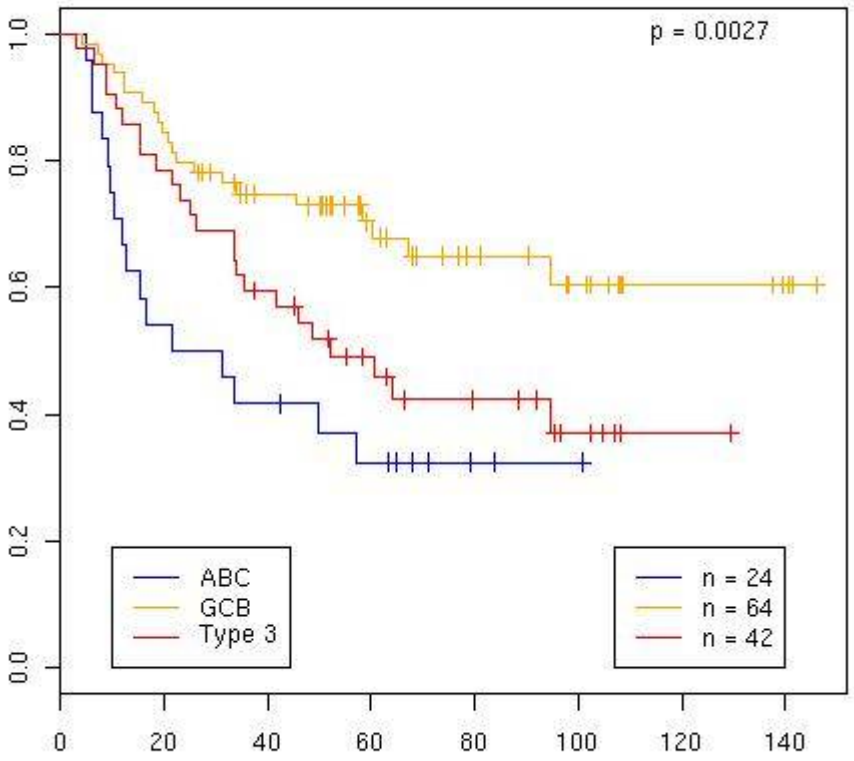

Figure 10: Kaplan-Meier curves corresponding to the COO-based partition of our DLBCL series.

Immunohistochemistry (IHC)

Two representative 0.6mm cores were obtained from diagnostic areas of available paraffin- embedded formalin- or B5-fixed DLBCLs (80 tumors) and inserted into a grid pattern using a automated tissue array. Three -- 5μ micron sections were cut from the tissue arrays for subsequent immunohistochemical analyses with mouse monoclonals anti-CD2 (LFA-2) (Novocastra Laboratories LTD, Newcastle upon Tyne, UK), anti-CCR7 and –CD123 (Bioscience, San Diego, CA) and anti CD1a (Dako, Carpintera, CA) rabbit polyclonal anti-CD3 and anti-S100 (Dako), and anti-gamma interferon-induced lysosomal transferase (GILT, Gift from Peter Cresswell, Yale University School of Medicine, New Haven, CT [19]).

In brief, slides were deparaffinized and pre-treated with 1.0 mM EDTA pH 8.0 (for CD2 CD3 and CD1a immunostaining) or 10-mM citrate, pH 6.0 (for GILT, CCR7 and CD123 immunostaining) in a steam pressure cooker (Decloaking Chamber, BioCare Medical, Walnut Creek, CA). No pretreatment was required for S100 immunostaining. Next, slides were treated with Peroxidase Block (DAKO USA, Carpinteria, CA) for 5 minutes to quench endogenous peroxidase activity and incubated with 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 antibodies were diluted in 50-mM Tris-Cl, pH 7.4 with 3% goat serum (anti-CD2 1:250, anti-CD3 1:400, anti-S100 1:3000, anti-GILT 1:2000, anti-CD1a 1:100, anti-CD123 1:25 and anti-CCR7 1:2000) and applied to slides at room temperature for 1 hour. Goat anti-mouse or rabbit horseradish peroxidase-conjugated antibody (Envision detection kit, DAKO) was applied for 30 minutes and developed using a diaminobenzidine (DAB) chromogen kit (DAKO). Harris hematoxylin counterstain was applied and slides were analyzed in blinded fashion by two expert hematopathologists, without information regarding consensus cluster designations.

The numbers of CD2+ and CD3+ cells/core were separately recorded for duplicate samples and represented in 5 categories: 1) <50 cells/core; 2) 50-150 cells/core; 3) 150-250 cells/core; 4) 250-500 cells/core; 5) >500 cells/core. Separate analyses of GILT immunostaining of dendritic cells and tumor cells were performed. The number of GILT+ dendritic cells/core was assessed in duplicate samples and represented in 3 categories: 1) 0-25 cells/core; 2) 25-100 cells/core; and 3) >100 cells/core. The number of S100+ dendritic cells/core was assessed in duplicate samples and represented in 4 categories: 1) 0-25 cells; 2) 25-50 cells; 3) 50-100 cells; and 4) > 100 cells. The associations between comprehensive consensus clusters and numbers of CD2+ and CD3+ cells and GILT+ and S100+ dendritic cells were measured with a Kurskal-Wallis exact test. The associations between numbers of CD2+ and CD3+ infiltrating T-cells and GILT+ and S100+ dendritic cells in individual tumors were determined using a Jonckheere-Terpstrat test.

Acknowledgments

We would like to thank Aravind Subramanian, Pablo Tamayo, and all the members of the Cancer Genomics group at the Broad Institute and Glenn Dranoff at the Dana-Farber Cancer Institute for useful discussions and input.

References

1. Monti, S., et al., *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Machine Learning, 2003. **52**(1-2): p. 91-118.
2. Eisen, M.B., et al., *Cluster Analysis and Display of Genome-Wide Expression Patterns.* Proceedings of the National Academy of Sciences, 1998. **95**: p. 14863-14868.
3. Kohonen, T., *Self-Organizing Maps.* 1997.
4. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation.* Proceedings of the National Academy of Sciences, March 1999. **96**: p. 2907-2912.
5. Titterington, D.M., A.F. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions.* 1985.
6. Yeung, K.Y., et al., *Model-based clustering and data transformations for gene expression data.* Bioinformatics, 2001. **17**(10): p. 977-987.
7. Cheeseman, P. and J. Stutz, *Bayesian Classification (AutoClass): Theory and Results,* in *Advances in Knowledge Discovery and Data Mining*, U.M. Fayyad, et al., Editors. 1996. p. 153-180.
8. Rand, W.M., *Objective Criteria for the Evaluation of Clustering Methods.* Journal of the American Statistical Association, 1971. **66**(336): p. 846-850.
9. Milligan, G.W. and M.C. Cooper, *An examination of procedures for determining the number of clusters in a data set.* Psyochometrika, 1985. **50**: p. 159-179.
10. Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis.* 1973.
11. Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression.* Science, 1999. **286**(5439): p. 531-537.
12. 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.
13. 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.
14. Rosenwald, A., et al., *The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma.* New England Journal of Medicine, 2002. **346**(25): p. 1937-1947.
15. Savage, K.J., et al., *The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin's lymphoma.* Blood, 2003. **102**(12): p. 3871-3879.
16. Subramanian, A., *Xtools: A software package for Gene Set Enrichment Analysis*. 2003, Broad Institute of MIT & Harvard. Unpublished.
17. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
18. Wright, G., et al., *A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma.* Proceedings of the National Academy of Sciences, 2003. **100**(17): p. 9991-9996.

19. Arunachalam, B., et al., *Enzymatic reduction of disulfide bonds in lysosomes: Characterization of a Gamma-interferon-inducible lysosomal thiol reductase (GILT).* PNAS, 2000. **97**(2): p. 745-750.