Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning

MARGARET A. SHIPP[1], KEN N. ROSS[2], PABLO TAMAYO[2], ANDREW P. WENG[3], JEFFERY L. KUTOK[3], RICARDO C.T. AGUIAR[1], MICHELLE GAASENBEEK[2], MICHAEL ANGELO[2], MICHAEL REICH[2], GERALDINE S. PINKUS[3], TANE S. RAY[6], MARGARET A. KOVAL[1], KIM W. LAST[4], ANDREW NORTON[5], T. ANDREW LISTER[4], JILL MESIROV[2], DONNA S. NEUBERG[1], ERIC S. LANDER[2,7], JON C. ASTER[3] & TODD R. GOLUB[1,2]

[1]*Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA*
[2]*Whitehead Institute for Biomedical Research/Massachusetts Institute of Technology Center for Genome Research, Cambridge, Massachusetts, USA*
[3]*Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA*
[4]*ICRF Medical Oncology Unit and* [5]*Pathology Unit, St. Bartholomew's Hospital, London, UK*
[6]*Department of Computer Science, Maths and Physics, University of West Indies, Bridgetown, Barbados*
[7]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA*
K.N.R. and P.T. contributed equally to this study.
Correspondence should be addressed to M.A.S.; email: margaret_shipp@dfci.harvard.edu, or T.R.G.; email: golub@genome.wi.mit.edu

Diffuse large B-cell lymphoma (DLBCL), the most common lymphoid malignancy in adults, is curable in less than 50% of patients. Prognostic models based on pre-treatment characteristics, such as the International Prognostic Index (IPI), are currently used to predict outcome in DLBCL. However, clinical outcome models identify neither the molecular basis of clinical heterogeneity, nor specific therapeutic targets. We analyzed the expression of 6,817 genes in diagnostic tumor specimens from DLBCL patients who received cyclophosphamide, adriamycin, vincristine and prednisone (CHOP)-based chemotherapy, and applied a supervised learning prediction method to identify cured versus fatal or refractory disease. The algorithm classified two categories of patients with very different five-year overall survival rates (70% versus 12%). The model also effectively delineated patients within specific IPI risk categories who were likely to be cured or to die of their disease. Genes implicated in DLBCL outcome included some that regulate responses to B-cell–receptor signaling, critical serine/threonine phosphorylation pathways and apoptosis. Our data indicate that supervised learning classification techniques can predict outcome in DLBCL and identify rational targets for intervention.

Diffuse large B-cell lymphomas (DLBCLs) are the most common lymphoid neoplasms, composing 30–40% of adult non-Hodgkin lymphomas[1]. Although a subset of DLBCL patients are cured with current chemotherapeutic regimens, most succumb to the disease[2]. Clinical prognostic models such as the International Prognostic Index (IPI) have been developed to identify DLBCL patients who are unlikely to be cured with standard therapy[3]. However, the clinical factors of the IPI (age, performance status, stage, number of extranodal sites and serum lactate dehydrogenase (LDH))[3] are likely to be surrogate markers for the intrinsic molecular heterogeneity in this disease. Therefore, it is not surprising that IPI is imperfect in its identification of high-risk patients. In addition, in the absence of molecular insights into the clinical heterogeneity of DLBCL, therapeutic approaches to high-risk patients have primarily included increased doses of conventional chemotherapeutic agents and additional stem-cell support[4]. However, the value of high-dose therapy has not been confirmed in this setting[4], underscoring the need to identify more rational, molecularly defined approaches to treatment.

Molecular analyses of clinical heterogeneity in DLBCL have largely focused on individual candidate genes, with particular emphasis on genes with known functions in other malignancies or in normal lymphocyte development. Examples include adhesion molecules that influence the trafficking of normal activated B cells and tumor cells[5,6], proteins that regulate apoptosis in other B-cell lymphomas and normal B-cell subpopulations[7–9], and angiogenic peptides that promote the development of an effective tumor vasculature[10]. Additional individual genes, such as *BAL* (B-aggressive lymphoma), have been identified on the basis of their differential expression in fatal high-risk DLBCL and cured low-risk tumors[11]. Although some of these candidate genes correlate with DLBCL treatment outcome, a comprehensive molecular approach to outcome prediction is still lacking.

The recent development of DNA microarrays provides an opportunity to take a genome-wide approach to predicting DLBCL treatment outcome. One strategy is to use gene-expression profiling to extend current biological insights into the disease. Such an approach was recently described by Alizadeh *et al.*, who built on the hypothesis that DLBCL derives from normal B cells located within the germinal centers (GCs) of lymphoid organs[12,13]. Customized cDNA ('lymphochip') microarrays enriched in genes related to the GCs were used to obtain the gene-expression pat-

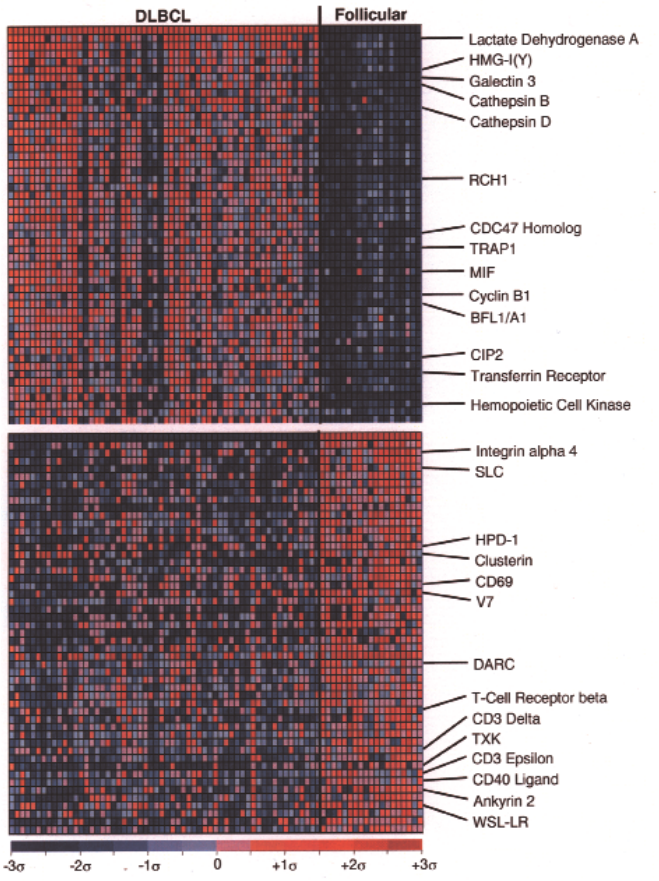

Fig. 1 Expression profiles of DLBCLs and FLs. The genes that were expressed at higher levels in DLBCL are shown on top, the ones which were more abundant in FL, on bottom. Red indicates high relative expression; blue, low expression. Color scale at bottom indicates relative expression in standard deviations from the mean. Each column is a sample, each row is a gene. Expression profiles of the 58 DLBCLs are on the left (58 columns); profiles of the 19 FLs are on the right (19 columns).

terns of DLBCL and normal lymphocytes, including B cells from GC B cells and *in vitro*-activated peripheral blood (PB) B cells. Using the unsupervised learning technique of hierarchical clustering, Alizadeh *et al.*[12] demonstrated that the DLBCLs fell into two groups: those with expression patterns similar to normal GC B cells, and those with expression patterns similar to *in vitro*-activated PB B cells. Alizadeh *et al.*[12] found the GC-like DLBCLs to have a more favorable outcome compared with the PB-like DLBCLs, suggesting that putative cell of origin might be predictive of response to treatment in this disease.

An alternative strategy for the prediction of DLBCL outcome is to use supervised learning methods to directly develop a gene-expression–based outcome model that is independent of *a priori* hypotheses. Here we report the successful prediction of outcome in a series of 58 DLBCL patients using gene-expression data from oligonucleotide microarrays together with supervised learning methods. Notably, this supervised approach identifies molecular correlates of outcome that are independent of the previously described putative cell of origin[12].

Delineating DLBCL from follicular lymphoma

We have described a supervised learning classification algorithm ('weighted voting') which delineated acute leukemias that arise from different lineages (lymphoid versus myeloid)[14]. Before attempting to apply this method to distinguish cured versus fatal/refractory DLBCLs, we investigated whether the algorithm could identify tumors within a single (B-cell) lineage. Specifically, we asked whether we could distinguish DLBCL from a related GC B-cell lymphoma, follicular lymphoma (FL). Although these two malignancies have very different clinical presentations, natural histories and responses to therapy[1,2], FLs frequently evolve over time and acquire the morphologic and clinical features of DLBCLs. In addition, a subset of *de novo* DLBCLs have the t(14;18) chromosomal translocation characteristic of most FLs (ref. 7). The t(14;18) results in overexpression of the anti-apoptotic protein BCL2 (ref. 15); however, the mechanism by which most DLBCLs circumvent normal apoptotic signals is not known.

Pre-treatment biopsies obtained from 77 patients with DLBCL (n = 58) or FL (n = 19) were subjected to transcriptional profiling using oligonucleotide microarrays containing probes for 6,817 genes. The gene-expression data are available in their entirety in Supplementary Information (www.genome.wi.mit.edu/MPR/lymphoma). The 6,817 genes were sorted by their degree of correlation with the DLBCL versus FL distinction, and the most highly correlated genes are shown in Fig. 1. Genes expressed at higher levels in DLBCL patients than in FL patients included known DLBCL markers such as lactate dehydrogenase[3] and transferrin receptor (Fig. 1). Genes associated with cellular proliferation (cyclin B1 and a CDC47 homolog) and invasion and metastasis (cathepsins B and D) were also expressed at higher levels in DLBCLs versus FLs. DLBCLs also overexpressed: 1) the high-mobility group protein isoforms I and Y (HMGIY), known to be a MYC target and encoded by a potential oncogene[16]; 2) the hematopoietic cell kinase (HCK) which has been linked with CD44 signaling[17]; and 3) inhibitors of apoptosis such as the carbohydrate-binding protein, galectin 3 (ref. 18) and the B-cell lymphoma-2 (BCL2)-related protein, BFL1A1 (ref. 19; also known as BCL2A1).

BFL1A1 overexpression in DLBCL is of particular interest because this anti-apoptotic molecule is induced by CD40 signaling and is required for CD40-mediated B-cell survival[20]. BFL1A1 is also a direct transcriptional target of nuclear factor-κB (NF-κB), which suppresses both chemotherapy- and tumor necrosis factor–associated apoptosis[21,22]. These observations raise the possibility that BFL1A1 overexpression may represent an important anti-apoptotic mechanism for reducing the chemosensitivity of DLBCLs.

Genes overexpressed in FLs compared with DLBCLs included additional regulators of apoptosis such as human programmed death-1 (*HPD1*)[23] and *WSL-LR* (also known as *TNFRSF12*)[24]. FLs also had more abundant expression of genes encoding cytoskeletal components (ankyrin 2) and adhesion molecules (α_4 integrin) and genes expressed by follicular dendritic cells (clusterin[25]) and infiltrating T cells (T-cell receptor-β, CD3-ε, CD3-δ, CD40 ligand, TXK tyrosine kinase[26], T-cell activation antigens, CD69 (ref. 27) and V7 (ref. 28)) and the T-cell chemoattractant, SLC (ref. 29; also known as SCYA21). The presence of a prominent T-cell and follicular dendritic-cell signature in the FLs also indicates that microarray profiling can be used to capture additional non-malignant components of the tumor microenvironment. This non-malignant component of the FL versus DLBCL signature would have been missed had purified tumor cells, rather than primary tumor specimens, been analyzed.

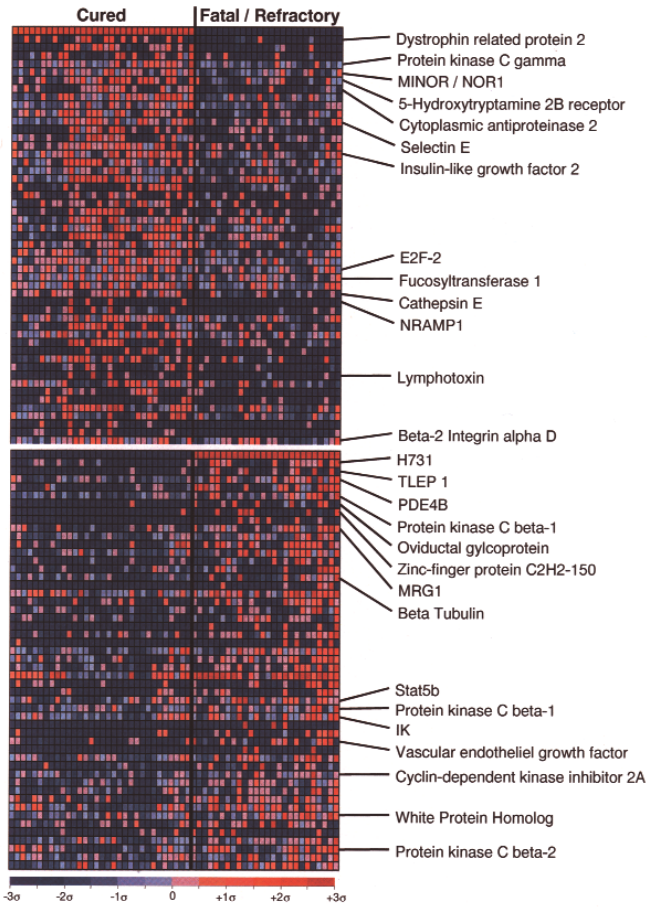

Fig. 2 Expression profiles of cured and fatal/refractory DLBCLs. The genes that were expressed at higher levels in cured disease are shown on top, those that were more abundant in fatal disease are shown on bottom. Red indicates high level expression; blue, low level expression. Color scale at bottom indicates relative expression in standard deviations from the mean. Each column is a sample, each row a gene. Expression profiles of the 32 cured DLBCLs are on the left; profiles of the 26 fatal/refractory tumors are on the right.

To determine whether the gene-expression patterns associated with DLBCL and FL (Fig. 1) were sufficiently robust to predict the lymphoma type of an unknown sample, we used the weighted-voting algorithm, which calculates the weighted combination of informative marker genes to make a class distinction (that is, DLBCL versus FL)[14]. To avoid the statistical problem of over-estimating prediction accuracy that occurs when a model is trained and evaluated with the same samples, we used a 'leave-one-out' cross-validation testing method. In this procedure, 1 of the 77 samples is withheld, and the remaining 76 samples are used to train a gene-expression–based model and predict the class of the withheld sample. The process is repeated until all 77 samples are predicted in turn. A 30-gene predictor correctly classified 71 of 77 tumors (91%) with respect to the DLBCL versus FL distinction ($P < 1 \times 10^{-9}$ compared with random prediction).

Predicting outcome in DLBCL

The success in distinguishing DLBCL from FL with supervised learning suggested that a similar approach might be used to delineate clinically relevant subsets of DLBCL. Long-term clinical follow-up was available for all 58 DLBCL patients in the study. These patients were divided into two groups: those with cured disease (n = 32) and those with fatal or refractory disease (n = 26).

The genes most highly correlated with the cured versus fatal/refractory distinction included genes that have been previously associated with DLBCL outcome, such as VEGF, linked with adverse outcome[10] and overexpressed in fatal/refractory DLBCLs, and E2F, associated with favorable outcome[30] and overexpressed in cured DLBCLs (Fig. 2). The presence of known prognostic markers among our outcome-correlated genes indicates that the gene-expression signatures are likely to be *bona fide*.

We next used a supervised learning classification approach (weighted-voting algorithm and cross-validation testing) to develop a DLBCL outcome predictor and assess its accuracy. Predictors containing between 8 and 16 genes all yielded statistically significant outcome predictions, with the highest accuracy obtained using 13 genes (Fig. 3). Although each of the cross-validation loops generated a new 13-gene model, each of these models contained mostly the same genes (see Methods and www.genome.wi.mit.edu/MPR/lymphoma).

The predictor separated the 58 patients, who had a 5-year overall survival (OS) of 54% (Fig. 4*a*), into 2 groups: those predicted to be cured and those predicted to have fatal/refractory disease. Kaplan–Meier survival analyses indicated that the patients predicted to be cured had significantly improved long-term survival compared with those predicted to have fatal/refractory disease (5-year OS, 70% versus 12%; nominal log rank P-value = 0.00004; Fig. 4*b*). Other classification algorithms, including support vector machines (SVM) and k-nearest neighbors (k-NN) performed similarly (5-year OS for SVM, 72% versus 12%, P = 0.00002; for k-NN, 68% versus 23%, P = 0.001; www.genome.wi.mit.edu/MPR/lymphoma). These results indi-

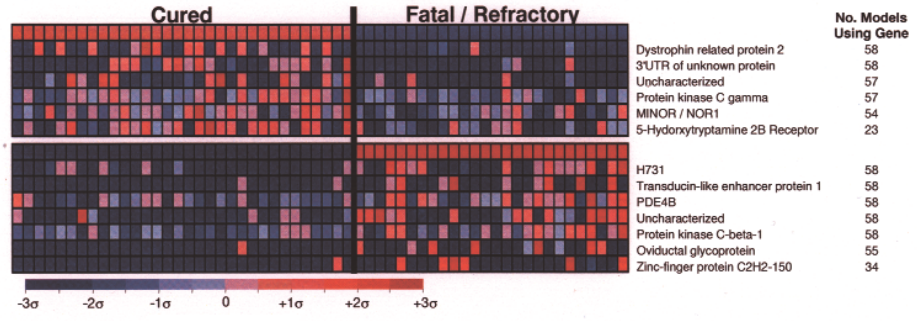

Fig. 3 Genes included in the DLBCL outcome model. Genes expressed at higher levels in cured disease are listed on top and those that were more abundant in fatal disease are shown on bottom. Red indicates high expression; blue, low expression. Color scale at bottom indicates relative expression in standard deviations from the mean. Each column is a sample, each row a gene. Expression profiles of the 32 cured DLBCLs are on the left; profiles of the 26 fatal/refractory tumors are on the right. Models with the highest accuracy were obtained using 13 genes. Although each of the 58 cross-validation loops generates a new 13-gene model, 7 of the genes were common to all 58 models; 4 additional genes were included in 54 or more models and 2 genes were included in 23–34 models.

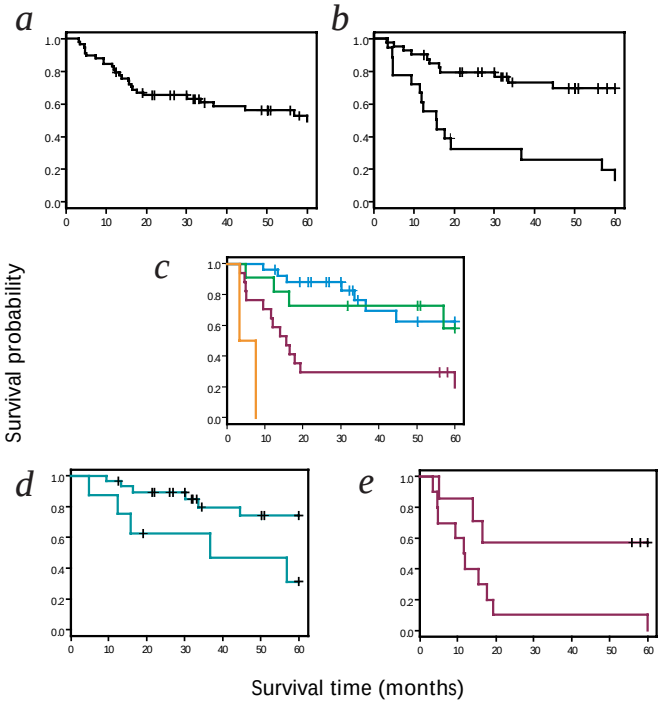

Fig. 4 Overall survival predictions for DLBCL study patients. ***a***, 5-year OS for the entire study group. 33 of 58 DLBCL study patients remained alive at a median of a 58-month follow-up. The predicted 5-year OS for the group as a whole was 54%. ***b***, 5-year OS for favorable and unfavorable risk groups defined by the 13-gene model (70% versus 12%, $P = 0.00004$). Top line, cured; bottom, fatal/refractory. ***c***, 5-year OS for patients in L-risk (green line), LI-risk (blue line), HI-risk (red line) and H-risk (orange line) categories as defined by the IPI: L, 26 pts; LI, 11 pts; HI, 17 pts; H, 2 pts. ***d***, 5-year OS for combined L/LI-risk patients with favorable or unfavorable disease as defined by the molecular model (75% versus 32%, nominal $P = 0.02$) Top line, cured; bottom, fatal/refractory. ***e***, 5-year OS for HI-risk patients with favorable or unfavorable disease as defined by the molecular model (57% versus 0%; nominal $P = 0.02$). Top line, cured; bottom, fatal/refractory.

cate the existence, at diagnosis, of a gene-expression signature of outcome in DLBCL.

The clinically based IPI outcome predictor is effective in predicting the outcome of subsets of DLBCL patients[3]. In the current series, all of the IPI-defined H-risk patients died of their disease (Fib. 4*c*). However, the IPI incorrectly predicted the outcome of many of the patients in the other IPI risk groups (high intermediate (HI), low intermediate (LI) and low (L)) (Fig. 4*c*). For this reason, we investigated whether the gene-expression–based outcome predictor contained additional information not captured by the IPI. L/LI-risk patients with the 'cured' gene-expression signature had significantly higher OS rates than L/LI-risk patients with the 'fatal/refractory' signature (5-year OS, 75 versus 32%; $P = 0.02$) (Fig. 4*d*). Similarly, the outcome of HI-risk patients could be further predicted by the application of the gene-expression model (5-year OS, 57 versus 0%; $P = 0.02$) (Fig. 4*e*). These results indicate that the microarray-based outcome predictor provides additional information not reflected in the clinical prognostic model and suggests a possible strategy for further individualization of patient treatment. However, the gene-expression–based predictor did not eliminate outcome differences between L/LI-risk and HI-risk patients (Fig. 4*d* and *e*), suggesting that the clinical and molecular models contain at least partially independent information. Additional studies will be required to determine how to optimally combine such models.

Validating the model

Having defined an outcome predictor for DLBCL, we investigated the connection, if any, between this model and the cell-of-origin classification described by Alizadeh *et al.* Such comparisons are admittedly difficult, given that, 1) different genes were measured on the arrays, 2) the microarray technology was different (oligonucleotide versus cDNA arrays), 3) different computational approaches were employed, and 4) different patient samples were studied. Nevertheless, we determined that 90 of the previously described cell-of-origin signature genes[12] were also represented on our oligonucleotide arrays (see Methods and www.genome.wi.mit.edu/MPR/lymphoma).

We first used a hierarchical clustering algorithm to sort the DLBCL samples of Alizadeh *et al.* based on expression of the 90 cell-of-origin signature genes represented on both the cDNA and oligonucleotide microarrays. Two major branches of the hierarchical tree were observed; these branches were closely associ-

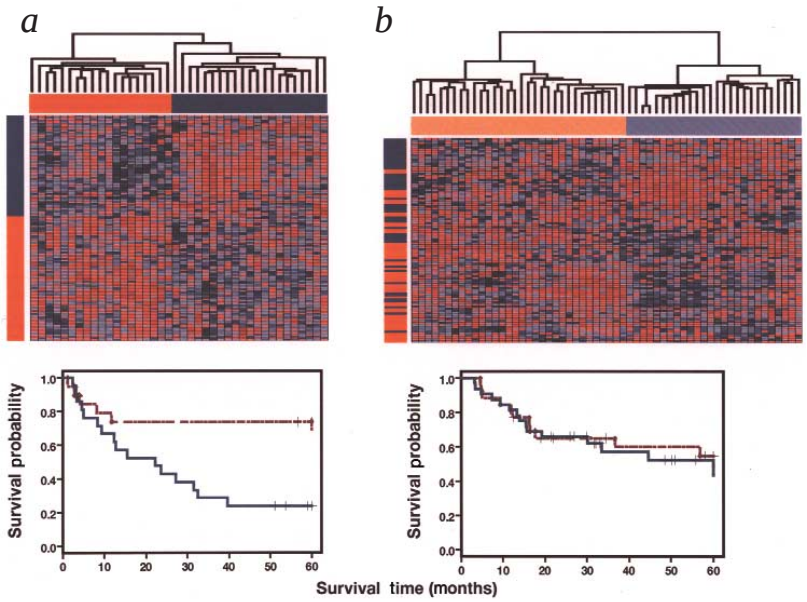

Fig. 5 Predictive value of GC-B-cell and activated B-cell signatures. ***a*** and ***b***, The GC and activated B-cell markers common to our dataset (*b*) and that of Alizadeh *et al.*[12] (*a*). (90 common UniGene clusters) were hierarchically clustered[45] with respect to patient samples. In each dataset, 2 major branches of the hierarchical tree were observed. In the Alizadeh *et al.* dataset, the 2 major branches corresponded exactly to the previously described cell-of-origin distinction (GC-like DLBCLs, orange and activated B-like DLBCLs, blue). Genes (rows) correlated with these 2 categories are similarly indicated with the same color scheme. In our dataset, the 2 major branches of the hierarchical tree were also associated with putative cell of origin. Genes correlated with the left branch were predominantly GC-like (orange) genes, whereas genes selectively expressed in the right branch were predominantly activated B-like (blue) genes ($P = .00001$, χ^2 test). In the bottom panels, the 5-year OS for patients whose tumors exhibited the GC (top lines) and activated B-cell (bottom lines) signature are shown.

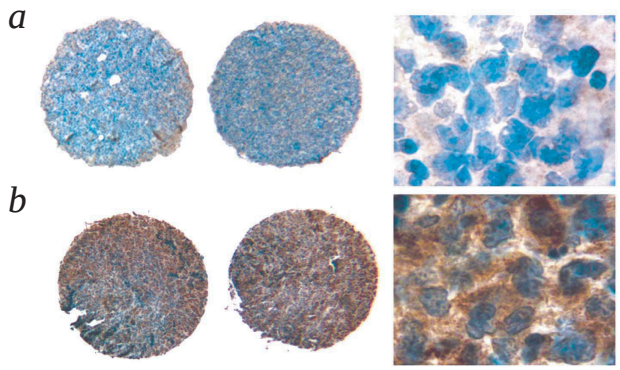

Fig. 6 Immunohistochemical staining for PKC-β. ***a*** and ***b***, Representative PKC-β immunostaining of duplicate core samples from a cured DLBCL (*a*) and a fatal DLBCL (*b*) are shown at low (×4, left) and high (×1000, right) power.

ated with the cell-of-origin distinction[12], confirming that the 90 overlapping signature genes were sufficient to make this determination (Fig. 5*a*). As expected, the 90-gene cell-of-origin distinction was associated with outcome in the DLBCL samples of Alizadeh *et al.* (Fig. 5*a*).

We then used these same 90 genes to cluster our own 58 DLBCL samples (Fig. 5*b*). Again, two major branches of the hierarchical tree were observed, and these branches were highly correlated with the cell-of-origin distinction ($P = 0.00001$, χ^2 test) (Fig 5*b*). However, this distinction was not significantly correlated with patient outcome in our DLBCL series (Fig. 5*b*). This observation suggests that although the signature genes may reflect cell of origin, they do not explain a significant portion of the clinical variability seen in this DLBCL dataset.

We next investigated whether we could find support for our outcome predictor in the expression data of Alizadeh *et al.* Of the 13 genes in our supervised DLBCL outcome predictor, 3 were represented on the lymphochip: *NOR1* (also known as *NR4A3*), *PDE4B* and *PKC*-β (also known as *PRKCB1*). When evaluated as single markers in the dataset of Alizadeh *et al.*, *NOR1* ($P = 0.05$) and *PDE4B* ($P = 0.07$) were clearly correlated with outcome. Multiple *PKC*-β cDNAs are present on the lymphochip; these clones gave discordant expression results in the DLBCL patients, perhaps reflecting varying degrees of specificity for the β isoforms of PKC. However, two clones (clone 1308435 and 685194), specific for the PKC-β2 isoform, were indeed correlated with outcome in the DLBCL patient series of Alizadeh *et al.* ($P = 0.04$). These results from an independent dataset confirm our initial observations and highlight the value of publicly accessible gene-expression databases for rapid, computational validation of hypotheses.

The potential extension of microarray-based outcome prediction to the clinical setting was further explored using immunohistochemical detection methods. For this purpose, we generated a tissue array containing the study DLBCLs for which formalin-fixed, paraffin-embedded tumor tissue was available ($n = 21$). PKC-β protein expression was analyzed because of the critical role of PKC pathways in B-cell signaling and the commercial availability of a monoclonal antibody against PKC-β known to function in immunohistochemistry assays. PKC-β protein expression was highly associated with microarray-determined transcript abundance in the DLBCL specimens ($P = 0.08$, Fisher exact test; Fig. 6). In addition, PKC-β protein expression was closely associated with clinical outcome in the DLBCL patients ($P = 0.03$). This result both validates the microarray measurements, and demonstrates how microarray-based studies can be extended using methods that are more widely available in routine clinical practice.

Discussion

Herein, we report the successful prediction of outcome in a series of DLBCL patients using oligonucleotide microarray gene-expression data and supervised learning methods. Genes implicated in DLBCL outcome included ones that regulate responses to B-cell–receptor signaling, critical serine/threonine phosphorylation pathways and apoptosis. For example, all three of the computationally validated microarray-based outcome genes, *NOR1*, *PDE4B* and *PKC*-β, regulate apoptotic responses to antigen-receptor engagement and, potentially, cytotoxic chemotherapy.

The mitogen-inducible nuclear orphan receptor (MINOR) or NOR1 is overexpressed in cured, as opposed to fatal/refractory, DLBCL (Figs. 2 and 4). NOR1 is a member of the nerve growth factor-1B (NGF1B, also known as NR4A1) subfamily (NGF1B/TR3/Nur77, Nurr-1) of nuclear orphan receptors[31]. NGF1B family members are induced by antigen-receptor engagement and external stressors such as seizures or ischemia[31,32]; in addition, these factors directly promote the apoptosis of affected cells[31,33,34]. Recent studies indicate that at least one NGF1B family member (NGF1B/TR3/Nur77) translocates from the nucleus to the mitochondria where it directly exerts its proapoptotic effects[31,35]. Given the functions of related NGF1B family members, it is possible that NOR1 increases the apoptotic response to chemotherapy in curable DLBCL.

The cyclic AMP (cAMP)-specific phosphodiesterase PDE4B is overexpressed in fatal/refractory, as opposed to cured, DLBCL (Figs. 2 and 4). PDE4s are the predominant class of phosphodiesterases in lymphocytes[36,37], catalyzing the hydrolysis of cAMP and terminating its activity[37]. cAMP-dependent protein kinase A (PKA) signaling inhibits lymphocyte chemotaxis, cytokine release and cellular proliferation[36]. Because PDE4B reduces cAMP-availability, the phosphodiesterase also limits the negative effects of PKA signaling in lymphocytes. For this reason, PDE4A and -4B inhibitors are being evaluated in the treatment of certain B-cell malignancies where they are reported to induce B-cell apoptosis[37,38]. Together, these data suggest that PDE4B may also be an attractive therapeutic target in fatal/refractory DLBCLs.

Like PDE4B, PKC-β is overexpressed in fatal/refractory, rather than cured, DLBCL (Figs. 2,3 and 6). The alternatively-spliced PKC-β1 and -β2 isoforms are the major PKC isoforms expressed by B-lymphocytes[39]. The pivotal role of PKC-β in B-cell signaling and survival was recently demonstrated in PKC-β-deficient mice which have profoundly impaired humoral and B-cell proliferative responses[40]. In additional *in vitro* analyses, the consequences of B-cell–receptor signaling were dependent upon associated activation of PKC-β (ref. 41). In the presence of an intact PKC-β pathway, B-cell–receptor engagement resulted in B-cell proliferation; however, B-cell–receptor signaling induced apoptosis when mature B cells are either PKC depleted or stimulated in the presence of PKC inhibitors[41]. Taken together, these studies suggest that PKC-β activity enhances B-cell proliferation and survival, consistent with our observation that the enzyme is overexpressed in fatal/refractory DLBCLs. Recently, synergy between PKC-β inhibitors and chemotherapeutic agents in murine tumor

models[42] has been reported, further suggesting that pharmacologic inhibition of PKC-β may have a therapeutic role in the future treatment of fatal/refractory DLBCL.

These studies demonstrate the potential of DNA microarray-based recognition of gene-expression patterns for the prediction of outcome in DLBCL patients. This work also illustrates the important difference between unsupervised (clustering) and supervised machine learning analytical approaches. The previously reported cell-of-origin distinction[12] was originally identified using an unsupervised clustering algorithm and this distinction was subsequently associated with disease outcome. In our series, the cell-of-origin distinction was not associated with significant outcome differences (Fig. 5), suggesting that additional factors may be important in determining DLBCL response to therapy.

One limitation of the supervised classification method employed here is that it reduces the classification problem to a dichotomous distinction (cured versus fatal/refractory disease). However, it is likely that these distinct clinical behaviors are explained by different molecular mechanisms in different patients. More refined outcome prediction may thus require the use of alternative feature selection algorithms capable of capturing more complex DLBCL substructure, or the application of non-linear classification strategies. Moreover, optimal outcome prediction may require not only gene-expression data but also the inclusion of tumor genotype information.

Nevertheless, the DLBCL outcome-correlated genes described here were highly informative, including key intermediaries in signaling pathways that regulate apoptotic responses to receptor engagement, and potentially, to cytotoxic therapy. These studies suggest strategies for both optimizing the use of existing therapy for DLBCL and developing more rationally designed therapies in this disease. The computational validation of our DLBCL outcome predictor using publicly available gene-expression databases further illustrates the important role of computational genomics in biomedical research.

Methods

Samples. Frozen diagnostic nodal tumor specimens from 58 DLBCL patients and 19 FL patients were analyzed according to an Institutional Review Board approved protocol. The histopathology and immunophenotype of each tumor were centrally reviewed to confirm diagnosis and uniform involvement with tumor. The DLBCL study patients were those for whom frozen tumor tissue and complete clinical information (presenting clinical characteristics, treatment records and long-term follow-up) were available. Treatment records of all 58 DLBCL patients were reviewed to confirm that patients had received adequate doses of cyclophosphamide, adriamycin, vincristine and prednisone (CHOP)-like combination chemotherapy[2] for 6 or more cycles or until documented disease progression and to document outcome and clinical IPI risk group[3]. The IPI was not determined in 2 patients because of missing LDH levels in these patients. DLBCL study patients (predicted 5-year OS 54%, median follow-up 58 months) were divided into 2 discrete categories: 1) 29 patients who achieved CR and remained free of disease plus 3 additional patients who died of other causes (total of 32 'cured' patients); and 2) 23 patients who died of lymphoma plus 3 additional patients who remained alive with recurrent refractory or progressive disease (total of 'fatal/refractory' 26 patients).

Target cRNAs and oligonucleotide microarrays. Total RNA was extracted from each frozen tumor specimen and biotinylated cRNAs were generated as described[43] and as detailed (see Supplementary Information). Samples were hybridized overnight to Affymetrix HU6800 oligonucleotide arrays (Affymetrix, Santa Clara, California)[14]. Arrays were subsequently developed with phycoerythrin-conjugated streptavidin (SAPE) and biotinylated antibody against streptavidin, and scanned to obtain quantitative gene-expression levels[14,43]. The raw gene-expression values were then scaled in order to account for any minor differences in global chip intensity. Expression levels below 20 units were assigned a value of 20, and those exceeding 16,000 units were assigned a value of 16,000. Genes whose expression did not vary across the dataset were removed (see Supplementary Information).

Supervised prediction. Classes (classes 0 and 1) were defined based on morphology (DLBCL versus FL) or treatment outcome (cured versus fatal/refractory disease). Marker genes were then identified using a signal-to-noise calculation: $S_x = (\mu_{class0} - \mu_{class1}) \div (\sigma_{class0} + \sigma_{class1})$ where, for each gene, μ_{class0} represents the mean value of arrays with true class equal to class 0, and σ_{class0} represents the standard deviation of class 0 samples[14]. Thereafter, a weighted-voting classification algorithm was applied as previously described, and was tested by 'leave-one-out' cross-validation[14]. The total number of prediction errors in cross-validation was calculated using a variable number of genes, and a final model chosen which minimized cross-validation errors. Analyses of error rates, confusion matrices (false negatives versus false positives) and Kaplan–Meier survival curves were performed using S-Plus (http://www.splus.mathsoft.com/products/splus/splusintro.html). The log-rank test was used to assess the differences between the survival curves and nominal *P*-values were calculated.

The *P*-value for the prediction of lymphoma type (DLBCL versus FL) was predicted using the proportional chance calculation[44] as described (http://marketing.byu.edu/htmlpages/tutorials/discriminant.htm).

Analysis of lymphochip microarray data. The raw lymphochip data from the 40 DLBCL specimens and the associated outcome information was obtained from the Lymphoma/Leukemia Molecular Profiling Project (http://llmpp.nih.gov/lymphoma). RAT2 values were pre-processed by setting minimum values to 0 and normalizing arrays to a mean value of 0 and variance of 1. Computational model validation was performed by identifying genes from our outcome predictor (Fig. 3) that were represented on the lymphochip. For the lymphochip data, we mapped the clone IMAGE (integrated molecular analysis of genomes and their expression) numbers to GenBank accession numbers (using the list at http://llmpp.nih.gov/lymphoma/data/clones.txt) and then mapped the accession numbers to UniGene clusters (National Center for Biotechnology Information, Bethesda, Maryland). Similarly, we mapped accession numbers for our oligonucleotide array data to UniGene clusters. Predictors using single genes (*PKC*-β, *PDE4B*, MINOR/*NOR1*) were constructed by finding the boundary halfway between the classes ($b_x = (\mu_{class0} + \mu_{class1}) \div 2$) in the dataset and predicting the unknown sample according to its gene-expression value with respect to that boundary. This method is equivalent to performing weighted voting with only 1 gene.

The 90 UniGene clusters common to both arrays are represented by 139 clones in the data of Fig. 3*c* of Alizadeh *et al.* and by 100 probe sets on the oligonucleotide arrays. The DLBCL series of Alizadeh *et al.* and our DLBCL series were separately clustered using these common cell-of-origin signature genes by average linkage hierarchical clustering, and the results visualized using TreeView (from M. Eisen)[45].

Immunohistochemical staining. Five representative 0.6-mm cores were obtained from diagnostic areas of each paraffin-embedded formalin-fixed DLBCL tumor and inserted in a grid pattern in a single-recipient paraffin block using a tissue arrayer (Beecher Instruments, Silver Spring, Maryland). Five-micron sections cut from this 'tissue array' were stained for PKC-β using an immunoperoxidase method. Briefly, slides were deparaffinized and pre-treated in 1 mM EDTA (pH 8.0) for 20 min at 95 °C. All further steps were performed at room temperature in a hydrated chamber. Slides were pre-treated with Peroxidase Block (DAKO, Carpinteria, California) for 5 min to quench endogenous peroxidase activity, and a 1:5 dilution of goat serum in 50 mM Tris-Cl (pH 7.4) for 20 min to block non-specific binding sites. Primary antibody (murine monoclonal antibody specific for PKC-β (Serotec, Oxford, UK) was applied at a 1:1000 dilution in 50 mM Tris-Cl (pH 7.4) with 3% goat serum for 1 h. After washing, secondary goat anti-mouse horseradish-peroxidase–conjugated antibody (Envision Detection Kit, DAKO) was applied for 30 min. After further washing, immunoperoxidase staining was developed using a DAB chromogen kit (DAKO) according to manufacturer's instructions. Following counterstaining with hema-

toxylin, immunoperoxidase staining within the malignant cell population of each core was scored in a blinded fashion with respect to clinical outcome and expression profile results by 3 experienced hematopathologists (J.C.A., A.P.W. and J.L.K.). The intensity of staining on each core was graded from 0 (no staining) to 3 (maximal staining), and an average staining intensity (mean of all 5 cores) was generated for each tumor. Median values were used to divide both the PKC immunostaining intensities and the array-based transcript levels into two categories. The Fisher exact test was then used to evaluate the association between these measurements.

Acknowledgments

We thank members of the Center for Genome Research and members of the Shipp and Aster laboratories for technical assistance and helpful comments. This work was supported in part by grants from Bristol–Myers Squibb, Millennium Pharmaceuticals and Affymetrix (E.S.L.).

RECEIVED 18 MAY; ACCEPTED 26 NOVEMBER 2001

1. A clinical evaluation of the International Lymphoma Study Group classification of non-Hodgkin's lymphoma. The Non-Hodgkin's Lymphoma Classification Project. *Blood* **89**, 3909–3918 (1997).
2. Shipp, M., Harris, N. & Mauch, P. The non-Hodgkin's lymphomas. in *Cancer Principles & Practices of Oncology* (eds. DeVita, V.T., Hellman, S. & Rosenberg, S.A) 2165–2220 (Lippincott, Philadelphia, 1997).
3. A predictive model for aggressive non-Hodgkin's lymphoma. The International NHL Prognostic Factors Project. *N. Engl. J. Med.* **329**, 987–994 (1993).
4. Shipp, M.A. *et al.* International consensus conference on high-dose therapy with hematopoietic stem cell transplantation in aggressive non-Hodgkin's lymphomas: Report of the jury. *J. Clin. Oncol.* **17**, 423–429 (1999).
5. Yakushijin, Y. *et al.* A directly spliced exon 10-containing CD44 variant promotes the metastasis and homotypic aggregation of aggressive non-Hodgkin's lymphoma. *Blood* **91**, 4282–4291 (1998).
6. Terol, M.-J. *et al.* Expression of the adhesion molecule ICAM-1 in non-Hodgkin's lymphoma: Relationship with tumor dissemination and prognostic importance. *J. Clin. Oncol.* **16**, 35–40 (1998).
7. Gascoyne, R. *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).
8. Kramer, M. *et al.* Clinical relevance of BCL2, BCL6, and MYC rearrangements in diffuse large B-cell lymphoma. *Blood* **92**, 3152–3162 (1998).
9. Ambrosini, G., Adida, C. & Altieri, D. A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma. *Nature Med.* **3**, 917–921 (1997).
10. Salven, P., Teerenhovi, L. & Joensuu, H. A high pretreatment serum vascular endothelial growth factor concentration is associated with poor outcome in non-Hodgkin's lymphoma. *Blood* **90**, 3167–3172 (1997).
11. Aguiar, R. *et al.* BAL is a novel risk-related gene in diffuse large B-cell lymphomas which enhances cellular migration. *Blood* **96**, 4328–4334 (2000).
12. Alizadeh, A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **4051**, 503–511 (2000).
13. Kuppers, R., Klein, U., Hansmann, M.-L. & Rajewsky, K. Cellular origin or human B-cell lymphomas. *N. Engl. J. Med.* **341**, 1520–1529 (1999).
14. Golub, T. *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
15. Cleary, M. *et al.* Clustering of extensive somatic mutations in the variable region of an immunoglobulin heavy chain gene from a human B cell lymphoma. *Cell* **44**, 97–106 (1986).
16. Wood, L. *et al.* HMG-I/Y, a new c-Myc target gene and potential oncogene. *Mol. Cell Biol.* **20**, 5490–5502 (2000).
17. Ilangumaran, S., Borisch, B. & Hoessli, D. Signal transduction via CD44: role of plasma membrane microdomains. *Leuk. Lymphoma* **35**, 455–469 (1999).
18. Matarrese, P. *et al.* Galectin-3 overexpression protects from apoptosis by improving cell adhesion properties. *Internal J. Cancer* **85**, 545–554 (2000).
19. Zhang, H. *et al.* Structural basis of BFL-1 for its interaction with BAX and its anti-apoptotic action in mammalian and yeast cells. *J. Biol. Chem.* **275**, 11092–11099 (2000).
20. Lee, H., Dadgostar, H., Cheng, Q., Shu, J. & Cheng, G. NF-κB-mediated up-regulation of Bcl-x and Bfl-1/A1 is required for CD40 survival signaling in B lymphocytes. *Proc. Natl. Acad. Sci. USA* **96**, 9136–9141 (1999).
21. Wang, J., Saunthararajah, Y., Redner, R.L. & Liu, J.M. Inhibitors of histone deacetylase relieve ETO-mediated repression and induce differentiation of AML1-ETO leukemia cells. *Cancer Res.* **59**, 2766–2769 (1999).
22. Zong, W., Edelstein, L., Chen, C., Bash, J. & Gelinas, C. The prosurvival Bcl-2 homolog Bfl-1/A1 is a direct transcriptional target of NF-κB that blocks TNFα-induced apoptosis. *Genes Devel.* **13**, 382–387 (1999).
23. Finger, L. *et al.* The human PD-1 gene: complete cDNA, genomic organization, and developmentally regulated expression in B cell progenitors. *Gene* **197**, 177–187 (1997).
24. Kitson, J. *et al.* A death-domain-containing receptor that mediates apoptosis. *Nature* **384**, 372–375 (1996).
25. Wellmann, A. *et al.* Detection of differentially expressed genes in lymphomas using cDNA arrays: identification of clusterin as a new diagnostic marker for anaplastic large-cell lymphomas. *Blood* **96**, 398–404 (2000).
26. Sommers, C. *et al.* A role for the Tec family tyrosine kinase Txk in T cell activation and thymocyte selection. *J. Exp. Med.* **190**, 1427–1438 (1999).
27. Testi, R., D'Ambrosio, D., De Maria, R. & Santoni, A. The CD69 receptor: a multipurpose cell-surface trigger for hematopoietic cells. *Immunol. Today* **15**, 479–483 (1994).
28. Ruegg, C. *et al.* V7, a novel leukocyte surface protein that participates in T cell activation. II. Molecular cloning and characterization of the V7 gene. *J. Immunol.* **154**, 4434–4443 (1995).
29. Saeki, H., Moore, A., Brown, M. & Hwang, S. Cutting edge: secondary lymphoid-tissue chemokine (SLC) and CC chemokine receptor 7 (CCR7) participate in the emigration pathway of mature dendritic cells from the skin to regional lymph nodes. *J. Immunol.* **162**, 2472–2475 (1999).
30. Moller, M. *et al.* Frequent disruption of the RB1 pathway in diffuse large B cell lymphoma: prognostic significance of E2F-1 and p16NK4A. *Leukemia* **14**, 898–904 (2000).
31. Brenner, C. & Kroemer, G. Mitochondria—the death signal integrators. *Science* **289**, 1150–1151 (2000).
32. Youn, H., Sun, L., Prywes, R. & Liu, J. Apoptosis of T cells mediated by Ca^{2+}-induced release of the transcription factor MEF2. *Science* **286**, 790–793 (1999).
33. Kuang, A., Cado, D. & Winoto, A. Nur77 transcription activity correlates with its apoptotic function *in vivo*. *Eur. J. Immunol.* **29**, 3722–3728 (1999).
34. Xue, Y. *et al.* Positive and negative thymic selection in T cell receptor-transgenic mice correlate with Nur77 mRNA expression. *Eur. J. Immunol.* **27**, 2048–2056 (1997).
35. Li, H. *et al.* Cytochrome c release and apoptosis induced by mitochondrial targeting of nuclear orphan receptor TR3. *Science* **289**, 1159–1164 (2000).
36. Manning, C. *et al.* Suppression of human inflammatory cell function by subtype-selective PDE4 inhibitors correlates with inhibition of PDE4A and PDE4B. *Br. J. Pharmacol.* **128**, 1393–1398 (1999).
37. Lerner, A., Kim, D. & Lee, R. The cAMP signaling pathway as a therapeutic target in lymphoid malignancies. *Leuk. Lymphoma* **37**, 39–51 (2000).
38. Kim, D. & Lerner, A. Type 4 cyclic adenosine monophosphate phosphodiesterase as a therapeutic target in chronic lymphocytic leukemia. *Blood* **92**, 2484–2494 (1998).
39. Mischak, H. *et al.* Expression of protein kinase C genes in hemopoietic cells is cell-type- and B cell-differentiation stage specific. *J. Immunol.* **147**, 3981–3987 (1991).
40. Leitges, M. *et al.* Immunodeficiency in protein kinase Cb-deficient mice. *Science* **273**, 788–791 (1996).
41. King, L., Norvell, A. & Monroe, J. Antigen receptor-induced signal transduction imbalances associated with the negative selection of immature B cells. *J. Immunol.* **162**, 2655–2662 (1999).
42. Teicher, B. *et al.* Enzymatic rationale and preclinical support for a potent protein kinase C beta inhibitor in cancer therapy. *Adv. Enzyme Reg.* **39**, 313–327 (1999).
43. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
44. Huberty, C.J. *Applied Discriminant Analysis*. (John Wiley and Sons, Inc., 1994).
45. 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).