Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2]*† D. K. Slonim,[1]† P. Tamayo,[1] C. Huard,[1]
M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2]
J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4]
E. S. Lander[1,5]*

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.

The challenge of cancer treatment has been to target specific therapies to pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity. Improvements in cancer classification have thus been central to advances in cancer treatment. Cancer classification has been based primarily on morphological appearance of the tumor, but this has serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. In a few cases, such clinical heterogeneity has been explained by dividing morphologically similar tumors into subtypes with distinct pathogeneses. Key examples include the subdivision of acute leukemias, non-Hodgkin's lymphomas, and childhood "small round blue cell tumors" [tumors with variable response to chemotherapy (*1*) that are now molecularly subclassified into neuroblastomas, rhabdomyosarcoma, Ewing's sarcoma, and other types (*2*)]. For many more tumors, important subclasses are likely to exist but have yet to be defined by molecular markers. For example, prostate cancers of identical grade can have widely variable clinical courses, from indolence over decades to explosive growth causing rapid patient death. Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes. Here we describe such an approach based on global gene expression analysis.

We divided cancer classification into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes, which could reflect current states or future outcomes.

We chose acute leukemias as a test case. Classification of acute leukemias began with the observation of variability in clinical outcome (*3*) and subtle differences in nuclear morphology (*4*). Enzyme-based histochemical analyses were introduced in the 1960s to demonstrate that some leukemias were periodic acid-Schiff positive, whereas others were myeloperoxidase positive (*5*). This provided the first basis for classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, ALL) or from myeloid precursors (acute myeloid leukemia, AML). This classification was further solidified by the development in the 1970s of antibodies recognizing either lymphoid or myeloid cell surface molecules (*6*). Most recently, particular subtypes of acute leukemia have been found to be associated with specific chromosomal translocations—for example, the t(12;21)(p13;q22) translocation occurs in 25% of patients with ALL, whereas the t(8;21)(q22;q22) occurs in 15% of patients with AML (*7*).

Although the distinction between AML and ALL has been well established, no single test is currently sufficient to establish the diagnosis. Rather, current clinical practice involves an experienced hematopathologist's interpretation of the tumor's morphology, histochemistry, immunophenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remains imperfect and errors do occur.

Distinguishing ALL from AML is critical for successful treatment; chemotherapy regimens for ALL generally contain corticosteroids, vincristine, methotrexate, and L-asparaginase, whereas most AML regimens rely on a backbone of daunorubicin and cytarabine (*8*). Although remissions can be achieved using ALL therapy for AML (and vice versa), cure rates are markedly diminished, and unwarranted toxicities are encountered.

We set out to develop a more systematic approach to cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays (*9*). It has been suggested (*10*) that such microarrays could provide a tool for cancer classification. Microarray studies to date (*11*), however, have primarily been descriptive rather than analytical and have focused on cell culture rather than primary patient material, in which genetic noise might obscure an underlying reproducible expression pattern.

We began with class prediction: How could one use an initial collection of samples belonging to known classes (such as AML and ALL) to create a "class predictor" to classify new, unknown samples? We developed an analytical method and first tested it on distinctions that are easily made at the morphological level, such as distinguishing normal kidney from renal cell carcinoma (*12*). We then turned to the more challenging problem of distinguishing acute leukemias, whose appearance is highly similar.

Our initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis (*13*). RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes (*14*). For each gene, we obtained a quantitative expression level. Samples were subjected to a priori quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image (*15*).

The first issue was to explore whether

[1] Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139, USA. [2] Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02115, USA. [3] St. Jude Children's Research Hospital, Memphis, TN 38105, USA. [4] Comprehensive Cancer Center and Cancer and Leukemia Group B, Ohio State University, Columbus, OH 43210, USA. [5] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: golub@genome.wi.mit.edu; lander@genome.wi.mit.edu.

†These authors contributed equally to this work.

there were genes whose expression pattern was strongly correlated with the class distinction to be predicted. The 6817 genes were sorted by their degree of correlation (*16*). To establish whether the observed correlations were stronger than would be expected by chance, we developed a method called "neighborhood analysis" (Fig. 1A). Briefly, one defines an "idealized expression pattern" corresponding to a gene that is uniformly high in one class and uniformly low in the other. One tests whether there is an unusually high density of genes "nearby" (that is, similar to) this idealized pattern, as compared to equivalent random patterns.

For the 38 acute leukemia samples, neighborhood analysis showed that roughly 1100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance (Fig. 2) (*17*). This suggested that classification could indeed be based on expression data.

The second issue was how to use a collection of known samples to create a "class predictor" capable of assigning a new sample to one of two classes. We developed a procedure that uses a fixed subset of "informative genes" (chosen based on their correlation with the class distinction) and makes a prediction on the basis of the expression level of these genes in a new sample. Each informative gene casts a "weighted vote" for one of the classes, with the magnitude of each vote dependent on the expression level in the new sample and the degree of that gene's correlation with the class distinction (Fig. 1B) (*18*, *19*). The votes were summed to determine the winning class, as well as a "prediction strength" (PS), which is a measure of the margin of victory that ranges from 0 to 1 (*20*). The sample was assigned to the winning class if PS exceeded a predetermined threshold, and was otherwise considered uncertain. On the basis of previous analysis, we used a threshold of 0.3 (*21*).

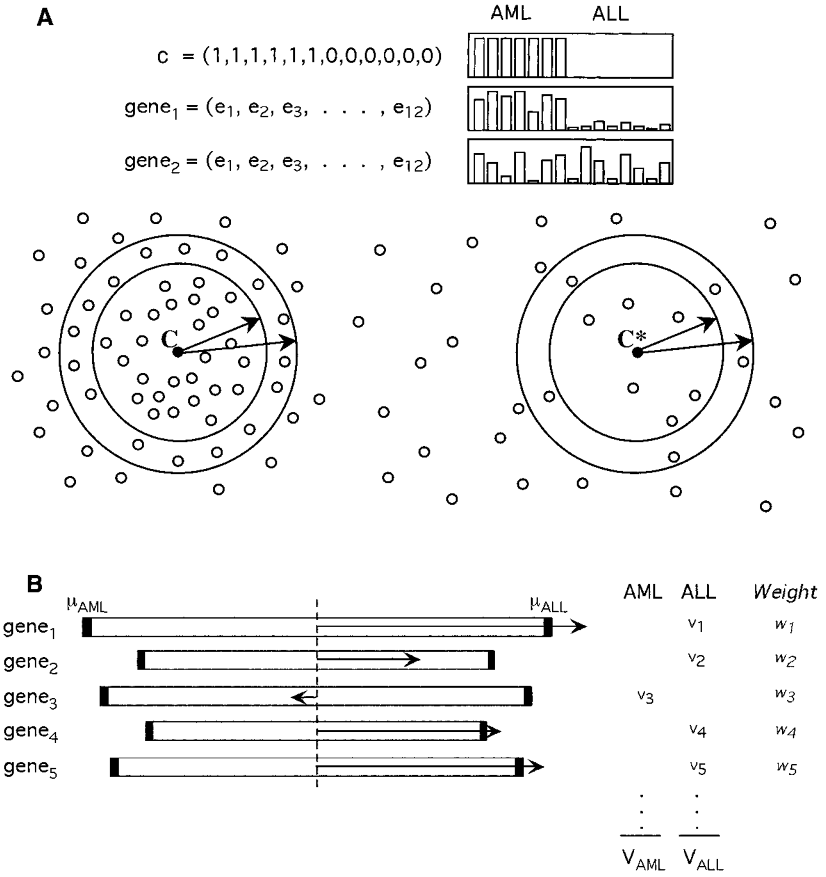
A
AML
ALL
c = (1,1,1,1,1,1,0,0,0,0,0,0)
gene₁ = (e₁, e₂, e₃, . . . , e₁₂)
gene₂ = (e₁, e₂, e₃, . . . , e₁₂)
C
C*
B
μAML
μALL
gene₁
gene₂
gene₃
gene₄
gene₅
AML
ALL
Weight
v₁
w₁
v₂
w₂
v₃
w₃
v₄
w₄
v₅
w₅
V_AML
V_ALL

Fig. 1. Schematic illustration of methodology. (**A**) Neighborhood analysis. The class distinction is represented by an "idealized expression pattern" *c*, in which the expression level is uniformly high in class 1 and uniformly low in class 2. Each gene is represented by an expression vector, consisting of its expression level in each of the tumor samples. In the figure, the data set is composed of six AMLs and six ALLs. Gene g_1 is well correlated with the class distinction, whereas g_2 is poorly correlated. Neighborhood analysis involves counting the number of genes having various levels of correlation with *c*. The results are compared to the corresponding distribution obtained for random idealized expression patterns *c**, obtained by randomly permuting the coordinates of *c*. An unusually high density of genes indicates that there are many more genes correlated with the pattern than expected by chance. The precise measure of distance and other methodological details are described in (*16*, *17*) and on our Web site (www.genome.wi.mit.edu/MPR). (**B**) Class predictor. The prediction of a new sample is based on "weighted votes" of a set of informative genes. Each such gene g_i votes for either AML or ALL, depending on whether its expression level x_i in the sample is closer to μ_{AML} or μ_{ALL} (which denote, respectively, the mean expression levels of AML and ALL in a set of reference samples). The magnitude of the vote is w_iv_i, where w_i is a weighting factor that reflects how well the gene is correlated with the class distinction and $v_i = |x_i - (\mu_{AML} + \mu_{ALL})/2|$ reflects the deviation of the expression level in the sample from the average of μ_{AML} and μ_{ALL}. The votes for each class are summed to obtain total votes V_{AML} and V_{ALL}. The sample is assigned to the class with the higher vote total, provided that the prediction strength exceeds a predetermined threshold. The prediction strength reflects the margin of victory and is defined as $(V_{win} - V_{lose})/(V_{win} + V_{lose})$, where V_{win} and V_{lose} are the respective vote totals for the winning and losing classes. Methodological details are described in (*19*, *20*) and on the Web site.

The third issue was how to test the validity of class predictors. We used a two-step procedure. The accuracy of the predictors was first tested by cross-validation on the initial data set. (Briefly, one withholds a sample, builds a predictor based only on the remaining samples, and predicts the class of the withheld sample. The process is repeated for each sample, and the cumulative error rate is calculated.) One then builds a final predictor based on the initial data set and assesses its accuracy on an independent set of samples.

We applied this approach to the 38 acute leukemia samples. The set of informative genes to be used in the predictor was chosen to be the 50 genes most closely correlated with AML-ALL distinction in the known samples. The parameters of the predictor were determined by the expression levels of these 50 genes in the known samples. The predictor was then used to classify new samples, by applying it to the expression levels of these genes in the sample.

The 50-gene predictors derived in cross-validation tests assigned 36 of the 38 samples as either AML or ALL and the remaining two as uncertain ($PS < 0.3$) (*22*). All 36 predictions agreed with the patients' clinical diagnosis.

We then created a 50-gene predictor on the basis of all 38 samples and applied it to an independent collection of 34 leukemia samples. The specimens consisted of 24 bone marrow and 10 peripheral blood samples (*23*). In total, the predictor made strong predictions for 29 of the 34 samples, and the accuracy was 100%. The success was notable because the collection included a much broader range of samples, including samples from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sample preparation protocols. Overall, the prediction strengths were quite high (median $PS = 0.77$ in cross-validation and 0.73

in independent test) (Fig. 3A). The average prediction strength was lower for samples from one laboratory that used a very different protocol for sample preparation. This suggests that clinical implementation of such an approach should include standardization of sample preparation.

The choice to use 50 informative genes in the predictor was somewhat arbitrary. The number was well within the total number of genes strongly correlated with the class distinction (Fig. 2), seemed likely to be large enough to be robust against noise, and was small enough to be readily applied in a clinical setting. In fact, the results were insensitive to the particular choice: Predictors based on between 10 and 200 genes were all found to be 100% accurate, reflecting the strong correlation of genes with the AML-ALL distinction (*24*).

The list of informative genes used in the AML versus ALL predictor was highly instructive (Fig. 3B). Some, including *CD11c*, *CD33*, and *MB-1*, encode cell surface proteins for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells (*25*). Others provide new markers of acute leukemia subtype. For example, the leptin receptor, originally identified through its role in weight regulation, showed high relative expression in AML. The leptin receptor was recently demonstrated to have anti-apoptotic function in hematopoietic cells (*26*). Similarly, the zyxin gene has been shown to encode a LIM domain protein important in cell adhesion in fibroblasts, but a role in hematopoiesis has not been reported (*27*).

We had expected that the genes most useful in AML-ALL class prediction would simply be markers of hematopoietic lineage, and would not necessarily be related to cancer pathogenesis. However, many of the genes encode proteins critical for S-phase cell cycle progression (Cyclin *D3*, *Op18*, and *MCM3*), chromatin remodeling (*RbAp48* and *SNF2*), transcription (*TFIIE*β), and cell adhesion (zyxin and *CD11c*) or are known oncogenes (*c-MYB*, *E2A* and *HOXA9*). In addition, one of the informative genes encodes topoisomerase II, which is the principal target of the antileukemic drug etoposide (*28*). Together, these data suggest that genes useful for cancer class prediction may also provide insight into cancer pathogenesis and pharmacology.

The methodology of class prediction can be applied to any measurable distinction among tumors. Importantly, such distinctions could concern a future clinical outcome—such as whether a prostate cancer turns out to be indolent or a breast cancer responds to a given chemotherapy. We explored the ability to predict response to chemotherapy among the 15 adult AML patients who had been treated with an anthracycline-cytarabine regimen and for whom long-term clinical follow-up was available (*29*). Eight patients failed to achieve remission after induction chemotherapy, while the remaining seven remained in remission for 46 to 84 months. Neighborhood analysis found no striking excess of genes correlated with response to chemotherapy, in contrast to the situation for the AML-ALL distinction, and class predictors that used 10 to 50 genes were not highly accurate in cross-validation. We thus found no evidence of a strong multigene expression signature correlated with clinical outcome, although this could reflect the relatively small sample size. Nonetheless, we examined the most highly correlated genes for potential biological significance. The single most highly correlated gene out of the 6817 genes was the homeobox gene *HOXA9*, which was overexpressed in patients with treatment failure. Notably, *HOXA9* is rearranged by a t(7; 11)(p15;p15) chromosomal translocation in a rare subset of AML patients, who tend to have poor outcomes (*30*). Furthermore, *HOXA9* overexpression has been shown to transform myeloid cells in vitro and to cause leukemia in animal models (*31*). A general role for *HOXA9* expression in predicting AML outcome has not been previously suggested. Larger studies will be needed to test this hypothesis.

We next turned to the question of class discovery. The initial identification of cancer classes has been slow, typically evolving through years of hypothesis-driven research. We explored whether cancer classes could be discovered automatically. For example, if the AML-ALL distinction were not already known, could it have been discovered simply on the basis of gene expression?

Class discovery entails two issues: (i) developing algorithms to cluster tumors by gene expression and (ii) determining whether putative classes produced by such clustering algorithms are meaningful—that is, whether they reflect true structure in the data rather than simply random aggregation.

To cluster tumors, we used a technique called self-organizing maps (SOMs), which is particularly well suited to the task of identifying a small number of prominent classes in a data set (*32*). In this approach, the user specifies the number of clusters to be identified. The SOM finds an optimal set of "centroids" around which the data points appear to aggregate. It then partitions the data set, with each centroid defining a cluster consisting of the data points nearest to it.

We applied a two-cluster SOM to automatically group the 38 initial leukemia samples into two classes on the basis of the expression pattern of all 6817 genes (*33*). We first evaluated the clusters by comparing them to the known AML-ALL classes (Fig. 4A). The SOM paralleled the known classes closely: Class A1 contained mostly ALL (24 of 25 samples) and class A2 contained mostly AML (10 of 13 samples). The SOM was thus quite effective, albeit not perfect, at automatically discovering the two types of leukemia.

We then considered how one could evaluate such putative clusters if the "right" answer were not already known. We reasoned

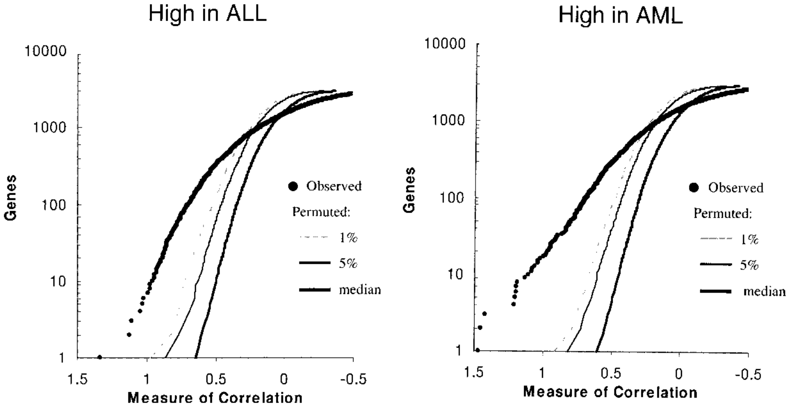

Fig. 2. Neighborhood analysis: ALL versus AML. For the 38 leukemia samples in the initial data set, the plot shows the number of genes within various "neighborhoods" of the ALL-AML class distinction together with curves showing the 5 and 1% significance levels for the number of genes within corresponding neighborhoods of the randomly permuted class distinctions (*16*, *17*). Genes more highly expressed in ALL compared to AML are shown in the left panel; those more highly expressed in AML compared to ALL are shown in the right panel. The large number of genes highly correlated with the class distinction is apparent. In the left panel (higher in ALL), the number of genes with correlation $P(g,c) > 0.30$ was 709 for the AML-ALL distinction, but had a median of 173 genes for random class distinctions. $P(g,c) = 0.30$ is the point where the observed data intersect the 1% significance level, meaning that 1% of random neighborhoods contain as many points as the observed neighborhood around the AML-ALL distinction. Similarly, in the right panel (higher in AML), 711 genes with $P(g,c) > 0.28$ were observed, whereas a median of 136 genes is expected for random class distinctions.

that class discovery could be tested by class prediction: If putative classes reflect true structure, then a class predictor based on these classes should perform well.

To test this hypothesis, we evaluated the clusters A1 and A2. We constructed predictors to assign new samples as "type A1" or "type A2." Predictors that used a wide range of different numbers of informative genes performed well in cross-validation. For example, a 20-gene predictor gave 34 accurate predictions with high prediction strength, one error, and three uncertains (*34*). The one "error" was the assignment of the sole AML sample in class A1 to class A2, and two of the three uncertains were ALL samples in class A2. The cross-validation thus not only showed high accuracy, but actually refined the SOM-defined classes: With one exception, the subset of samples accurately classified in cross-validation were those perfectly subdivided by the SOM into ALL and AML classes. The results suggest an iterative procedure for refining clusters, in which an SOM is used to initially cluster the data, a predictor is constructed, and samples not correctly predicted in cross-validation are removed. The edited data set could then be used to generate an improved predictor to be tested on an independent data set (*35*).

We then tested the class predictor of the A1-A2 distinction on the independent data set. In the general case of class discovery, predictors for novel classes cannot be assessed for "accuracy" on new samples, because the "right" way to classify the independent samples is not known. Instead, however, one can assess whether the new samples are assigned a high prediction strength. High prediction strengths indicate that the structure seen in the initial data set is also seen in the independent data set. The prediction strengths, in fact, were quite high: The median PS was 0.61, and 74% of samples were above threshold (Fig. 4B). To assess these results, we performed the same analyses with random clusters. Such clusters consistently yielded predictors with poor accuracy in cross-validation and low prediction strength on the independent data set (Fig. 4B). On the basis of such analysis (*36*), the A1-A2 distinction can be seen to be meaningful, rather than simply a statistical artifact of the initial data set. The results thus show that the AML-ALL distinction could have been automatically discovered and confirmed without previous biological knowledge.

We then sought to extend the class discovery by searching for finer subclasses of the leukemias. We used a SOM to divide the samples into four clusters (denoted B1 to B4). We subsequently obtained immunophenotype data on the samples and found that the four classes largely corresponded to AML, T-lineage ALL, B-lineage ALL, and B-lineage ALL, respectively (Fig. 4C). The four-cluster SOM thus divided the samples along

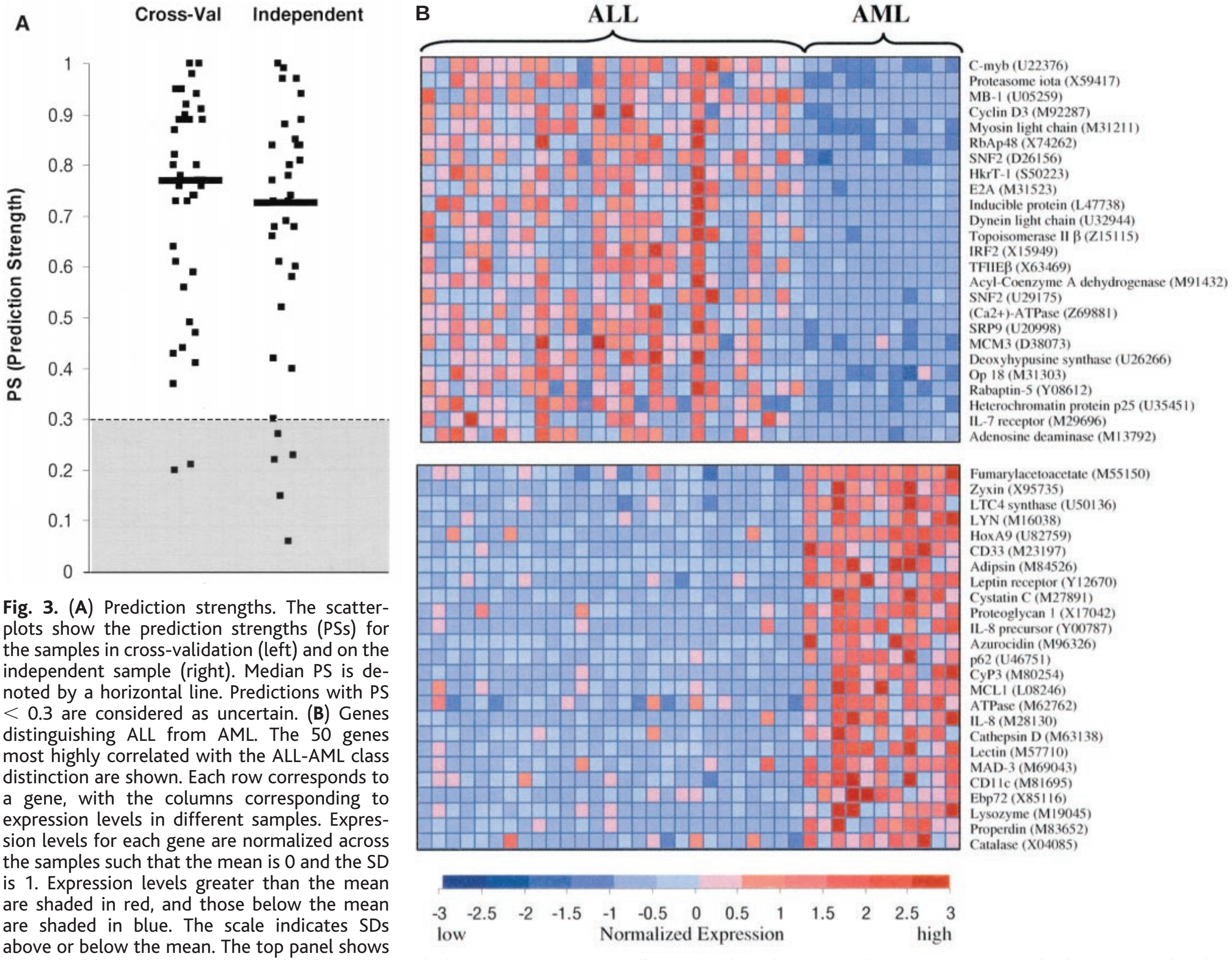

Fig. 3. (**A**) Prediction strengths. The scatterplots show the prediction strengths (PSs) for the samples in cross-validation (left) and on the independent sample (right). Median PS is denoted by a horizontal line. Predictions with PS < 0.3 are considered as uncertain. (**B**) Genes distinguishing ALL from AML. The 50 genes most highly correlated with the ALL-AML class distinction are shown. Each row corresponds to a gene, with the columns corresponding to expression levels in different samples. Expression levels for each gene are normalized across the samples such that the mean is 0 and the SD is 1. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue. The scale indicates SDs above or below the mean. The top panel shows genes highly expressed in ALL, the bottom panel shows genes more highly expressed in AML. Although these genes as a group appear correlated with class, no single gene is uniformly expressed across the class, illustrating the value of a multigene prediction method. For a complete list of gene names, accession numbers, and raw expression values, see www.genome.wi.mit.edu/MPR.

another key biological distinction.

We again evaluated these classes by constructing class predictors (*37*). The four classes could be distinguished from one another, with the exception of B3 versus B4 (Fig 4D). The prediction tests thus confirmed the distinctions corresponding to AML, B-ALL, and T-ALL, and suggested that it may be appropriate to merge classes B3 and B4, composed primarily of B-lineage ALL.

The class discovery approach thus automatically discovered the distinction between AML and ALL, as well as the distinction between B-cell and T-cell ALL. These are the most important distinctions known among acute leukemias, both in terms of underlying biology and clinical treatment. With larger sample collections, it would be possible to search for finer subclassifications. It will be interesting to see whether they correspond to existing subclassifications for AML and ALL or define new groupings perhaps based on fundamental similarities in mechanism of transformation.

In principle, the class discovery techniques above can be used to identify fundamental subtypes of any cancer. In general, such studies will require careful experimental design to avoid potential experimental artifacts—especially in the case of solid tumors. Biopsy specimens, for example, might have gross differences in the proportion of surrounding stromal cells. Blind application of class discovery could result in identifying classes reflecting the proportion of stromal contamination in the samples, rather than underlying tumor biology. Such "classes" would be real and reproducible, but would not be of biological or clinical interest. Various approaches could be used to avoid such artifacts—such as microscopic examination of tumor samples to ensure comparability, purification of tumor cells by flow sorting or laser-capture microdissection, computational analysis that excludes genes expressed in stromal cells, and confirmation of candidate marker genes by RNA in situ hybridization or immunohistochemistry to tumor sections.

Class discovery methods could also be used to search for fundamental mechanisms that cut across distinct types of cancers. For example, one might combine different cancers (for example, breast tumors and prostate tumors) into a single data set, eliminate those genes that correlate strongly with tissue type, and then cluster the samples based on the remaining genes.

We also describe techniques for class prediction, whereby samples can be automatically assigned to already-recognized classes. Creation of a new predictor involves expression analysis of thousands of genes to select a set of informative genes (we used 50 genes, although other choices also performed well) and then validating the accuracy of the assignments made on the basis of these genes. Subsequent application of the predictor then requires only monitoring the expression level of these informative genes. We described a class predictor able to accurately assign samples as AML or ALL. We have also similarly constructed a class predictor that accurately assigns ALL samples as either T-ALL or B-ALL (*38*). These class predictors could be adapted to a clinical setting, with appropriate steps to standardize the protocol for sample preparation. We envisage such a test supplementing rather than replacing existing leukemia diagnostics. Indeed, this would provide an opportunity to gain clinical experience with the use of expression-based class predictors in a well-studied cancer, before applying them to cancers with less well-developed diagnostics.

More generally, class predictors may be useful in a variety of settings. First, class pre-

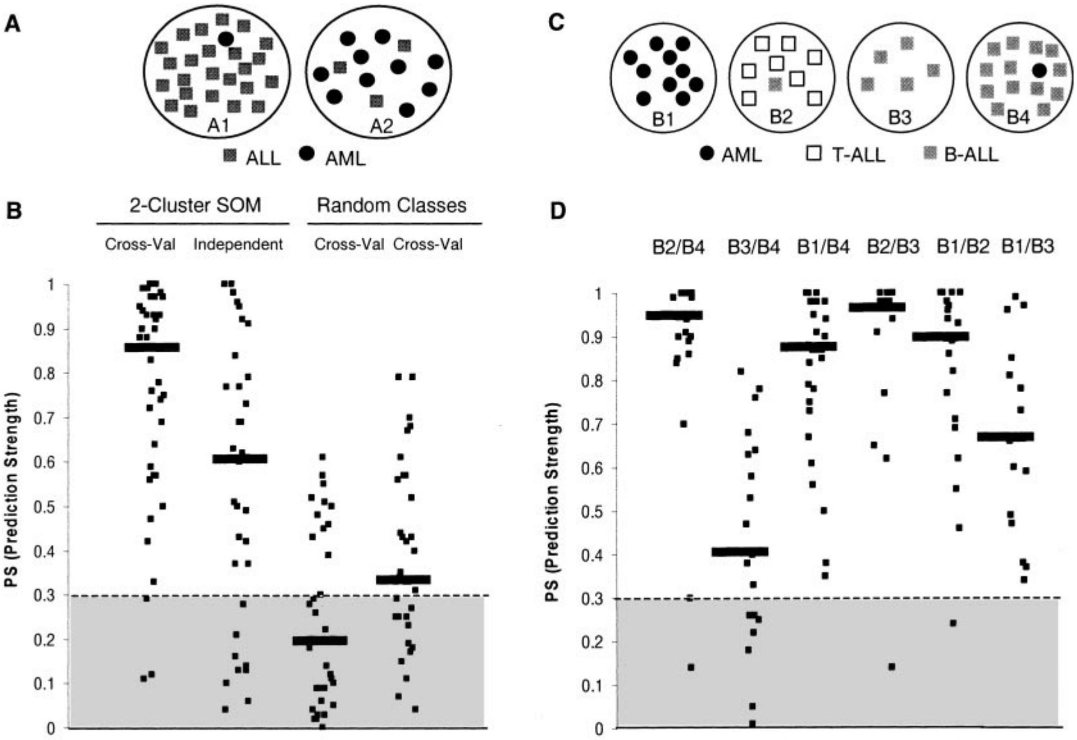

Fig. 4. ALL-AML class discovery. (**A**) Schematic representation of two-cluster SOM. A two-cluster (2 by 1) SOM was generated from the 38 initial leukemia samples, with a modification of the GENECLUSTER computer package (*32*). Each of the 38 samples is thereby placed into one of two clusters on the basis of patterns of gene expression for the 6817 genes assayed in each sample. Cluster A1 contains the majority of ALL samples (gray squares) and cluster A2 contains the majority of AML samples (black circles). (**B**) Prediction strength (PS) distributions. The scatterplots show the distribution of PS scores for class predictors. The first two plots show the distribution for the predictor created to classify samples as "A1-type" or "A2-type" tested in cross-validation on the initial data set (median PS = 0.86) and on the independent data set (median PS = 0.61). The remaining plots show the distribution for two predictors corresponding to random classes. In these cases, the PS scores are much lower (median PS = 0.20 and 0.34, respectively), and about half of the samples fall below the threshold for prediction (PS = 0.3). A total of 100 such random predictors were examined, to calculate the distribution of median PS scores to evaluate the statistical significance of the predictor for A1-A2 (*36*). (**C**) Schematic representation of the four-cluster SOM. AML samples are shown as black circles, T-lineage ALL as open squares, and B-lineage ALL as gray squares. T- and B-lineages were differentiated on the basis of cell-surface immunophenotyping. Class B1 is exclusively AML, class B2 contains all eight T-ALLs, and classes B3 and B4 contain the majority of B-ALL samples. (**D**) Prediction strength (PS) distributions for pair-wise comparison among classes. Cross-validation prediction studies show that the four classes could be distinguished with high prediction scores, with the exception of classes B3 and B4. These two classes could not be easily distinguished from one another, consistent with their both containing primarily B-ALL samples, and suggesting that B3 and B4 might best be merged into a single class.

dictors can be constructed for known pathological categories—reflecting a tumor's cell of origin, stage, or grade. Such predictors could provide diagnostic confirmation or clarify unusual cases. This point was illustrated by a recent anecdotal experience. A patient with a classic leukemia presentation (pancytopenia, circulating "blasts") was diagnosed with AML, but with atypical morphology. We took the opportunity to apply our class predictor to a bone marrow sample from this patient. The classifier produced extremely low vote totals for both AML and ALL: Neither lymphoid- nor myeloid-specific genes were highly expressed, thus bringing into question the diagnosis of acute leukemia. Examination of the expression profile revealed that genes more highly expressed relative to the leukemias included those encoding tropomyosin, muscle-specific actin, decorin, and *IGF-2*, suggestive of a mesenchymal origin, such as muscle (*39*). In fact, independent cytogenetic analysis identified a t(2; 13)(q35;q14) translocation characteristic of the muscle tumor alveolar rhabdomyosarcoma (*40*). The patient's diagnosis was revised accordingly, and treatment was changed from AML therapy to rhabdomyosarcoma therapy. This experience underscores the fact that leukemia diagnosis remains imperfect and could benefit from a battery of expression-based predictors for various cancers.

Most importantly, the technique of class prediction can be applied to distinctions relating to future clinical outcome, such as drug response or survival. Class prediction provides an unbiased, general approach to constructing such prognostic tests, provided that one has a collection of tumor samples for which eventual outcome is known.

References and Notes

1. T. J. Triche *et al.*, *Prog. Clin. Biol. Res.* **271**, 475 (1988).
2. C. F. Stephenson, J. A. Bridge, A. A. Sandberg, *Hum. Pathol.* **23**, 1270 (1992); O. Delattre *et al.*, *N. Engl. J. Med.* **331**, 294 (1994); C. Turc-Carel *et al.*, *Cancer Genet. Cytogenet.* **19**, 361 (1986); E. C. Douglass *et al.*, *Cytogenet. Cell Genet.* **45**, 148 (1987); R. Dalla-Favera *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 7824 (1982); R. Taub *et al.*, *ibid.*, p. 7837; G. Balaban-Malenbaum and F. Gilbert, *Science* **198**, 739 (1977).
3. S. Farber, L. K. Diamond, R. D. Mercer, R. F. Sylvester, J. A. Wolff, *N. Engl. J. Med.* **238**, 787 (1948).
4. C. E. Forkner, *Leukemia and Allied Disorders* (Macmillan, New York, 1938); E. Frei *et al.*, *Blood* **18**, 431 (1961); Medical Research Council, *Br. Med. J.* **1**, 7 (1963).
5. D. Quaglino and F. G. J. Hayhoe, *J. Pathol.* **78**, 521 (1959); J. M. Bennett and T. F. Dutcher, *Blood* **33**, 341 (1969); R. C. Graham, U. Lundholm, M. J. Karnovsky, *J. Histochem. Cytochem.* **13**, 150 (1965).
6. I. Tsukimoto, R. Y. Wong, B. C. Lampkin, *N. Engl. J. Med.* **294**, 245 (1976); S. F. Schlossman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 1288 (1976); M. Roper *et al.*, *Blood* **61**, 830 (1983); B. S. E. Sallan *et al.*, *ibid.* **55**, 395 (1980); J. M. Pesando *et al.*, *ibid.* **54**, 1240 (1979).
7. T. R. Golub *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 4917 (1995); T. W. McLean *et al.*, *Blood* **88**, 4252 (1996); S. A. Shurtleff *et al.*, *Leukemia* **9**, 1985 (1995); S. P. Romana *et al.*, *Blood* **86**, 4263 (1995); J. D. Rowley, *Ann. Genet.* **16**, 109 (1973).
8. Recent reviews of ALL and AML therapy can be found in C. H. Pui and W. E. Evans, *N. Engl. J. Med.* **339**, 605 (1998); J. F. Bishop, *Med. J. Aust.* **170**, 39 (1999); R. M. Stone and R. J. Mayer, *Hematol. Oncol. Clin. N. Am.* **7**, 47 (1993).
9. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996); D. J. Lockhart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996); V. R. Iyer *et al.*, *Science* **283**, 83 (1999); L. Wodicka, H. Dong, M. Mittmann, M. H. Ho, D. J. Lockhart, *Nature Biotechnol.* **15**, 1359 (1997); P. T. Spellman *et al.*, *Mol. Biol. Cell* **9**, 3273 (1998); M. Schena *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10614 (1996); G. P. Yang, D. T. Ross, W. W. Kuang, P. O. Brown, R. J. Weigel, *Nucleic Acids Res.* **27**, 1517 (1999).
10. E. S. Lander, *Science* **274**, 536 (1996); J. DeRisi, *et al.*, *Nature Genet.* **14**, 457 (1996); J. Kononen *et al.*, *Nature Med.* **4**, 844 (1998); J. Khan *et al.*, *Cancer Res.* **58**, 5009 (1998); K. A. Cole *et al.*, *Nature Genet.* **21** (suppl. 1), 38 (1999).
11. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996); G. P. Yang, D. T. Ross, W. W. Kuang, P. O. Brown, R. J. Weigel, *Nucleic Acids Res.* **27**, 1517 (1999); J. Khan *et al.*, *Cancer Res.* **58**, 5009 (1998); J. Khan *et al.*, *Electrophoresis* **20**, 223 (1999).
12. We compared six normal human kidney biopsies and six kidney tumors (renal cell carcinomas, RCCs) using the methods described for the leukemias. Neighborhood analysis showed a high density of genes correlated with the distinction. Class predictors were constructed using 50 genes, and the predictions proved to be 100% accurate in cross-validation. The informative genes more highly expressed in normal kidney as compared to RCCs included 13 metabolic enzymes, two ion channels, and three isoforms of the heavy-metal chelator metallothionein, all of which function in normal kidney physiology. Those more highly expressed in RCC than normal kidney included interleukin-1, an inflammatory cytokine responsible for the febrile response experienced by patients with RCC, and CCND1, a D-type cyclin amplified in some cases of RCC.
13. The initial 38 samples were all derived from bone marrow aspirates performed at the time of diagnosis, before chemotherapy. After informed consent was obtained, mononuclear cells were collected by Ficoll sedimentation and total RNA extracted with either Trizol (Gibco/BRL) or RNAqueous reagents (Ambion). The 27 ALL samples were derived from childhood ALL patients treated on Dana-Farber Cancer Institute (DFCI) protocols between 1980 and 1999. Samples were randomly selected from the leukemia cell bank based on availability. The 11 adult AML samples were similarly obtained from the Cancer and Leukemia Group B (CALGB) leukemia cell bank. Samples were selected without regard to immunophenotype, cytogenetics, or other molecular features. The independent samples used to confirm the results contained a broader range of samples, including peripheral blood samples and childhood AML cases (*23*).
14. A total of 3 to 10 μg of total RNA from each sample was used to prepare biotinylated target essentially as previously described, with minor modifications [P. Tamayo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2907 (1999); L. Wodicka, H. Dong, M. Mittmann, M. Ho, D. Lockhart, *Nature Biotechnol.* **15**, 1359 (1997)]. A complete description of the biochemical and mathematical procedures used in this paper is available through our Web site at www.genome.wi.mit.edu/MPR.
15. Samples were excluded if they yielded less than 15 μg of biotinylated RNA, if the hybridization was weak (see our Web site for quantitative criteria), or if there were visible defects in the array (such as scratches). A total of 80 leukemia samples were analyzed during the course of the experiments reported here. Of these, eight were excluded on the basis of these a priori quality control criteria.
16. Each gene is represented by an expression vector $v(g) = (e_1, e_2, \ldots, e_n)$, where e_i denotes the expression level of gene g in ith sample in the initial set S of samples. A class distinction is represented by an idealized expression pattern $c = (c_1, c_2, \ldots, c_n)$, where $c_i = +1$ or 0 according to whether the i-th sample belongs to class 1 or class 2. One can measure "correlation" between a gene and a class distinction in a variety of ways. One can use the Pearson correlation coefficient or the Euclidean distance. We used a measure of correlation, $P(g,c)$, that emphasizes the "signal-to-noise" ratio in using the gene as a predictor. Let $[\mu_1(g),\sigma_1(g)]$ and $[\mu_2(g),\sigma_2(g)]$ denote the means and SDs of the log of the expression levels of gene g for the samples in class 1 and class 2, respectively. Let $P(g,c) = [\mu_1(g) - \mu_2(g)]/[\sigma_1(g) + \sigma_2(g)]$, which reflects the difference between the classes relative to the SD within the classes. Large values of $|P(g,c)|$ indicate a strong correlation between the gene expression and the class distinction, while the sign of $P(g,c)$ being positive or negative corresponds to g being more highly expressed in class 1 or class 2. Unlike a standard Pearson correlation coefficient, $P(g,c)$ is not confined to the range $[-1, +1]$. Neighborhoods $N_1(c,r)$ and $N_2(c,r)$ of radius r around class 1 and class 2 were defined to be the sets of genes such that $P(g,c) = r$ and $P(g,c) = -r$, respectively. An unusually large number of genes within the neighborhoods indicates that many genes have expression patterns closely correlated with the class vector.
17. A permutation test was used to calculate whether the density of genes in a neighborhood was statistically significantly higher than expected. We compared the number of genes in the neighborhood to the number of genes in similar neighborhoods around idealized expression patterns corresponding to random class distinctions, obtained by permuting the coordinates of c. We performed 400 permutations and determined the 5 and 1% significance levels for the number of genes contained within neighborhoods of various levels of correlation with c. See also the legend to Fig. 2.
18. The set of informative genes consists of the $n/2$ genes closest to a class vector high in class 1 [that is, $P(g,c)$ as large as possible] and the $n/2$ genes closest to class 2 [that is, $-P(g,c)$ as large as possible]. The number n of informative genes is the only free parameter in defining the class predictor.
19. The class predictor is uniquely defined by the initial set S of samples and the set of informative genes. Parameters (a_g, b_g) are defined for each informative gene. The value $a_g = P(g,c)$ reflects the correlation between the expression levels of g and the class distinction. The value $b_g = [\mu_1(g) + \mu_2(g)]/2$ is the average of the mean log expression values in the two classes. Consider a new sample X to be predicted. Let x_g denote the normalized log (expression level) of gene g in the sample (where the expression level is normalized by subtracting the mean and dividing by the SD of the expression levels in the initial set S). The vote of gene g is $v_g = a_g(x_g - b_g)$, with a positive value indicating a vote for class 1 and a negative value indicating a vote for class 2. The total vote V_1 for class 1 is obtained by summing the absolute values of the positive votes over the informative genes, while the total vote V_2 for class 2 is obtained by summing the absolute values of the negative votes.
20. The prediction strength PS is defined as $\text{PS} = (V_{\text{win}} - V_{\text{lose}})/(V_{\text{win}} + V_{\text{lose}})$, where V_{win} and V_{lose} are the vote totals for the winning and losing classes. The measure PS reflects the relative margin of victory of the vote.
21. The appropriate PS threshold depends on the number n of genes in the predictor, because the PS is a sum of n variables corresponding to the individual genes, and thus its fluctuation for random input data scales inversely with $\sqrt{n}$. See our Web site concerning the specific choice of PS threshold.
22. In cross-validation, the entire prediction process is repeated from scratch with 37 of the 38 samples. This includes identifying the 50 informative genes to be used in the predictor and defining parameters for weighted voting.
23. The independent set of leukemia samples comprised 24 bone marrow and 10 peripheral blood specimens, all obtained at the time of leukemia diagnosis. The ALL samples were obtained from the DFCI childhood ALL bank ($n = 17$) or St. Jude Children's Research Hospital (SJCRH) ($n = 3$). Whereas the AML samples in the initial data set were all derived from adult patients, the AML samples in the independent data set were derived from both adults and children. The samples were obtained from either the CALGB (adult

AML, $n = 4$), SJCRH (childhood AML, $n = 5$), or the Children's Cancer Group (childhood AML, $n = 5$) leukemia banks. The samples were processed as described (*13*), with the exception of the samples from SJCRH, which used a very different protocol. The SJCRH samples were subjected to hypotonic lysis (rather than Ficoll sedimentation), and RNA was prepared by an aqueous extraction (Qiagen).

24. Although the number of genes used had no significant effect on the outcome in this case (median PS for cross-validation ranged from 0.81 to 0.68 over a range of predictors using 10 to 200 genes, all with 0% error), it may matter in other instances. One approach is to vary the number of genes used, select the number that maximizes the accuracy rate in cross-validation, and then use the resulting model on the independent data set. In any case, we recommend using at least 10 genes for two reasons. Class predictors using a small number of genes may depend too heavily on any one gene and can produce spuriously high prediction strengths (because a large "margin of victory" can occur by chance due to statistical fluctuation resulting from a small number of genes). In general, we also considered the 99% confidence line in neighborhood analysis to be the upper bound for gene selection.
25. P. A. Dinndorf *et al.*, *Med. Pediatr. Oncol.* **20**, 192 (1992); P. S. Master, S. J. Richards, J. Kendall, B. E. Roberts, C. S. Scott, *Blut* **59**, 221 (1989); V. Buccheri *et al.*, *Blood* **82**, 853 (1993).
26. M. Konopleva *et al.*, *Blood* **93**, 1668 (1999).
27. A. W. Crawford and M. C. Beckerle, *J. Biol. Chem.* **266**, 5847 (1991).
28. W. Ross, T. Rowe, B. Glisson, J. Yalowich, L. Liu, *Cancer Res.* **44**, 5857 (1984).
29. Treatment failure was defined as failure to achieve a complete remission after a standard induction regimen including 3 days of anthracycline and 7 days of cytarabine. Treatment successes were defined as patients in continuous complete remission for a minimum of 3 years. FAB subclass M3 patients were excluded but samples were otherwise not selected with regard to FAB criteria.
30. J. Borrow *et al.*, *Nature Genet.* **12**, 159 (1996); T. Nakamura, *et al.*, *ibid.*, p. 154; S. Y. Huang *et al.*, *Br. J. Haematol.* **96**, 682 (1997).
31. E. Kroon *et al.*, *EMBO J.* **17**, 3714 (1998).
32. P. Tamayo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2907 (1999).
33. The SOM was constructed using our GENECLUSTER software (*32*), with a variation filter excluding genes with less than fivefold variation across the collection of samples.
34. For testing putative clusters derived from the SOM or chosen at random, we constructed class predictors with various number of genes (ranging from 10 to 100) and selected the one with the highest cross-validation accuracy rate (in this case, 20 genes).
35. A related approach would be to represent each cluster only as the subset of points lying near the centroid of the cluster.
36. Various statistical methods can be used to compare the predictors derived from the SOM-derived clusters with predictors derived from random classes. We compared the median prediction strength. Specifically, 100 predictors corresponding to random classes of comparable size were constructed, and the median PS for each predictor was determined. The performance for the actual predictor was then compared to the distribution of these 100 median PSs, to obtain empirical significance levels. The observed median PS in the initial data set was 0.86, which exceeded the median PS for all 100 random predictors; the empirical significance level was thus $<1\%$. The observed median PS for the independent data set was 0.61, which exceed the median PS for all but 4 of the 100 random permutations; the empirical significance level was thus 4%.
37. Various approaches can be used to test classes $C_1, C_2, \ldots, C_n$ arising from a multinode SOM. One can construct predictors to distinguish each pair of classes (C_i versus C_j) or to distinguish each class for the complement of the class (C_i versus not C_i). Here we used the pair-wise approach (C_i versus C_j). For cross-validation, one can restrict attention to samples known to lie in the union of C_i and C_j. For an independent data set, one must examine all samples (because it is unknown which samples lie in the union of C_i and C_j). It may be possible to improve the statistical power of this test by using techniques for multiclass prediction.
38. Thirty-three ALL samples were tested by cross-validation using a 50-gene predictor. Thirty-two of 33 samples were correctly assigned as T-ALL or B-ALL; the remaining sample received a PS < 0.3, and no prediction was therefore made. Details are provided on our Web site.
39. T. R. Golub, unpublished results.
40. S. Turc-Carel *et al.*, *Cancer Genet. Cytogenet.* **19**, 361 (1986); E. C. Douglass *et al.*, *ibid.* **45**, 148 (1987).
41. We are grateful to S. Sallan, J. Ritz, K. Loughlin, S. Shurtleff, P. Kourlas, F. Smith, the Cancer and Leukemia Group B, and Children's Cancer Group for providing valuable patient samples. We thank R. Klausner, D. G. Gilliland, D. Nathan, G. Daley, J. Staunton, M. Angelo, A. Leblanc, P. Lee, Z. Kikinis, G. Acton, and members of the Lander and Golub laboratories for helpful discussions. This work was supported in part by the Leukemia Society of America (T.R.G); the National Institutes of Health and the Leukemia Clinical Research Foundation (C.D.B); and Affymetrix, Millennium Pharmaceuticals, and Bristol-Myers Squibb (E.S.L).

27 May 1999; accepted 17 September 1999

Sequencing Complex Polysaccharides

Ganesh Venkataraman,[1] Zachary Shriver,[2] Rahul Raman,[2] Ram Sasisekharan[2]*

Although rapid sequencing of polynucleotides and polypeptides has become commonplace, it has not been possible to rapidly sequence femto- to picomole amounts of tissue-derived complex polysaccharides. Heparin-like glycosaminoglycans (HLGAGs) were readily sequenced by a combination of matrix-assisted laser desorption ionization mass spectrometry and a notation system for representation of polysaccharide sequences. This will enable identification of sequences that are critical to HLGAG biological activities in anticoagulation, cell growth, and differentiation.

The chemical heterogeneity of polysaccharides, their structural complexity, and the lack of effective tools and methods have seriously limited the development of a sequencing approach that is rapid and practical, like that used for polynucleotides and polypeptides. This limitation is especially relevant in the study of glycosaminoglycan (GAG) complex polysaccharides, which are present at the cell surface and in the extracellular matrix (*1*, *2*). Heparin or heparan sulfate–like glycosaminoglycans (HLGAGs), a subset of GAGs, are currently used clinically as anticoagulants, and this function of HLGAGs has been assigned to a specific pentasaccharide sequence that is responsible for binding to antithrombin III (*3*). Recent progress in developmental biology, genetics, and other fields has resulted in a virtual explosion in the discovery of important roles for HLGAGs in the biological activity of morphogens (*4*) (for example, Wingless, Decapentaplegic, and Hedgehog); growth factors, cytokines, and chemokines (*5*); enzymes (*1*, *6*); and surface proteins of microorganisms (*7*). Although it is increasingly recognized that a specific sequence, typically from a tetra- to a decasaccharide in size, is responsible for HLGAGs' modulation of biological activity, in only a few cases is there any structural information regarding sequences (*8*). Therefore, accelerating our understanding of structure-function relationships for HLGAGs requires the development of rapid yet thorough sequencing methodologies.

There are many issues that have limited the development of sequencing techniques for HLGAGs. HLGAGs are chemically complex and heterogeneous, because the HLGAG chain can vary in terms of the number of disaccharide repeat units and possesses, within the disaccharide repeat unit, four potential sites for chemical modification. The basic disaccharide repeat unit of HLGAG is a uronic acid [α-L-iduronic acid (I) or β-D-glucuronic acid (G)] linked 1,4 to α-D-hexosamine (H) (Fig. 1A). Together, the four different modifications ($2^4 = 16$) for an I or G uronic acid isomer containing disaccharide give rise to $16 \times 2 = 32$ different plausible disaccharide units for HLGAGs. In contrast, four bases make up DNA, and 20 amino acids make up proteins. With these 32 building blocks, an octasaccharide could have over a million possible sequences, thereby making HLGAGs not only the most acidic but also the most information-dense biopolymers found in nature. There are no methods available to amplify or produce HLGAGs in large amounts, unlike the techniques that are available for DNA or proteins.

To handle the enormous information den-

[1] Harvard-MIT Division of Health Sciences and Technology, [2] Division of Bioengineering and Environmental Health, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: ramnat@mit.edu