Identification of Molecularly and Clinically Distinct Types of Diffuse Large B-Cell Lymphoma By Gene Expression Profiling


Depts. of *Biochemistry, ‡Genetics, ||Medicine, ‡‡‡Pediatrics, ***Health Research & Policy and Statistics, and ## the Howard Hughes Medical Institute Stanford University School of Medicine, Stanford, California 94305, USA
§Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
¶Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA
#CBER, FDA, Bethesda, Maryland 20892, USA
**Research Genetics, Huntsville, Alabama 35801, USA
Departments of ††Pathology and Microbiology, and §§Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA
§§Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
||||Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA
¶¶Walter Reed Army Medical Center, Washington, DC, USA

†These authors contributed equally to this work.

Proofs should be sent to:
Louis M. Staudt
Metabolism Branch, NCI
Bldg. 10, Rm. 4N114
National Institutes of Health
9000 Rockville Pike
Bethesda, Maryland 20892, USA
Tel (301) 402-1892
Fax (301) 496-9956
lstaudt@box-l.nih.gov
SUMMARY

Diffuse Large B-Cell Lymphoma (DLBCL), the most common subtype of non-Hodgkin’s lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival whereas the remainder succumb to the disease. We hypothesized that this variability in natural history reflects unrecognized molecular heterogeneity in the tumors. Using DNA microarrays, we have conducted a systematic characterization of gene expression in B cell malignancies. Here we show striking diversity in gene expression among the tumors of DLBCL patients, apparently reflecting variation in tumor proliferation rate, host response, and differentiation state of the tumor. We identified two molecularly distinct forms of DLBCL which had gene expression patterns indicative of different stages of B cell differentiation. One type expressed genes characteristic of germinal center B cells (GC B-like DLBCL); the second type expressed genes normally induced during in vitro activation of peripheral blood B cells (Activated B-like DLBCL). GC B-like DLBCL patients had a significantly better overall survival than Activated B-like DLBCL patients. Molecular classification of tumors based on gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

INTRODUCTION

Despite the variety of clinical, morphological, and molecular parameters used to classify human malignancies today, patients receiving the same diagnosis can have markedly different clinical courses and treatment responses. The history of cancer diagnosis has been punctuated by reassortments and subdivisions of diagnostic categories. There is little doubt that our current taxonomy of cancer still lumps together molecularly distinct diseases with distinct clinical phenotypes. Molecular heterogeneity within individual cancer diagnostic categories is already evident in the variable presence of chromosomal translocations, deletions of tumor suppressor genes and numerical chromosomal abnormalities. The classification of human cancer is likely to become increasingly more informative and clinically useful as more detailed molecular analyses of the tumors are conducted.

The classification of human lymphomas has steadily evolved since their initial recognition by Thomas Hodgkin in 1832 ¹. Beginning with the distinction of Hodgkin's disease from other malignant and non-malignant conditions by Sternberg ² and Reed ³, a variety of
lymphoma classifications have been advanced based on both morphologic and molecular parameters. The most recent classification scheme, the Revised European-American Lymphoma (REAL) Classification was introduced to categorize distinct clinical-pathological entities. However, within this classification system, various morphologic subtypes were unified into groups despite the suspicion that they “include more than one disease entity”.

Diffuse Large B-Cell Lymphoma (DLBCL) is one disease in which attempts to define subgroups based on morphology have largely failed due to diagnostic discrepancies arising from inter- and intra-observer irreproducibility. DLBCL is an aggressive malignancy of mature B lymphocytes, with an annual incidence of over 25,000 cases, accounting for roughly 40% of cases of non-Hodgkin's lymphoma. Patients with DLBCL have highly variable clinical courses: while the majority of patients respond initially to chemotherapy, fewer than half of the patients achieve a durable remission. Although a combination of clinical parameters is currently used to assess a patient’s risk profile, these prognostic variables are considered to be proxies for the underlying cellular and molecular variation within DLBCL.

An important component of the biology of a malignant cell is inherited from its non-transformed cellular progenitor. Each of the currently recognized categories of B cell malignancy has been tentatively traced to a particular stage of B cell differentiation, although the extent to which these malignancies maintain the molecular and physiological properties of normal B cell subsets is not clear. The rearranged immunoglobulin genes in DLBCL and most other non-Hodgkin's lymphomas bear mutations that are characteristic of somatic hypermutation, an antibody diversification mechanism that normally occurs only within the germinal center of secondary lymphoid organs. This evidence suggests that DLBCL arises either from germinal center B cells or from B cells at a later stage of differentiation.

In the present study, we examined the extent to which genomic-scale gene expression profiling can further our understanding of B cell malignancies. We addressed whether we could: 1) generate a molecular portrait of distinct types of B cell malignancies, 2) identify distinct types of B cell malignancies not recognized by the current classification system, and 3) relate each malignancy to normal stages in B cell development and physiology. We focused particularly on DLBCL to determine if gene expression profiling could subdivide this clinically heterogeneous diagnostic category into molecularly distinct diseases with more homogeneous clinical behaviors.
Construction of a specialized DNA microarray

Recent technical and analytical advances make it practical to quantify the expression of thousands of genes in parallel using cDNA microarrays. This mode of analysis has been used to observe gene expression variation in a variety of human tumors. To apply this method to questions in normal and malignant lymphocyte biology, we designed a specialized microarray, the "Lymphochip", by selecting genes that are preferentially expressed in lymphoid cells and genes with known or suspected roles in processes important in immunology or cancer. Because of the suspected importance of the germinal center B cell to the genesis of non-Hodgkin's lymphomas, 12,069 of the 17,856 cDNA clones on this microarray were chosen from a germinal center B cell library. An effort was made to include all distinct genes that were initially discovered in this library. We included an additional 2,338 cDNA clones from libraries derived from DLBCL, follicular lymphoma (FL), mantle cell lymphoma and chronic lymphocytic leukemia (CLL). Finally, we added clones representing a variety of genes that are induced or repressed during B and T lymphocyte activation by mitogens or cytokines and a curated set of 3,186 genes of importance to lymphocyte and/or cancer biology. About a quarter of the genes included in this microarray were represented by two or more different cDNA clones, providing internal controls for the reproducibility of gene expression quantitation. The complete annotated list of these cDNAs can be found at http://llmpp.nih.gov/lymphoma.

Analysis of gene expression in lymphoid malignancies

We used these microarrays to characterize gene expression patterns in the three most prevalent adult lymphoid malignancies: DLBCL, FL, and CLL (Fig. 1). To provide a framework for interpretation of the gene expression in these patient samples, we also profiled gene expression in purified normal lymphocyte subpopulations under a range of activation conditions, in normal human tonsil and lymph node, and in a variety of lymphoma and leukemia cell lines. Fluorescent cDNA probes, labeled with the Cy5 dye, were prepared from each experimental mRNA sample. A reference cDNA probe, labeled with the Cy3 dye, was prepared from a pool of mRNAs isolated from 9 different lymphoma cell lines. Each Cy5-labeled experimental cDNA probe was combined with the Cy3-labeled reference probe and the mixture was hybridized to the microarray. The fluorescence ratio was quantified for each gene and reflected the relative
abundance of the gene in each experimental mRNA sample compared with the reference mRNA pool. The use of a common reference probe allowed us to treat these fluorescent ratios as measurements of the relative expression level of each gene across all of our experimental samples.

In all, roughly 1.8 million measurements of gene expression were made in 96 normal and malignant lymphocyte samples using 128 Lymphochip microarrays. Figure 1 provides an overview of the variation in gene expression across these samples. A hierarchical clustering algorithm was used to group genes based on similarity in the pattern with which their expression varied over all samples. The same clustering method was used to group tumor and cell samples based on similarities in their expression of these genes. The data are shown in a matrix format, with each row representing all the hybridization results for a single cDNA element of the array, and each column representing the measured expression levels for all genes in a single sample. To visualize the results, the expression level of each gene (relative to its median expression level across all samples) was represented by a color, with red representing expression greater than the mean, green representing expression less than the mean, and the color intensity representing the magnitude of the deviation from the mean.

Distinct clones representing the same gene were typically clustered in adjacent rows in this gene map, indicating that these genes have characteristic and individually distinct patterns of expression and demonstrating that the effects of experimental noise or artifact are negligible. Likewise, where different tumor samples from the same patient were analyzed, they were invariably found clustered in immediately adjacent columns. For example, in 3 cases of FL in which the malignant cells were separated from the normal host cells by magnetic cell sorting, the purified and unpurified samples from the same patient clustered next to each other. Two samples of leukemic cells from the same CLL patient were obtained 18 months apart, and these samples were more highly correlated in gene expression with each other than with any other patient's CLL cells. The observed patterns of gene expression thus reflected intrinsic differences between the tumors, rather than variation in handling or experimental artifacts. Moreover, these results demonstrate that even within a diagnostic category, each cancer patient has a unique tumor with a characteristic gene expression profile.

Figure 1 paints a complex, yet remarkably ordered, picture of the variation in gene expression patterns in lymphoid malignancies, with large sets of genes displaying coordinate
expression in related biological samples. Strikingly, although no information on the identity of the samples was used in the clustering, the algorithm segregated, with few exceptions, the recognized classes of lymphoid malignancies based on global similarities in their gene expression patterns. Examination of the coordinately expressed genes in each of the B cell malignancies, and comparison to the normal lymphocyte cell populations, yielded considerable insights into the biology of these malignancies. The colored bars at the right of Figure 1 indicate clusters of coordinately expressed genes that we operationally defined as gene expression "signatures". A gene expression signature was named either for the cell type in which its component genes were expressed (e.g. the "T cell" signature) or for the biological process in which its component genes are known to function (e.g. the "proliferation" signature). Thus, the overall gene expression profile of a complex clinical sample such as a DLBCL lymph node biopsy can be understood, in a first approximation, as a collection of gene expression signatures that reveal different biological features of the sample.

Gene expression patterns and phenotypic variation among B cell malignancies

One of clearest distinctions between the gene expression patterns of the three B cell malignancies involved genes that vary in expression with cellular proliferation rates. Both CLLs and FLs were clustered next to resting B cell samples reflecting, in part, the fact that both of these malignancies are relatively indolent, with very low proliferation rates. Correspondingly, these malignancies expressed cell cycle inhibitory genes such as p27kip1 and lacked expression of a large number of genes which comprise the "proliferation" signature (Fig 2). This gene expression signature included diverse cell cycle control genes, cell cycle checkpoint genes, DNA synthesis and replication genes, and the gene Ki67, commonly used to gauge the “proliferation index” of a tumor biopsy, as previously noted. The more rapidly proliferating DLBCLs had higher expression, in general, of the genes in the proliferation signature. Nonetheless, marked differences in the expression of these genes was evident between individual DLBCL samples, corresponding to the variability in proliferation index that has been previously observed in DLBCL.

The most prominent distinction between CLL and FL was provided by genes that are characteristic of germinal center B cells (Fig. 2). An extensive cluster of genes distinguished germinal center B cells from both resting blood B cells and in vitro activated blood B cells. This
is remarkable since the stimuli used to activate the blood B cells were chosen to mimic those known to be important for germinal center formation: cross-linking of the immunoglobulin receptor and CD40 signaling. However, it has thus far not been possible to mimic exactly the germinal center phenotype in vitro, as evidenced by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal center B cells 24,25. The germinal center B cell gene expression signature demonstrates that germinal center B cells represent a distinct stage of B cell differentiation and not merely one specific form of B cell activation. Support for this notion comes from the fact that the characteristic gene expression program of germinal center B cells was maintained in a cultured DLBCL cell line in the absence of the germinal center microenvironment (Figs. 1, 2).

The observation that FLs show a pattern of ongoing somatic hypermutation of immunoglobulin genes has led to the suggestion that the transformation event leading to FL occurs while the B cell is in the germinal center microenvironment 26. The gene expression signature of germinal center B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B cell differentiation (Fig. 2).

The gene expression profiles of DLBCLs were largely distinct from those of CLL and FL and demonstrated additional biological complexity in these biopsy samples. Prominent features of the DLBCL profiles appeared to reflect the non-malignant cells in these tumors. A large group of genes defined a “lymph node” signature that was shared by most of the DLBCLs and samples of normal lymph node and tonsil (Fig. 2). This signature featured genes encoding known markers of monocytes and macrophages (CD14, CD105, CSF-1 receptor) and NK cells (NK4). In addition, genes involved in the production and remodeling of the extracellular matrix were abundantly expressed (fibronectin, MMP9 matrix metalloproteinase and TIMP-3). All but one DLBCL biopsy displayed the lymph node signature, but intensity of this signature varied, possibly reflecting to the relative proportion of tumor and host cells in the lymph node biopsy.

The variable presence of T lymphocytes in DLBCL biopsies was readily discernable by a T cell gene expression signature that featured components of the T cell receptor (TCR beta, CD3 epsilon) and genes downstream of T cell receptor signaling (fyn, LAT, PKC theta) (Fig. 2). While this T cell expression signature was readily apparent in some DLBCLs, in others it was virtually undetectable.
Discovery of DLBCL subtypes

The structure of the hierarchical dendrogram in Figure 1 suggested that gene expression patterns in DLBCLs were inhomogeneous. Three “branches” of the dendrogram captured most of the DLBCL cases with only two outlying cases. Clearly, the position of any given DLBCL case in the dendrogram is determined in a complicated fashion by the influences of several distinct biological themes reflected in the expression pattern. Inspection of the gene expression map shown in Figure 1 suggested that several independent sets of genes were responsible for much of the DLBCL substructure. The expression signatures related to proliferation, T cells and lymph node biology were differentially represented in the three DLBCL branches. In addition, we noted that the genes that distinguished germinal center B cells from other stages in B cell ontogeny were also differentially expressed among DLBCLs, suggesting that B cell differentiation genes might also be used to subdivide DLBCL. Indeed, the clustering of the germinal center B cell samples with a subset of the DLBCLs in a major branch of the dendrogram in Figure 1 suggested that this group of DLBCLs might be distinguished by the similarity of its gene expression program to that of the normal germinal center B cell.

To test this hypothesis, we reclustered the DLBCL cases based only on their pattern of expression of the genes that define the germinal center B cell signature (Fig. 3A). Two large branches were evident in the resulting dendrogram. We will refer to the groups defined by these branches as GC B-like DLBCL and Activated B-like DLBCL, for reasons detailed below. The same two branches were also evident in the dendrogram in Figure 1: Activated B-like DLBCL includes all cases in the branch labeled "A" and GC B-like DLBCL includes all cases in branch labeled "G". The largest DLBCL branch in Figure 1 is a mixture of the cases assigned to the two subgroups. Normal germinal center B cells were clustered with the GC B-like DLBCL group. Indeed, the DLBCL cases in GC B-like DLBCL group expressed, to a varying degree, all of the genes that define the germinal center B cell signature. In contrast, the Activated B-like DLBCL group expressed these genes at low or undetectable levels, for the most part. These data clearly suggested that a distinct class of DLBCLs was derived from the germinal center B cell and retained the gene expression program, and presumably many of the phenotypic characteristics, of this stage of B cell differentiation.

We next searched for genes that were selectively expressed in the Activated B-like DLBCL group. This search excluded genes that were readily assigned to the proliferation, T cell
and lymph node signatures (Fig. 1) in order to focus attention on more subtle intrinsic molecular features of this group of tumors. We used hierarchical clustering to reorder this set of 2984 genes while maintaining the Fig. 3A order of the DLBCL cases (Fig. 3B). As is evident in Fig. 3C, a cluster of genes could be recognized based on their elevated expression in the Activated B-like DLBCLs, compared to GC B-like DLBCLs. It is important to note that considerable gene expression heterogeneity exists within each subgroup and that no single gene in either of these large clusters was absolutely correlated in expression with the DLBCL subgroup taxonomy. Rather, patients assigned by this method to either DLBCL subgroup shared a large gene expression program that distinguished them from the other subgroup.

**DLBCL subgroups and B cell differentiation**

We next examined how all of the genes that distinguish these DLBCL subgroups are expressed during B cell differentiation and activation. Fig. 4 shows that almost all of the genes that defined GC B-like DLBCL were highly expressed in normal germinal center B cells. Most of these genes were expressed at low or undetectable levels in peripheral blood B cells that had been activated *in vitro* by a variety of mitogenic signals. Some of the GC B-like DLBCL genes were expressed in resting blood B cells and germinal center B cells at comparable levels but not in activated peripheral blood B cells. Conversely, virtually all of the genes that were selectively expressed in germinal center B cells relative to resting or activated peripheral blood B cells were expressed by GC B-like DLBCL (data not shown).

By contrast, the majority of the genes that defined Activated B-like DLBCL were not expressed in normal germinal center B cells (Fig. 4). Instead, many of these genes, but not all, were induced during *in vitro* activation of peripheral blood B cells. The time course of expression of these genes during B cell activation varied, with some genes induced after 6 hours of activation and other genes only expressed after 48 hours of activation. Thus, the gene expression signature of Activated B-like DLBCLs is reminiscent of, but not identical to, the signature of activated peripheral blood B cells. Importantly, two DLBCL cell lines, OCI Ly3 and OCI Ly10, were among the Activated B-like DLBCLs. In fact, one or both of these cell lines expressed virtually all of the genes that defined the Activated B-like DLBCL signature. This observation suggests that signal transduction pathways that are inducibly engaged during
peripheral B cell activation and mitogenesis are constitutively active in Activated B-like DLBCLs.

The gene expression program that distinguishes GC B-like DLBCLs includes many known markers of germinal center differentiation (e.g., the genes encoding the cell surface proteins CD10 and CD38, the nuclear factor A-myb and the DNA repair protein 8-oxoguanine DNA glycosylase (OGG1)) and a host of novel genes. A particularly noteworthy gene in the GC B-like DLBCL signature is BCL-6, a well-established germinal center marker that is also the most frequently translocated gene in DLBCL. Although BCL-6 protein expression is invariably detected in DLBCL, its levels vary and are not correlated with the presence of BCL-6 translocations. Cytogenetic data are available on 16 of the DLBCL cases studied here and do not support a link between elevated BCL-6 mRNA levels in GC B-like DLBCL and BCL-6 translocations (data not shown). Thus, the higher expression of BCL-6 mRNA in GC B-like DLBCLs is most likely related to their derivation from germinal center B cells (Fig. 4).

Two other genes that can be altered by translocations in lymphoid malignancies, BCL-7A and LMO2 (TTG-2/RBTN2), have not previously been described as highly expressed in germinal center B cells. BCL-7A was cloned as part of a complex chromosomal translocation in a Burkitt's lymphoma cell line and was found to be rearranged in another cell line derived from mediastinal large B-cell lymphoma. The specific expression of BCL-7A in germinal center B cells has strong parallels with BCL-6. BCL-6 is required for germinal center formation during an antigen-driven immune response and is translocated in B cell malignancies that derive from germinal center B cells. Given the preferential expression of BCL-7A in germinal center B cells, it is conceivable that this gene also plays a role in normal germinal center physiology and in the pathophysiology of GC B-like DLBCL. LMO2 is translocated and overexpressed in a subset of T cell acute lymphoblastic leukemias and LMO2 transgenic mice have a block in early T cell differentiation and develop T cell leukemia. The selective expression of LMO2 in germinal center B cells suggests a possible role in inhibiting differentiation in the B cell lineage as well, and perhaps a corresponding role in the DLBCL malignant phenotype.

The Activated B-like DLBCL signature also includes a gene translocated in lymphoid malignancies, IRF4 (MUM1/LSIRF). IRF4 is fused to the immunoglobulin locus in some cases of multiple myeloma and can function as an oncogene in vitro. IRF4 is transiently induced
during normal lymphocyte activation \(^{39}\) (Fig. 4) and is critical for the proliferation of B lymphocytes in response to signals from the antigen receptor \(^{40}\). Thus, the constitutive expression of IRF4 in Activated B-like DLBCLs may contribute to the unchecked proliferation of the malignant cells in these tumors.

A notable feature of the gene expression pattern of Activated B-like DLBCLs was the expression of two genes whose products inhibit programmed cell death. FLIP (FLICE-like inhibitory protein/I-FLICE/FLAME-1/Casper/MRIT/CASH/CLARP) is a dominant negative mimic of caspase 8 (FLICE) which can block apoptosis mediated by fas and other death receptors \(^{41}\). FLIP is induced early during normal lymphocyte activation, presumably to block activation-induced apoptosis that occurs physiologically later in an immune response. FLIP is highly expressed in many tumor types and its constitutive expression in Activated B-like DLBCLs could inhibit apoptosis of tumor cells induced by host T cells expressing fas ligand \(^{42,43}\). The key anti-apoptotic gene BCL-2 is translocated in most cases of follicular lymphoma and in a subset of DLBCL. BCL-2 mRNA is not expressed in germinal center B cells but is induced more than 30-fold during activation of peripheral blood B cells (Fig. 4). The majority of Activated B-like DLBCLs (71%) had BCL-2 mRNA levels more than 4-fold higher than was observed in germinal center B cells (Fig. 4). This overexpression did not correlate with BCL-2 translocations (data not shown). A minority of GC B-like DLBCLs (29%) had similarly elevated BCL-2 mRNA levels suggesting that BCL-2 might also play an important role in some cases of this DLBCL subgroup.

**DLBCL gene expression subgroups define prognostic categories**

Does the taxonomy of DLBCL based on gene expression patterns define clinically distinct subgroups of patients? None of the patients included in this study had been treated prior to obtaining the biopsy sample. Further, these patients were "de novo" DLBCL cases that had not obviously arisen from pre-existing low-grade malignancies such as follicular lymphoma. Following biopsy, the patients were treated at two medical centers using comparable, standard multi-agent chemotherapy regimens. Fig. 5A presents a Kaplan-Meier plot of overall survival data from these patients, segregated according to gene expression subgroup. GC B-like and Activated B-like DLBCLs were associated with statistically significant differences in overall survival (p<0.01) and in event-free survival (data not shown). While the average 5 year survival
for all patients was 52%, 76% of GC B-like DLBCL patients were still alive at 5 years compared with only 16% of Activated B-like DLBCL patients. Thus, the molecular differences between these two kinds of lymphoma were accompanied by a remarkable divergence in clinical behavior, suggesting that GC B-like DLBCL and Activated B cell DLBCL should be regarded as distinct diseases.

A clinical indicator of prognosis, the International Prognostic Indicator (IPI), has been successfully used to define prognostic subgroups in DLBCL. This indicator takes into account the patient's age, performance status and extent and location of disease. As suspected, within our patient population a low IPI score (0-2) identified patients with better overall survival compared with patients with a high IPI score (3-5) (Fig. 5B). We next asked whether our molecular definition of DLBCL subgroups could add to the prognostic value of this clinical indicator of prognosis. Considering only patients with low clinical risk based on the IPI, patients in the Activated B-like DLBCL group had a distinctly worse overall survival than patients in the GC B-like DLBCL group (p<0.05) (Fig. 5C). Thus, the molecular dissection of DLBCL by gene expression profiling and the IPI apparently identify different features of these patients which influence their survival.

Conclusions

This study demonstrates that a genomic view of gene expression in cancer can bring clarity to previously muddy diagnostic categories. The precision of morphological diagnosis, even when supplemented with immunohistochemistry for a few markers, was insufficient in the case of DLBCL to identify believable diagnostic subgroups. While a number of individual markers have been used to define subsets of DLBCL, these studies do not provide the present overview that strongly implies that this single diagnostic category of lymphoma harbors at least two distinct diseases. Indeed, the new methods of gene expression profiling call for a revised definition of what is deemed a "disease". The two DLBCL subgroups are distinguished from each other by the differential expression of hundreds of different genes, and these genes relate each subgroup to a separate stage of B cell differentiation and activation. These molecular differences, in the light of accompanying clinical differences between these subgroups, suggest that these two subgroups of DLBCL should be considered separate diseases.
Nonetheless, we do not wish to imply that patients within a DLBCL subgroup defined here are monomorphic. As mentioned above, considerable molecular heterogeneity exists within each DLBCL subgroup. As many more DLBCL patients are studied by gene expression profiling, it is quite possible that more subgroups will emerge. Given that many current diagnostic categories of non-Hodgkin's lymphoma constitute less than 10% of the total cases, it seems likely that the DLBCL diagnostic category will also include a number of minor subgroups.

The classification scheme highlighted in this study divided DLBCL on the basis of genes that are differentially expressed within the B cell lineage. This particular classification identified patient groups that differed in survival following treatment with anthracycline-based multiagent chemotherapy regimens. It is unclear at present which of the genes that distinguish the GC B-like from Activated B-like DLBCL are the most important molecular determinants of chemotherapy responsiveness. Further, there is residual clinical heterogeneity that cannot be explained by the current classification. Despite the fact that patients with GC B-like DLBCL had an overall favorable prognosis, 5 patients died within the first two years of diagnosis. Likewise, 3 patients in the Activated B-like DLBCL subgroup were alive at 5 years following treatment, despite the poor outcome of most patients in this subgroup. By profiling the gene expression of many more DLBCLs, it may become possible to implicate a single gene or pathway in chemotherapy responsiveness with statistical certainty. More likely, however, a multivariate approach to prognosis will be needed which combines knowledge of the DLBCL subgroup, as defined here, with measurements of individual genes or pathways that contribute to treatment outcome.

Gene expression profiling presents a new way of approaching cancer therapeutics in the future. Current treatment of DLBCL typically begins with multiagent chemotherapy and then, if a complete remission cannot be maintained, patients are considered for bone marrow transplantation. The definition of prognostic groups by gene expression profiling, in combination with clinical indicators such as the IPI, may lead to the recommendation that some patients receive early bone marrow transplantations upon initial diagnosis. In testing cancer therapeutics in clinical trials, it is obviously beneficial to define homogeneous populations of patients to improve the likelihood of observing efficacy in specific disease entities. We anticipate that global surveys of gene expression in cancer, such as presented here, will identify a
small number of marker genes that will be used to stratify patients into molecularly relevant
categories that will improve the precision and power of clinical trials.

Finally, the genomic-scale view of gene expression in cancer provides a unique
perspective on the development of novel cancer therapeutics based on a molecular understanding
of the cancer phenotype. This study demonstrated that the two DLBCL subgroups differentially
expressed entire transcriptional modules comprised of hundreds of genes, many of which could
be expected to contribute to the malignant behavior of the tumor. This observation suggests that
successful new therapeutics might be aimed at the upstream signal transducing molecules whose
constitutive activity in these lymphomas leads to expression of pathological transcriptional
programs.
Methods

Microarray procedures

DNA microarray analysis of gene expression was performed essentially as described \(^{54}\). The cDNA clones on the Lymphochip microarray are listed at http://llmpp.nih.gov/lymphoma and are available from Research Genetics (Huntsville, AL). Fluorescent images of hybridized microarrays were obtained using a GenePix 4000 microarray scanner (Axon Instruments, Foster City, CA). Images were analyzed with ScanAlyze (Michael Eisen; http://www.microarrays.org/software), and fluorescence ratios (along with numerous quality control parameters; see ScanAlyze manual) were stored in a custom database. Single spots or areas of the array with obvious blemishes were flagged and excluded from subsequent analyses. Fluorescence ratios were calibrated independently by applying a uniform scaling factor to all fluorescent ratios from each array. This scaling factor was computed so that the median fluorescence ratio of well-measured spots on each array was 1.0.

All cDNA microarray analyses were performed using poly-A+ mRNA (Fast Track, Invitrogen). In each experiment, fluorescent cDNA probes were prepared from an experimental mRNA sample (Cy5-labeled) and a control mRNA sample (Cy3-labeled) isolated from a pool of 9 lymphoma cell lines (Raji, Jurkat, L428, OCI-Ly3, OCI-Ly8, OCI-Ly1, SUDHL5, SUDHL6 and WSU1). The use of a common control cDNA probe allows the relative expression of each gene to be compared across all samples \(^{20}\).

All array elements for which the fluorescent intensity in each channel was greater than 1.4 times the local background were considered well-measured and used to generate Fig. 1, with the exception of array elements that had been manually flagged as poor quality measurements by visual inspection of the array images. Fluorescence ratios were log (base 2) transformed, and stored in a table (rows=genes, columns=arrays). Where samples had been analyzed on multiple arrays, multiple observations for an array element for a single sample were averaged. Array elements that were not well-measured on at least 80% of the 128 arrays were excluded. Data for the remaining genes were normalized by subtracting the median observed value, to remove any effect of the amount of RNA in the reference pool. Fig. 1 depicts the gene expression measurements derived from 4,026 elements on the microarray. Average-linkage hierarchical
clustering was applied to both axes using the program Cluster and the results were analyzed with the program Tree View (Michael Eisen; http://www.microarrays.org/software).

**mRNA samples**

Total germinal center B cells and centroblasts were purified from human tonsils as described 27. Human blood B cells were purified from adult apheresis products or cord blood by magnetic enrichment for CD19+ cells (Miltenyi Biotec). Naive CD27+ B cells and memory CD27- blood B cells were isolated by fluorescent cell sorting starting with CD19+ adult peripheral blood B cells 55,56. Patient samples were obtained after informed consent and were treated anonymously during microarray analysis. DLBCL patients were treated at either University of Nebraska Medical Center or Stanford University School of Medicine using comparable, anthracycline-based, multiagent chemotherapeutic regimens with curative intent. DLBCL and FL lymph node biopsies were either snap frozen, frozen in OCT or disaggregated and frozen as a viable cell suspension. CLL cells were purified from untreated patients by magnetic selection for CD19+ cells (Miltenyi Biotec). *In vitro* stimulation of peripheral B cells was performed as described 24. Magnetic cell sorting was used to purify CD4+, CD45RA<sup>high</sup> T cells from human cord blood or adult peripheral blood and CD4+ thymocytes from human fetal thymus (Milteni Biotec). T cells were stimulated for 2 hr with phorbol ester (50 ng/ml) and ionomycin (1.5 µM).
ACKNOWLEDGEMENTS

The authors wish to acknowledge the support of the Cancer Genome Anatomy Project (CGAP), led by Bob Strausberg and Rick Klausner. They also thank Rick Klausner for comments on the manuscript. The authors thank Christa Prange for providing CGAP cDNA clones, Hans Messner for providing DLBCL cell lines, Howard Mostowski for sorting lymphocyte subpopulations by FACS, Holy Cross Hospital, Silver Spring, MD for providing human tonsils, Joe DeRisi for helpful advice on microarray technology and members of the Staudt, Brown and Botstein laboratories for helpful discussions. Research at Stanford was supported by grants from the National Cancer Institute to D.B., R.L. and P.O.B., who is an Associate Investigator of the Howard Hughes Medical Institute. A.A. was supported by the Howard Hughes Medical Institute Research Scholar Program while at the National Institutes of Health and by the Medical Scientist Training Program at Stanford University. M.B.E. was supported by a Computational Molecular Biology Postdoctoral Fellowship from the Alfred E. Sloan Foundation and his current address is Life Sciences Division, Lawrence Orlando Berkeley National Labs and Department of Molecular and Cellular Biology, University of California, Berkeley, CA.

Correspondence and requests for materials should be addressed to:

L.M.S. (lstaudt@box-l.nih.gov) or P.O.B. (pbrown@cmgm.stanford.edu)
FIGURE LEGENDS

Fig. 1. Heirarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression made on 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of relatedness of gene expression in each samples. The dendrogram is color coded based on the category of mRNA sample studied (see upper right key). Each row represents a separate cDNA clone on the microarray and each column a separate mRNA sample. The results presented represent the ratio of hybridization of fluorescent cDNA probes prepared from each experimental mRNA samples to a reference mRNA sample. These ratios are a measure of relative gene expression in each experimental sample and were depicted according to the color scale shown at the bottom. As indicated, the scale extends from fluorescence ratios of 0.25 to 4 (-2 to +2 in log base 2 units). Grey indicates missing or excluded data. See text for details. Full data for this figure is at http://llmpp.nih.gov/lymphoma.

Fig. 2. Expanded view of biologically distinct gene expression signatures defined by heirarchical clustering. Data are the same as in Fig. 1. The majority of the genes without designations at the right are novel genes of unknown function derived from various lymphoid cDNA libraries. See text for details.

Fig. 3 Discovery of DLBCL subtypes by gene expression profiling. A listing of the samples used in this clustering analysis is shown at the bottom. A. Heirarchical clustering of DLBCL cases (blue and orange) and germinal center B cells (black) based on the genes of the germinal center B cell gene expression signature shown in Figs. 1 and 2. Two DLBCL subgroups, GC B-like DLBCL (orange) and Activated B-like DLBCL (blue) were defined by this process. B. Discovery of genes that are selectively expressed in GC B-like DLBCL and Activated B-like DLBCL. All genes from Fig. 1, with the exception of the genes in the proliferation, T cell and lymph node gene expression signatures, were ordered by heirarchical clustering while maintaining the order of samples determined in Fig. 3A. Genes selectively expressed in GC B-like DLBCL (orange) and Activated B-like DLBCL (blue) are indicated. C. Heirarchical clustering of the genes selectively expressed in GC B-like DLBCL and Activated B-like DLBCL, discovered in Fig. 3B.
Fig. 4. Relationship of DLBCL subgroups to normal B lymphocyte differentiation and activation. The data in the left panel are taken from Fig. 3C. The right panel depicts gene expression data from the following normal B cell samples: 1: Total CD19+ Blood B cells; 2: Naive CD27- blood B cells; 3: Memory CD27+ blood B cells; 4: Cord Blood CD19+ B cells; 5: Blood B cells;anti-IgM 6h; 6: Blood B cells;anti-IgM+IL-4 6h; 7: Blood B cells;anti-IgM+CD40L 6h; 8: Blood B cells;anti-IgM+CD40L+IL-4 6h; 9: Blood B cells;anti-IgM 24h; 10: Blood B cells;anti-IgM+IL-4 24h; 11: Blood B cells;anti-IgM+CD40L 24h; 12: Blood B cells;anti-IgM+CD40L+IL-4 24h; 13: Blood B cells;anti-IgM+CD40L low 48h; 14: Blood B cells;anti-IgM+CD40L high 48h; 15: Tonsil germinal center B cells; 16: Tonsil germinal center centroblasts. Full data for this figure is at http://llmpp.nih.gov/lymphoma.

Fig. 5. Clinically distinct DLBCL subgroups defined by gene expression profiling. A. Kaplan-Meier plot of overall survival of DLBCL patients grouped based on gene expression profiling. B. Kaplan-Meier plot of overall survival of DLBCL patients grouped according to the International Prognostic Index (IPI). Low clinical risk patients (IPI 0-2) and high clinical risk patients (IPI 3-5) are plotted separately. C. Kaplan-Meier plot of overall survival of low clinical risk DLBCL patients (IPI 0-2) grouped based on gene expression profiling.
REFERENCES

CD49F=Integrin
PKC theta
LAT
CD2
CD3 epsilon
T cell receptor
b chain
fyn
Caspase 10
IRF-1
Cyclin A
BUB1 mitotic kinase
Cyclin B1
SOCS-1
p55CDC
p5k=polo-like kinase
CIP2/Cdi1/KAP1
aurora kinase
p16
Thymidine kinase
CDC21 homologue
RAD54
Dihydrofolate reductase
CD38
FAK=focal adhesion kinase
WIP=WASP interacting protein
FMR2
CD10
BCL-7A
A-myb
BCL-6
PI 3-kinase p110
RGS13
CD105
CD14
FGF-7
MMP9
fms=CSF-1 receptor
Cathepsin B
Fc receptor chain
Fibronectin 1
TIMP-3
Integrin beta 5
NK4= NK cell protein-4
SDF-1 chemokine
CD49F=Integrin 6
FAK theta
CD2
CD3 epsilon
T cell receptor chain
IRF-1
Caspase 10
A. All patients

- GC B-like
  - 21 patients, 6 deaths
- Activated B-like
  - 19 patients, 16 deaths

B. All patients

- Low Clinical Risk
  - 24 patients, 11 deaths
- High Clinical Risk
  - 14 patients, 11 deaths

C. Low clinical risk patients

- GC B-like
  - 14 patients, 3 deaths
- Activated B-like
  - 14 patients, 6 deaths

Overall Survival (years)

Probability

p = 0.01

p = 0.005

p = 0.05