Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling

Ash A. Alizadeh[1,2], Michael B. Eisen[2,3,4], R. Eric Davis[5], Chi Ma[5], Izidore S. Lossos[6], Andreas Rosenwald[5], Jennifer C. Boldrick[1], Hajeer Sabet[5], Truc Tran[5], Xin Yu[5], John I. Powell[7], Liming Yang[7], Gerald E. Marti[8], Troy Moore[9], James Hudson Jr[9], Lisheng Lu[10], David B. Lewis[10], Robert Tibshirani[11], Gavin Sherlock[4], Wing C. Chan[12], Timothy C. Greiner[12], Dennis D. Weisenburger[12], James O. Armitage[13], Roger Warnke[14], Ronald Levy[6], Wyndham Wilson[15], Michael R. Grever[16], John C. Byrd[17], David Botstein[4], Patrick O. Brown[1,18] & Louis M. Staudt[5]

Departments of [1]Biochemistry, [3]Genetics, [14]Pathology, [6]Medicine, [10]Pediatrics and [11]Health Research & Policy and Statistics, and [18]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA
[5] Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[7] Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland 20892, USA
[8] CBER, FDA, Bethesda, Maryland 20892, USA
[9] Research Genetics, Huntsville, Alabama 35801, USA
Departments of [12]Pathology and Microbiology, and [13]Internal Medicine, University of Nebraska Medical Center, Omaha, Nebraska 68198, USA
[15] Medicine Branch, Division of Clinical Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA
[16] Johns Hopkins Oncology Center, Johns Hopkins School of Medicine, Baltimore, Maryland 21287, USA
[17] Walter Reed Army Medical Center, Washington, DC 20307, USA
[2] These authors contributed equally to this work

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 proposed that this variability in natural history reflects unrecognized molecular heterogeneity in the tumours. Using DNA microarrays, we have conducted a systematic characterization of gene expression in B-cell malignancies. Here we show that there is diversity in gene expression among the tumours of DLBCL patients, apparently reflecting the variation in tumour proliferation rate, host response and differentiation state of the tumour. 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 centre B cells ('germinal centre B-like DLBCL'); the second type expressed genes normally induced during *in vitro* activation of peripheral blood B cells ('activated B-like DLBCL'). Patients with germinal centre B-like DLBCL had a significantly better overall survival than those with activated B-like DLBCL. The molecular classification of tumours on the basis of gene expression can thus identify previously undetected and clinically significant subtypes of cancer.

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 tumour 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 tumours are conducted.

The classification of human lymphomas has steadily evolved since their initial recognition by Thomas Hodgkin in 1832 (ref. 1). Beginning with the distinction of Hodgkin's disease from other malignant and non-malignant conditions[2,3], a variety of lymphoma classifications have been advanced on the basis of both morphologic and molecular parameters[4]. The most recent classification scheme, the Revised European–American Lymphoma (REAL) classification, was introduced to categorize distinct clinical–pathological entities[5]. However, within this classification system, various morphologic subtypes were unified into groups despite the suspicion that they "include more than one disease entity"[5].

Diffuse large B-cell lymphoma (DLBCL) is one disease in which attempts to define subgroups on the basis of morphology have largely failed owing to diagnostic discrepancies arising from inter- and intra-observer irreproducibility[5,6]. Diffuse large B-cell lymphoma 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: although most patients respond initially to chemotherapy, fewer than half of the patients achieve a durable remission[6,7]. 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[8].

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

[4] Present address: Life Sciences Division, Lawrence Orlando Berkeley National Labs and Department of Molecular and Cellular Biology, University of California, Berkeley, California 94720, USA.

mechanism that normally occurs only within the germinal centre of secondary lymphoid organs[9]. This evidence suggests that DLBCL arises either from germinal centre B cells or from B cells at a later stage of differentiation.

Here 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 malignancy; (2) identify distinct types of B-cell malignancy 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 whether gene expression profiling could subdivide this clinically heterogeneous diagnostic category into molecularly distinct diseases with more homogeneous clinical behaviours.

Construction of a specialized DNA microarray

Recent technical and analytical advances make it practical to quantitate the expression of thousands of genes in parallel using complementary DNA microarrays[10]. This mode of analysis has been used to observe gene expression variation in a variety of human tumours[11–17]. 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[18].

Because of the suspected importance of the germinal centre B cell to the genesis of non-Hodgkin's lymphomas, 12,069 out of the 17,856 cDNA clones on this microarray were chosen from a germinal centre B-cell library[18]. 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 leukaemia (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[19] 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. See Supplementary Information for the complete annotated list of these cDNAs.

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 leukaemia cell lines. Fluorescent cDNA probes, labelled with the Cy5 dye, were prepared from each experimental messenger RNA sample. A reference cDNA probe, labelled with the Cy3 dye, was prepared from a pool of mRNAs isolated from nine different lymphoma cell lines. Each Cy5-labelled experimental cDNA probe was combined with the Cy3-labelled 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, ~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 on the basis of similarity in the pattern with which their expression varied over all samples[20]. The same clustering method was used to group tumour and cell samples on the basis of 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 colour, with red representing expression greater than the mean, green representing expression less than the mean, and the colour intensity representing the magnitude of the deviation from the mean[20].

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 showing that the effects of experimental noise or artefact are negligible. Likewise, where different tumour samples from the same patient were analysed, they were invariably found clustered in immediately adjacent columns. For example, in three 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 leukaemic 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 tumours, rather than variation in handling or experimental artefacts. Moreover, these results show that even within a diagnostic category, each cancer patient has a unique tumour with a characteristic gene expression profile.

Figure 1 paints a complex, but 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. 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 gene expression patterns. Examination of the coordinately expressed genes in each of the B-cell malignancies and comparison with the normal lymphocyte cell populations yielded considerable insights into the biology of these malignancies. The coloured bars on the right of Fig. 1 indicate clusters of coordinately expressed genes that we operationally defined as gene expression 'signatures'. A gene expression signature was named by either the cell type in which its component genes were expressed (for example, the 'T-cell' signature) or the biological process in which its component genes are known to function (for example, 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 tumour phenotype

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, which reflects, in part, the fact that both of these malignancies are relatively indolent, with very low proliferation rates. Correspondingly, the genes that define the proliferation signature were not highly expressed in these malignancies (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 tumour biopsy, as previously noted[15]. In general, the more rapidly proliferating DLBCLs had higher expression of the genes in the proliferation signature. Nonetheless, marked differences in the expression of these genes were evident

between individual DLBCL samples, corresponding to the variability in proliferation index that has been previously observed in DLBCL[21].

The most prominent distinction between CLL and FL came from genes that are characteristic of germinal centre B cells (Fig. 2). An extensive cluster of genes distinguished germinal centre B cells from both resting blood B cells and *in vitro* activated blood B cells. This is remarkable because the stimuli used to activate the blood B cells were chosen to mimic those known to be important for germinal centre formation: crosslinking of the immunoglobulin receptor and CD40 signalling. However, it has thus far not been possible to mimic exactly the germinal centre phenotype *in vitro*, as determined by the failure of a variety of activation conditions to induce the expression of BCL-6 protein, a highly specific marker for germinal centre B cells[22]. The germinal centre B-cell gene expression signature shows that germinal centre 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 centre B cells was maintained in a cultured DLBCL cell line in the absence of the germinal centre microenvironment (Figs 1 and 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 centre microenvironment[23]. The gene expression signature of germinal centre B cells was reproduced virtually unchanged in FL, supporting the view that this lymphoma arises from this stage of B-cell differentiation (Fig. 2).

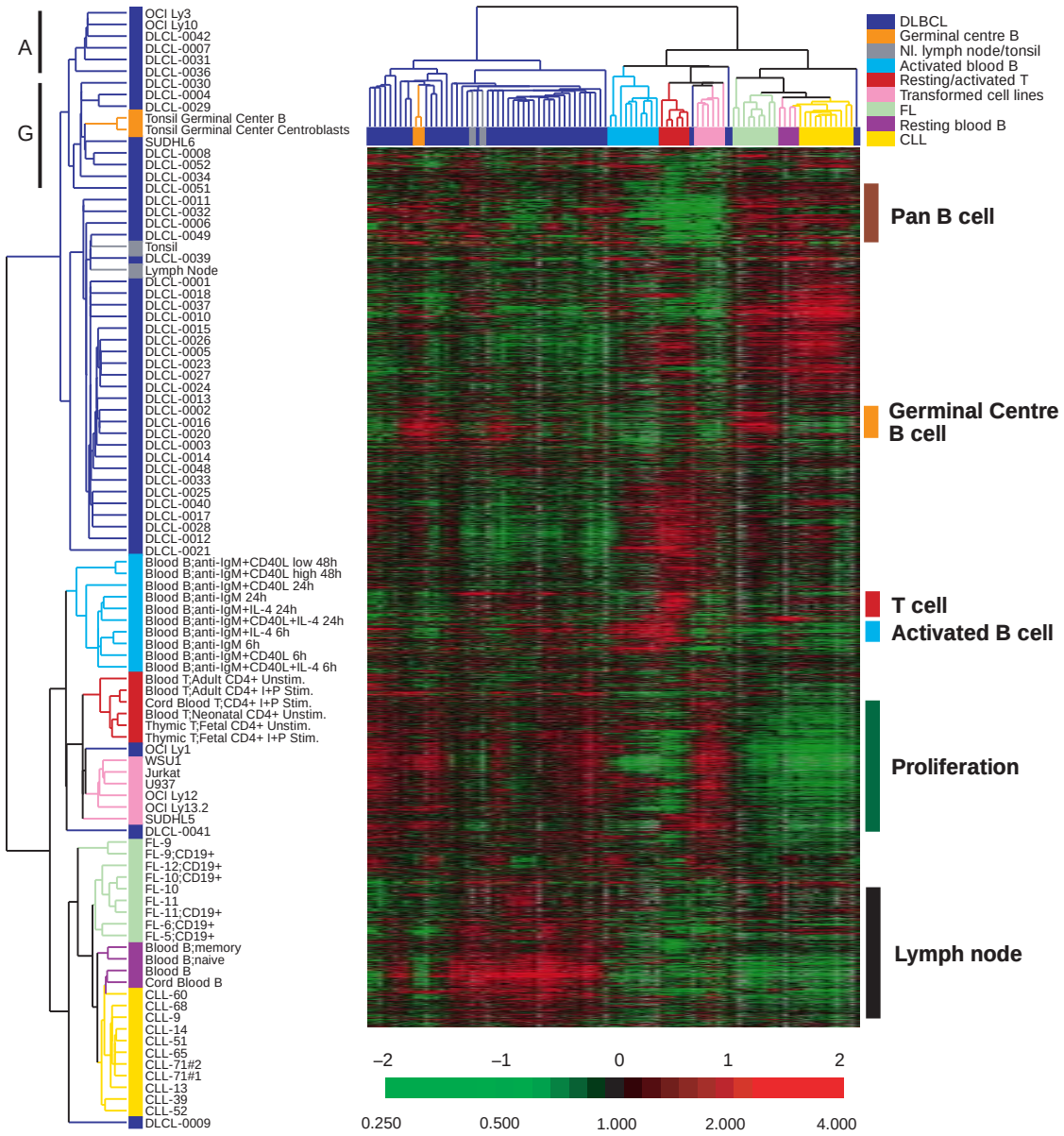

Figure 1 Hierarchical clustering of gene expression data. Depicted are the ~1.8 million measurements of gene expression from 128 microarray analyses of 96 samples of normal and malignant lymphocytes. The dendrogram at the left lists the samples studied and provides a measure of the relatedness of gene expression in each sample. The dendrogram is colour coded according to 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 colour 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 Supplementary Information for full data.

The gene expression profiles of DLBCLs were largely distinct from those of CLL and FL and showed additional biological complexity in these biopsy samples. Prominent features of the DLBCL profiles appeared to reflect the non-malignant cells in these tumours. A large group of genes defined a 'lymph-node' signature which 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 natural killer cells (NK4). In addition, genes involved in the remodelling of the extracellular matrix were abundantly expressed (MMP9 matrix metalloproteinase and TIMP-3). All but one DLBCL biopsy displayed the lymph-node signature, but the intensity of this signature varied, possibly reflecting the relative proportion of tumour and host cells in the lymph-node biopsy.

The variable presence of T lymphocytes in DLBCL biopsies was readily discernible by a T-cell gene expression signature that featured components of the T-cell receptor (TCR-β, CD3ϵ) and genes downstream of T-cell receptor signalling (fyn, LAT, PKC-θ) (Fig. 2). Although this T-cell expression signature was readily apparent in some DLBCLs, it was virtually undetectable in others.

Discovery of DLBCL subtypes

The structure of the hierarchical dendrogram in Fig. 1 indicated that gene expression patterns in DLBCLs might be inhomogeneous. Three branches of the dendrogram captured most of the DLBCLs with only three outlying samples. Clearly, the position of any given DLBCL sample in the dendrogram is determined in a complicated fashion by the influences of several distinct biological themes that are reflected in the expression pattern. Inspection of the gene expression map shown in Fig. 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 centre B cells from other stages in B-cell ontogeny were also differentially expressed among DLBCLs, suggesting that B-cell differentiation genes may also be used to subdivide DLBCL. The expression of the germinal centre B cell genes among DLBCLs varied independently from the expression of genes in the other gene expression signatures (Fig. 2; see Supplementary Information for details). In principle, each of these gene expression

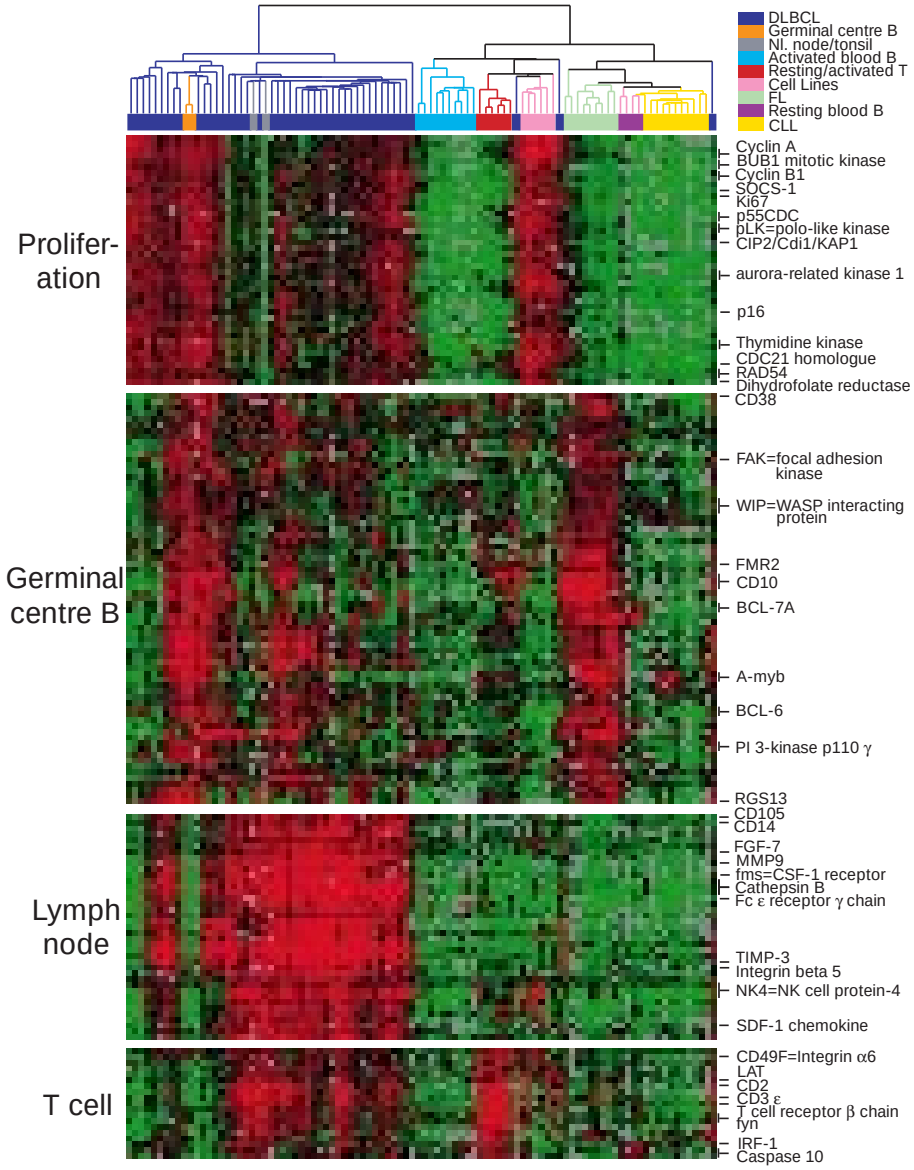

Figure 2 Expanded view of biologically distinct gene expression signatures defined by hierarchical clustering. Data are the same as in Fig. 1. Most genes without designations on the right are new genes of unknown function derived from various lymphoid cDNA libraries.

signatures could be used to define subsets of DLBCL. We decided to focus our attention initially on the germinal centre B-cell genes, however, because we suspected that these genes might identify DLBCL cases that were derived from distinct stages of normal B-cell differentiation. Indeed, the clustering of the germinal centre B-cell samples with a subset of the DLBCLs in a major branch of the dendrogram in Fig. 1 suggested that this group of DLBCLs might resemble normal germinal centre B cells.

To test this hypothesis, we reclustered the DLBCL cases using only the expression pattern of the genes that define the germinal centre 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 Fig. 1: activated B-like DLBCL includes all cases in the branch labelled 'A', and GC B-like DLBCL includes all cases in branch labelled 'G'. The largest DLBCL branch in Fig. 1 is a mixture of the cases assigned to the two subgroups. Normal germinal centre 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 centre B-cell signature. In contrast, the activated B-like DLBCL group expressed these genes at low or undetectable levels, for the most part. The gene expression subgroups defined here were not obviously related to histological subtypes of DLBCL: only two of the cases studied could be assigned to the immunoblastic histological subtype, according to the revised Kiel classification system. Furthermore, no evidence of normal germinal centres was found in the lymph-node biopsies. Indeed, one of the germinal centre B-cell markers described below, CD10, was expressed by the lymphoma cells using immunohistochemistry (data not shown). These data clearly suggested that a distinct class of DLBCLs was derived from the germinal centre B cell and retained the gene expression program, and presumably many of the phenotypic characteristics, of this stage of B-cell differentiation.

We 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 tumours. We used hierarchical clustering to reorder this set of 2,984 genes while maintaining the order shown in Fig. 3a of the DLBCL cases (Fig. 3b). As is evident in Fig. 3c, a cluster of genes could be recognized on the basis of their elevated expression in the activated B-like DLBCLs, as compared with 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.

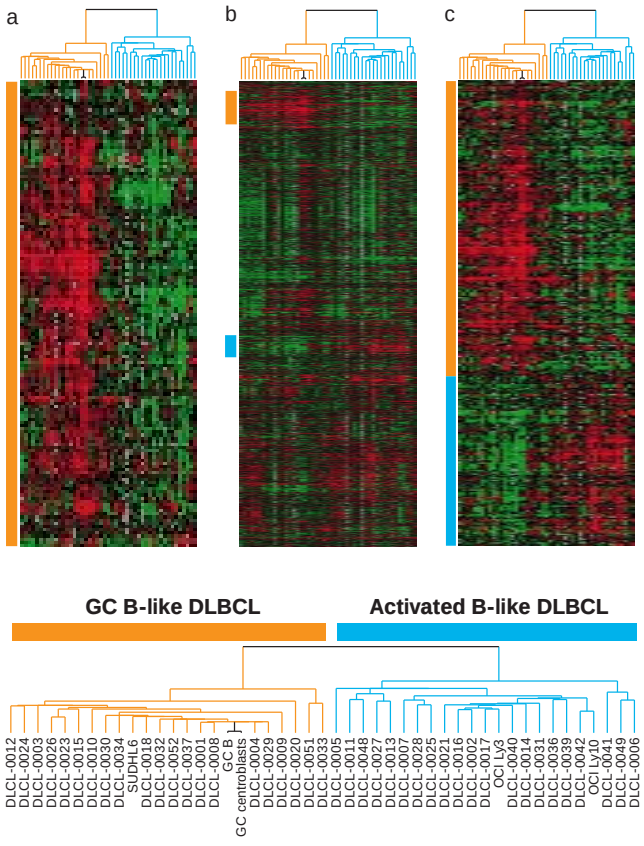

Figure 3 Discovery of DLBCL subtypes by gene expression profiling. The samples used in this clustering analysis are shown at the bottom. **a**, Hierarchical clustering of DLBCL cases (blue and orange) and germinal centre B cells (black) based on the genes of the germinal centre B-cell gene expression 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 hierarchical 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**, Hierarchical clustering of the genes selectively expressed in GC B-like DLBCL and activated B-like DLBCL, which was determined from Fig. 3b.

DLBCL subgroups and B-cell differentiation

We examined how all of the genes that distinguish these DLBCL subgroups are expressed during B-cell differentiation and activation. Figure 4 shows that almost all of the genes that defined GC B-like DLBCL were highly expressed in normal germinal centre 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 centre 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 centre B cells relative to resting or activated peripheral blood B cells were expressed by GC B-like DLBCL (data not shown).

By contrast, most of the genes that defined activated B-like DLBCL were not expressed in normal germinal centre 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 h of activation and others only expressed after 48 h 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. Notably, two DLBCL cell lines, OCI Ly3 and OCI Ly10, were among the activated B-like DLBCLs. In fact, one or both of these two 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 centre differentiation (for example, the genes encoding the cell-surface proteins CD10 and CD38 (ref. 24), the nuclear factor A-myb (ref. 25) and the DNA repair protein 8-oxoguanine DNA glycosylase (OGG1)[26]) and a host of new genes. A particularly noteworthy gene in the GC B-like DLBCL signature is BCL-6, a well-established

germinal centre marker that is also the most frequently translocated gene in DLBCL[22]. Although BCL-6 protein expression is invariably detected in DLBCL, its levels vary and are not correlated with the presence of BCL-6 translocations[27,28]. Cytogenetic data are available for 16 out 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 probably related to their derivation from germinal centre 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 centre 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[29]. The specific expression of BCL-7A in germinal centre B cells has strong parallels with BCL-6. BCL-6 is required for germinal centre formation during an antigen-driven immune response[30–32] and is translocated in B-cell malignancies that derive from germinal centre B cells. Given the preferential expression of BCL-7A in germinal centre B cells, it is conceivable that this gene is also involved in normal germinal centre physiology and in the pathophysiology of GC B-like DLBCL. LMO2 is translocated and overexpressed in a subset of T-cell acute lymphoblastic leukaemias

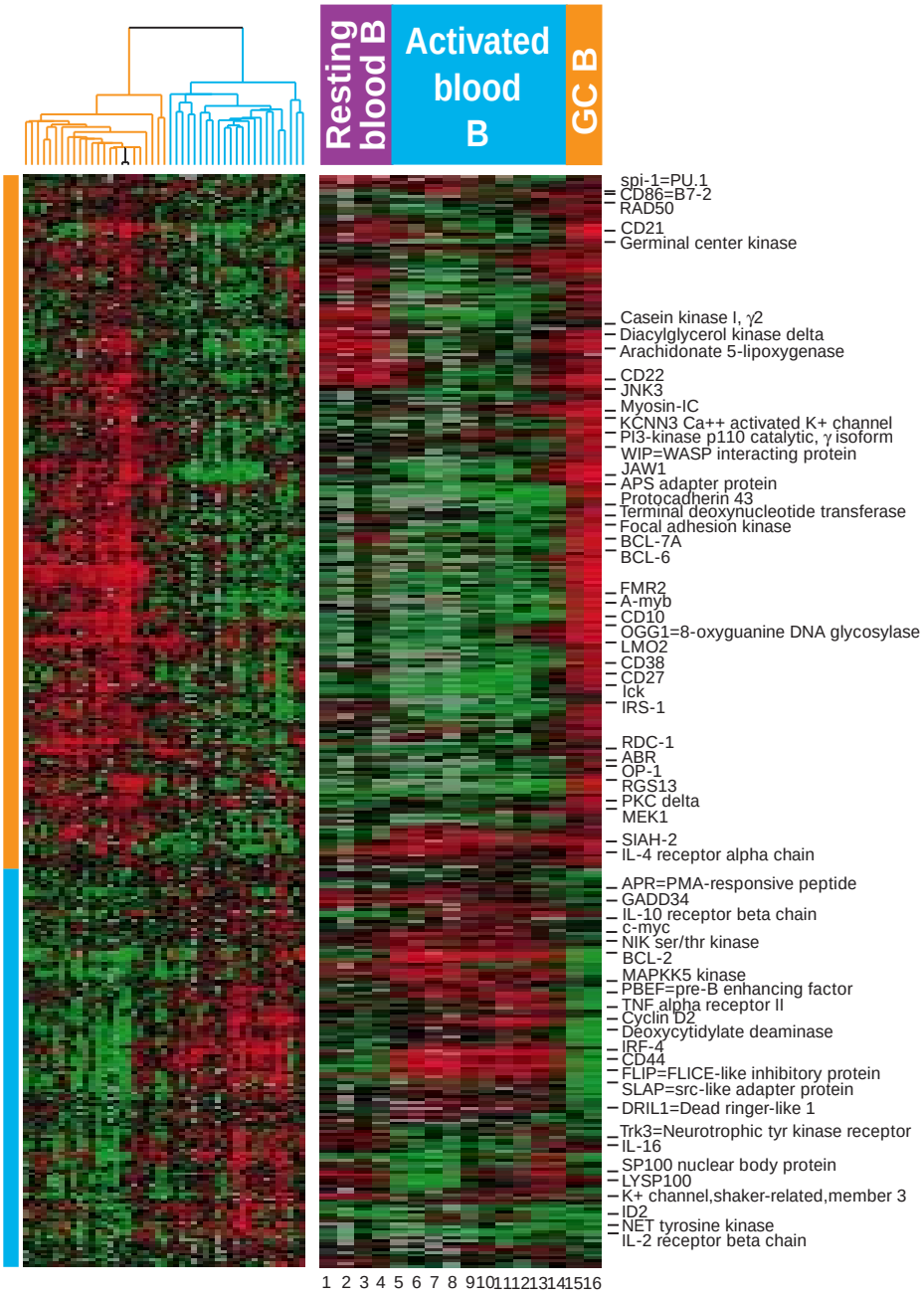

Figure 4 Relationship of DLBCL subgroups to normal B-lymphocyte differentiation and activation. The data in the left panel are taken from Fig. 3c. The right panel depicts gene expression data from the following normal B-cell samples: (1) Total CD19$^+$ blood B cells; (2) Naive CD27$^-$ blood B cells; (3) Memory CD27$^+$ blood B cells; (4) cord blood CD19$^+$ B cells; (5) blood B cells; anti-IgM 6 h; (6) blood B cells; anti-IgM + IL-4 6 h; (7) blood B cells; anti-IgM + CD40 ligand 6 h; (8) blood B cells; anti-IgM + CD40 ligand + IL-4 6 h; (9) blood B cells; anti-IgM 24 h; (10) blood B cells; anti-IgM + IL-4 24 h; (11) blood B cells; anti-IgM + CD40 ligand 24 h; (12) blood B cells; anti-IgM + CD40 ligand + IL-4 24 h; (13) blood B cells; anti-IgM + CD40 ligand (low concentration) 48 h; (14) blood B cells; anti-IgM + CD40 ligand (high concentration) 48 h; (15) tonsil germinal centre B cells; (16) tonsil germinal centre centroblasts. See Supplementary Information for full data.

and LMO2 transgenic mice have a block in early T-cell differentiation and develop T-cell leukaemia[33]. The selective expression of LMO2 in germinal centre B cells indicates that LMO2 may have a 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 that is 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*[34]. IRF4 is transiently induced during normal lymphocyte activation[35] (Fig. 4) and is critical for the proliferation of B lymphocytes in response to signals from the antigen receptor[36]. Thus, the constitutive expression of IRF4 in activated B-like DLBCLs may contribute to the unchecked proliferation of the malignant cells in these tumours.

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[37]. 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 tumour types, and its constitutive expression in activated B-like DLBCLs could inhibit apoptosis of tumour cells induced by host T cells expressing Fas ligand[38,39]. 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 centre B cells but is induced more than 30-fold during activation of peripheral blood B cells (Fig. 4). Most activated B-like DLBCLs (71%) had BCL-2 mRNA levels more than fourfold higher than were observed in germinal centre 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, indicating that BCL-2 may also be important in some cases of this DLBCL subgroup.

DLBCL gene expression subgroups define prognostic categories

Does the taxonomy of DLBCL derived from gene expression patterns define clinically distinct subgroups of patients? None of the patients included in this study had been treated before obtaining the biopsy sample. Furthermore, these patients were '*de novo*' DLBCL cases that had not obviously arisen from pre-existing low-grade malignancies such as follicular lymphoma. After biopsy, the patients were treated at two medical centres using comparable, standard multi-agent chemotherapy regimens. Figure 5a presents a Kaplan–Meier plot of overall survival data from these patients, segregated according to gene expression subgroup. Germinal centre 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). Although the average five-year survival for all patients was 52%, 76% of GC B-like DLBCL patients were still alive after five years, as compared with only 16% of activated B-like DLBCL patients. The differential survival of patients in the two DLBCL subgroups was apparently uninfluenced by the anthracyline-based chemotherapeutic regimen used (data not shown), which is not surprising as responses of DLBCL patients to various multi-agent chemotherapeutic regimens were found to be equivalent[40]. Thus, the molecular differences between these two kinds of lymphoma were accompanied by a remarkable divergence in clinical behaviour, 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[8]. This indicator takes into account the patient's age, performance status, and the extent and location of disease. As suspected, within our patient population a low IPI score (0–2) identified patients with better overall survival as compared with patients with a high IPI score (3–5) (Fig. 5b). We then determined 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, as judged by 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 that influence their survival.

Conclusions

This study shows 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. A number of individual markers have been used to define subsets of DLBCL[41–46], but these studies do not provide the present overview that strongly implies that this single diagnostic category of lymphoma harbours 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

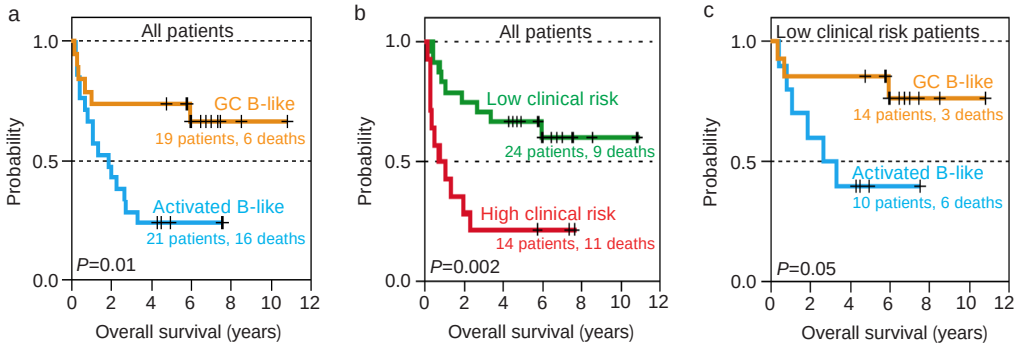

Figure 5 Clinically distinct DLBCL subgroups defined by gene expression profiling. **a**, Kaplan–Meier plot of overall survival of DLBCL patients grouped on the basis of gene expression profiling. **b**, Kaplan–Meier plot of overall survival of DLBCL patients grouped according to the International Prognostic Index (IPI). Low clinical risk patients (IPI score 0–2) and high clinical risk patients (IPI score 3–5) are plotted separately. **c**, Kaplan–Meier plot of overall survival of low clinical risk DLBCL patients (IPI score 0–2) grouped on the basis of their gene expression profiles.

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[5], 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 after treatment with anthracycline-based multi-agent chemotherapy regimens. It is unclear at present which of the genes that distinguish GC B-like DLBCL from activated B-like DLBCL are the most important molecular determinants of chemotherapy responsiveness. Furthermore, there is residual clinical heterogeneity which cannot be explained by the current classification. Despite the fact that patients with GC B-like DLBCL had an overall favourable prognosis, five patients died within the first two years of diagnosis. Likewise, three patients in the activated B-like DLBCL subgroup were alive five years after 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 probably, however, a multivariate approach to prognosis will be needed that 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 multi-agent chemotherapy, and then, if a complete remission cannot be maintained, patients are considered for bone marrow transplantation[7]. 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 we present here, will identify a small number of marker genes that will be used to stratify patients into molecularly relevant categories which 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 new cancer therapeutics that could be based on a molecular understanding of the cancer phenotype. Our study shows that the two DLBCL subgroups differentially expressed entire transcriptional modules composed of hundreds of genes, many of which could be expected to contribute to the malignant behaviour of the tumour. 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

Messenger RNA samples

Total germinal centre B cells and centroblasts were purified from human tonsils as described[24]. 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[47,48]. Magnetic cell sorting was used to purify $CD4^+$, $CD45RA^{high}$ T cells from human cord blood or adult peripheral blood and $CD4^+$ thymocytes from human fetal thymus (Milteni Biotec). All lymphocyte samples were purified to more than 98% homogeneity as determined by FACS analysis. For rare lymphoid subpopulations such as centroblasts or resting and naive peripheral blood B cells, purified samples from multiple donors were pooled for microarray analysis. *In vitro* stimulation of peripheral B cells was done as described[49] using anti-IgM antibody, IL-4 and/or CD40 ligand-containing membranes. Most experiments used a 1:1000 dilution of CD40 ligand membranes (designated 'low' concentration, Figs 1 and 4) but one experiment used a 1:200 dilution (designated 'high' concentration, Figs 1 and 4). T cells were stimulated for 2 h with phorbol ester (50 ng ml^{-1}) and ionomycin (1.5 μM). 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 ($n = 34$) or Stanford University School of Medicine ($n = 8$) using comparable, anthracycline-based, multi-agent chemotherapeutic regimens with curative intent. Clinical data were not available on two of the DLBCL cases presented in Fig. 1 (DLCL-51 and DLCL-52). For two additional patients (DLCL-25 and DLCL-36), the data needed to calculate the IPI were not available. DLBCL and FL lymph-node biopsies were either snap frozen, frozen in OCT or disaggregated and frozen as a viable cell suspension. Chronic lymphocyte leukaemia cells were purified from untreated patients by magnetic selection for $CD19^+$ cells (Miltenyi Biotec).

Microarray procedures

DNA microarray analysis of gene expression was done essentially as described[50]. The cDNA clones on the Lymphochip microarray are listed in Supplementary Information and are available from Research Genetics. Fluorescent images of hybridized microarrays were obtained using a GenePix 4000 microarray scanner (Axon Instruments). Images were analysed with ScanAlyze (M. 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. Raw data files for each array containing all measured values and manual flags are available in Supplementary Information. A set of clones that consistently behaved poorly across arrays was identified and excluded from all analyses (see Supplementary Information). Fluorescence ratios were calibrated independently for each array by applying a single 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-labelled) and a control mRNA sample (Cy3-labelled) isolated from a pool of nine 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[18].

Data analysis

All non-flagged array elements for which the fluorescent intensity in each channel was greater than 1.4 times the local background were considered well measured. The ratio values were log-transformed (base 2) and stored in a table (rows, individual cDNA clones; columns, single mRNA samples). Where samples had been analysed 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 96 mRNA samples were excluded. Data for the remaining genes were centred by subtracting (in log space) the median observed value, to remove any effect of the amount of RNA in the reference pool. This dataset contains 4,026 array elements (see Supplementary Information). Hierarchical clustering was applied to both axes using the weighted pair-group method with centroid average as implemented in the program Cluster (M. Eisen; http://www.microarrays.org/software)[20]. The distance matrixes used were Pearson correlation for clustering the arrays and the inner product of vectors normalized to magnitude 1 for the genes (this is a slight variant of Pearson correlation; see Cluster manual available at http://www.microarrays.org/software/ for computational details). The results were analysed with Tree View (M. Eisen; http://www.microarrays.org/software)[20]. All datasets and image files used to generate Figs 1–4 are included in the Supplementary Information, along with numerous supplementary and additional analyses.

Received 11 November 1999; accepted 10 January 2000.

1. Hodgkin, T. On some morbid appearances of the absorbant glands and spleen. *Med.-Chir. Trans.* **17,** 68–114 (1832).
2. Sternberg, C. Über eine eigenartige unter dem Bilde der Pseudoleukamie verlaufende Tuberculose des lymphatischen Apparates. *Heilk* **19,** 21–90 (1898).
3. Reed, D. M. On the pathological changes in Hodgkin's disease, with especial reference to its relation to tuberculosis. *Johns Hopkins Hosp. Rep.* **10,** 133–196 (1902).
4. Rosenberg, S. A. Classification of lymphoid neoplasms. *Blood* **84,** 1359–1360 (1994).
5. Harris, N. L. *et al.* A revised European–American classification of lymphoid neoplasms: a proposal from the International Lymphoma Study Group. *Blood* **84,** 1361–1392 (1994).
6. The Non-Hodgkin's Lymphoma Classification Project: A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. *Blood* **89,** 3909–3918 (1997).
7. Vose, J. M. Current approaches to the management of non-Hodgkin's lymphoma. *Semin. Oncol.* **25,** 483–491 (1998).
8. The International Non-Hodgkin's Lymphoma Prognostic Factors Project: A predictive model for aggressive non-Hodgkin's lymphoma. *N. Engl. J. Med.* **329,** 987–994 (1993).
9. Klein, U. *et al.* Somatic hypermutation in normal and transformed human B cells. *Immunol. Rev.* **162,** 261–280 (1998).

10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
11. Bubendorf, L. *et al.* Hormone therapy failure in human prostate cancer: analysis by complementary DNA and tissue microarrays. *J. Natl Cancer Inst.* **91,** 1758–1764 (1999).
12. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229,** 101–108 (1999).
13. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
14. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58,** 5009–5013 (1998).
15. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96,** 9212–217 (1999).
16. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14,** 457–460 (1996).
17. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96,** 6745–6750 (1999).
18. Alizadeh, A. *et al.* The Lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harbor Symp. Quant. Biol.* (in the press).
19. Alizadeh, A., Eisen, M., Botstein, D., Brown, P. O. & Staudt, L. M. Probing lymphocyte biology by genomic-scale gene expression analysis. *J. Clin. Immunol.* **18,** 373–379 (1998).
20. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
21. Grogan, T. M. *et al.* Independent prognostic significance of a nuclear proliferation antigen in diffuse large cell lymphomas as determined by the monoclonal antibody Ki-67. *Blood* **71,** 1157–1160 (1988).
22. Staudt, L. M., Dent, A. L., Shaffer, A. L. & Yu, X. Regulation of lymphocyte cell fate decisions and lymphomagenesis by BCL-6. *Int. J. Immunol.* **18,** 381–403 (1999).
23. Bahler, D. W. & Levy, R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. *Proc. Natl Acad. Sci. USA* **89,** 6770–6774 (1992).
24. Liu, Y. -J. & Banchereau, J. in *Handbook of Experimental Immunology* (eds Weir, D., Blackwell, C., Herzenberg, L. & Herzenberg, L.) 93.1–93.9 (Blackwell Scientific, Oxford, 1996).
25. Golay, J., Erba, E., Bernasconi, S., Peri, G. & Introna, M. The A-myb gene is preferentially expressed in tonsillar CD38+, CD39–, and sIgM– B lymphocytes and in Burkitt's lymphoma cell lines. *J. Immunol.* **153,** 543–553 (1994).
26. Kuo, F. C. & Sklar, J. Augmented expression of a human gene for 8-oxoguanine DNA glycosylase (MutM) in B lymphocytes of the dark zone in lymph node germinal centers. *J. Exp. Med.* **186,** 1547–1556 (1997).
27. Flenghi, L. *et al.* A specific monoclonal antibody (PG-B6) detects expression of the BCL-6 protein in germinal center B cells. *Am. J. Pathol.* **147,** 405–411 (1995).
28. Pittaluga, S. *et al.* BCL-6 expression in reactive lymphoid tissue and in B-cell non-Hodgkin's lymphomas. *J. Pathol.* **179,** 145–150 (1996).
29. Zani, V. J. *et al.* Molecular cloning of complex chromosomal translocation t(8;14;12)(q24. 1;q32. 3;q24. 1) in a Burkitt lymphoma cell line defines a new gene (BCL7A) with homology to caldesmon. *Blood* **87,** 3124–3134 (1996).
30. Fukuda, T. *et al.* Disruption of the Bcl6 gene results in an impaired germinal center formation. *J. Exp. Med.* **186,** 439–448 (1997).
31. Ye, B. H. *et al.* The BCL-6 proto-oncogene controls germinal-centre formation and Th2-type inflammation. *Nature Genet.* **16,** 161–170 (1997).
32. Dent, A. L., Shaffer, A. L., Yu, X., Allman, D. & Staudt, L. M. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. *Science* **276,** 589–592 (1997).
33. Rabbitts, T. H. LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes Dev.* **12,** 2651–2657 (1998).
34. Iida, S. *et al.* Deregulation of MUM1/IRF4 by chromosomal translocation in multiple myeloma. *Nature Genet.* **17,** 226–230 (1997).
35. Matsuyama, T. *et al.* Molecular cloning of LSIRF, a lymphoid-specific member of the interferon regulatory factor family that binds the interferon-stimulated response element (ISRE). *Nucleic Acids Res.* **23,** 2127–2136 (1995).
36. Mittrucker, H. W. *et al.* Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. *Science* **275,** 540–543 (1997).
37. Tschopp, J., Irmler, M. & Thome, M. Inhibition of fas death signals by FLIPs. *Curr. Opin. Immunol.* **10,** 552–558 (1998).
38. Djerbi, M. *et al.* The inhibitor of death receptor signaling, FLICE-inhibitory protein defines a new class of tumor progression factors. *J. Exp. Med.* **190,** 1025–1031 (1999).
39. Medema, J. P., de Jong, J., van Hall, T., Melief, C. J. M. & Offringa, R. Immune escape of tumors *in vivo* by expression of cellular FLICE-inhibitory protein. *J. Exp. Med.* **190,** 1033–1038 (1999).
40. Fisher, R. I. *et al.* Comparison of a standard regimen (CHOP) with three intensive chemotherapy regimens for advanced non-Hodgkin's lymphoma. *N. Engl. J. Med.* **328,** 1002–1006 (1993).
41. Jalkanen, S., Joensuu, H., Soderstrom, K. O. & Klemi, P. Lymphocyte homing and clinical behavior of non-Hodgkin's lymphoma. *J. Clin. Invest.* **87,** 1835–1840 (1991).
42. Harada, S. *et al.* Molecular and immunological dissection of diffuse large B cell lymphoma: CD5–, and CD5– with CD10+ groups may constitute clinically relevant subtypes. *Leukemia* **13,** 1441–1447 (1999).
43. Kramer, M. H. *et al.* Clinical significance of bcl2 and p53 protein expression in diffuse large B-cell lymphoma: a population-based study. *J. Clin. Oncol.* **14,** 2131–2138 (1996).
44. Preti, H. A. *et al.* Prognostic value of serum interleukin-6 in diffuse large-cell lymphoma. *Ann. Int. Med.* **127,** 186–194 (1997).
45. Gascoyne, R. D. *et al.* Prognostic significance of Bcl-2 protein expression and Bcl-2 gene rearrangement in diffuse aggressive non-Hodgkin's lymphoma. *Blood* **90,** 244–251 (1997).
46. Kramer, M. H. *et al.* Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large B-cell lymphoma. *Blood* **92,** 3152–3162 (1998).
47. Klein, U., Rajewsky, K. & Kuppers, R. Human immunoglobulin (Ig)M+IgD+ peripheral blood B cells expressing the CD27 cell surface antigen carry somatically mutated variable region genes: CD27 as a general marker for somatically mutated (memory) B cells. *J. Exp. Med.* **188,** 1679–1689 (1998).
48. Tangye, S. G., Liu, Y. J., Aversa, G., Phillips, J. H. & de Vries, J. E. Identification of functional human splenic memory B cells by expression of CD148 and CD27. *J. Exp. Med.* **188,** 1691–1703 (1998).
49. Allman, D. *et al.* BCL-6 expression during B-cell activation. *Blood* **87,** 5257–5268 (1996).
50. Eisen, M. B. & Brown, P. O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303,** 179–205 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com), on the authors' World-Wide Web site (http://llmpp.nih.gov/lymphoma) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We acknowledge the support of the Cancer Genome Anatomy Project (CGAP), led by B. Strausberg and R. Klausner. We also thank R. Klausner for comments on the manuscript; C. Prange for providing CGAP cDNA clones; H. Messner for providing DLBCL cell lines; H. Mostowski for sorting lymphocyte subpopulations by FACS; Holy Cross Hospital, Silver Spring, Maryland, for providing human tonsils; J. 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. and by the Howard Hughes Medical Institute. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute. A.A. was initially supported by the Howard Hughes Medical Institute Research Scholar Program while at the NIH and then 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.

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).