SUPPLEMENTARY APPENDIX 1

Statistical Analysis

Missing or Censored Data

Ten percent of the microarray elements were censored because they could not be identified during image analysis or they had intensities (expressed in arbitrary units with the use of GenePix software [Axon]) of less than 500 in both channels and an intensity of less than 50 in one of the channels. Patients with missing values for a particular microarray element were excluded from all analyses involving that element.

Data on one or more of the components required to compute the international prognostic index were missing for 37 patients. For 19 of these patients, the addition of the missing component would not have changed the risk-group assignment according to the international prognostic index, and so it was still possible to make an unambiguous assignment. The remaining 18 patients were excluded from all analyses involving the international prognostic index.

The Kaplan–Meier method was used to estimate all survival probabilities after the data were censored. The analysis included death from any cause.

Identification of Subgroups

To classify biopsy samples, a two-sided t-test was used to identify genes whose levels of expression differed significantly ($P<0.001$) between activated B-cell–like and germinal-center B-cell–like tumors in a previous analysis.[3] Of these genes, 100 were represented on the Lymphochip microarrays used in the present analysis. The samples were divided into three subgroups — activated B-cell–like, germinal-center B-cell–like, and type 3 diffuse large-B-cell lymphoma — according to the hierarchical clustering algorithm.[9] Two-sided t-tests were also used to identify genes whose levels of expression differed significantly among these three subgroups.

Formulation of Preliminary and Validation Groups

Two hundred forty patients were randomly assigned to the preliminary or validation group according to the following approach. At each participating institution, two thirds of the patients were randomly assigned to the preliminary group and one third were randomly assigned to the validation group, with the constraint that none of the 13 patients whose data were censored before three years were included in the validation group, in order to maximize the ability of this approach to evaluate the usefulness of the gene-expression–based outcome predictor. In addition, 30 patients from the previous study[3] again underwent gene-expression analysis and were included in the preliminary group in order to make the validation group fully independent of the previous study. No statistically significant differences ($P<0.05$) were found between the preliminary and validation groups with respect to the components of the international prognostic index. Furthermore, the overall survival rates did not differ significantly between groups ($P=0.75$).

Identification of Significant Variables

To determine the statistical significance of the large number of genes that were associated with good and poor prognoses, a permutation test was used. The associations between the level of gene expression in samples from individual patients and overall survival were permuted with use of a random-number generator (Splus, Insightful). We then fitted a univariate Cox model to each gene to determine the gene's association with the permuted information on survival and counted the number of genes that were significantly associated with either a good or a bad prognosis according to a one-sided Wald test ($P<0.01$). This procedure was repeated 4000 times, and in only 20 instances were there as many significant genes as were found in unpermutated data. Hence, a P value of 0.005 (20 of 4000) was reported.

Formulation of the Gene-Expression–Based Outcome Predictor

We defined the variance in the expression of a particular gene as the variance of the log ratio of the levels of expression of that gene across the samples of the preliminary group.

To be included in the gene-expression–based outcome predictor, a gene had to be represented by at least one microarray feature that met the following conditions with respect to the preliminary group: data on the microarray feature were significantly correlated with survival ($P<0.01$) according to a two-sided Wald test for the proportional-hazards model; data on the microarray feature were available for at least 90 percent of the patients; and the gene-expression variance for the gene was in the upper 33rd percentile of such variances. For a more accurate estimate of the level of gene expression and to reduce the number of missing values in our combined model, we calculated a single level of expression for each gene on the microarray as follows. For each microarray feature, the median value in the overall group of patients was determined and subtracted from each patient's value for that feature (i.e., the data were median-centered). Next, multiple microarray features representing a given gene were averaged. If data on one or more of the features were missing for a given gene, the features with data were averaged to provide a value. We then checked the averaged value

to determine whether it still met the conditions for significance and variance described above. This approach resulted in the identification of 35 genes. Of these 35 genes, 9 were from the major histocompatibility complex (MHC) class II gene-expression signature, 3 were from the germinal-center B-cell gene-expression signature, 3 were from the proliferation gene-expression signature, and 6 were from the lymph-node gene-expression signature. The remaining 14 genes were not associated with any particular gene-expression signature. The values for the 21 significant genes within each cluster were then averaged to yield a germinal-center B-cell, lymph-node, and proliferation gene-expression signature value for each patient. Since the MHC class II signature genes were very highly correlated, we believed that this group could be adequately represented by averaging only the four MHC class II genes that were most highly correlated with survival.

We used a Cox proportional-hazards model to determine which variables should be included in the gene-expression–based outcome predictor. Briefly, a multivariate Cox model is fitted by finding coefficients that best describe the effect of a given variable on censored survival data. A positive coefficient is associated with a variable for which high values are correlated with a low likelihood of survival. A negative coefficient is associated with a variable for which high values are correlated with a high likelihood of survival. To calculate a risk score for a given patient, we multiplied the patient's values for each variable in the model by the corresponding coefficients and then totaled the values. The risk score can be used to predict the patient's likelihood of survival relative to that of other patients.

An initial multivariate Cox model was fitted with the four gene-expression signature values. The P values reported for this model were based on a likelihood-ratio test. The remaining 14 genes were added to the model in a stepwise fashion until no genes were found that significantly added to the model ($P<0.05$ by the likelihood-ratio test). This approach resulted in a five-variable Cox model (i.e., the four gene-expression signatures plus the single gene *BMP6*), which we then used to calculate the risk scores for each patient in the preliminary group. For purposes of exposition, patients were ranked according to this score and then divided into four equal groups, or quartiles.

We used the same approach to calculate the outcome-predictor components for each patient in the validation group. To validate the model as a whole, the Cox-model coefficients were not refitted. Instead, coefficients that were determined in the preliminary group were applied to calculate the outcome-predictor score for each patient in the validation group. The patients in the validation group were then divided into four quartiles.

Once the model was validated, a final five-variable Cox model was fitted to the total data. The coefficients of the Cox model were as follows: germinal-center B-cell signature, −0.290; MHC class II signature, −0.311; lymph-node signature, −0.249; proliferation signature, 0.241; and *BMP6,* 0.310. We calculated a patient's gene-expression–based outcome-predictor score using the Cox-model coefficients as follows: (0.241 × proliferation-signature value) + (0.310 × *BMP6* value) − (0.290 × germinal-center B-cell signature value) − (0.311 × MHC class II value) − (0.249 × lymph-node signature value).

Methods for Calculating P Values

The chi-square test was used to calculate all P values related to the difference between the subgroups. To calculate the P values for the differences in the incidence of the histologic subtypes between the subgroups of diffuse large-B-cell lymphoma, either a chi-square test or Fisher's exact test was used. The chi-square test was used for subtypes identified in 20 or more samples (centroblastic, immunoblastic, and unclassified). Fisher's exact test was used for subtypes identified in fewer than 20 samples (Burkitt-like, plasmablastic, T-cell–rich, and anaplastic). The P values for the significance of signatures were calculated from univariate Cox models with use of the Wald test. For the preliminary group and the overall group of patients, a two-sided t-test was used. For the validation group, a one-sided t-test was used that was based on the direction of the effect of that gene in the preliminary group.

To determine the significance of the model in relation to the international-prognostic-index scores, we devised a multivariate Cox model in which the gene-expression–model value was included as a single continuous variable and the international-prognostic-index risk score was included as a categorical variable with three values (0 to 1, 2 to 3, and 4 to 5). The Wald test was used on the gene-expression component of this model to generate P values as well as confidence intervals for the relative risk.

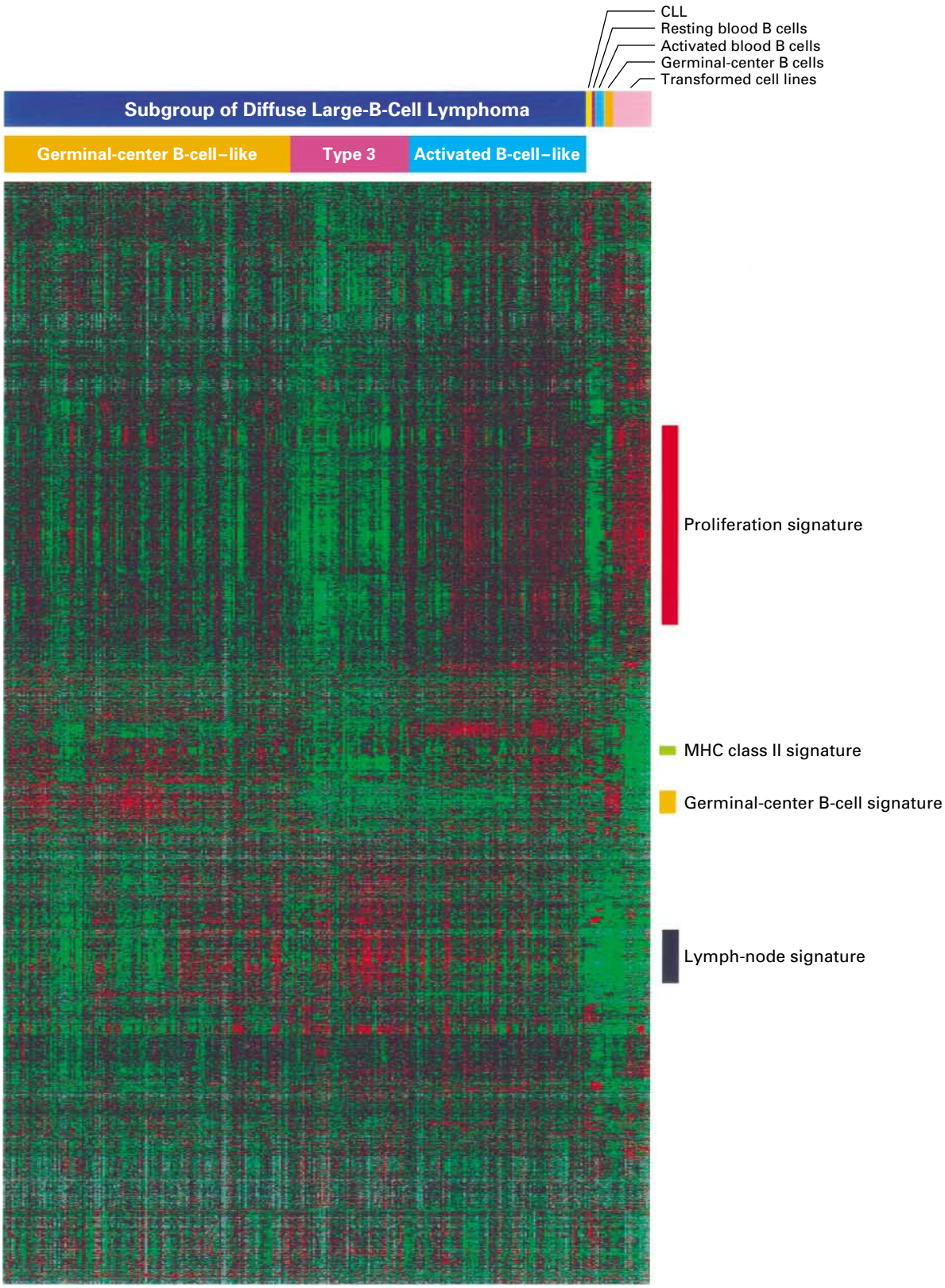

Figure 1. Gene-Expression Signatures Defined by Hierarchical Clustering.

Gene-expression data from 305 Lymphochip-microarray experiments are shown. A total of 7399 microarray elements representing approximately 4128 genes were organized with the use of hierarchical clustering based on the level of expression in the following numbers of samples: 274 samples of diffuse large-B-cell lymphoma; 3 samples of chronic lymphocytic leukemia (CLL); 2 samples of resting blood CD19+ B cells; 4 samples of blood CD19+ B cells stimulated with IgM antibodies (1 sample each obtained after 1 hour, 3 hours, 6 hours, and 24 hours of stimulation); 2 samples of germinal-center B cells (1 of CD77+ centroblasts and 1 of CD77− centrocytes); 4 samples of germinal-center B-cell–like diffuse large-B-cell lymphoma cell lines (SUDHL10, DB, SUDHL4, and SUDHL6); and 2 samples of activated B-cell–like diffuse large-B-cell lymphoma cell lines (OCI-Ly3 and OCI-Ly10). The four gene-expression signatures used in the outcome predictor are indicated at the right. MHC denotes major histocompatibility complex.