

Supplementary Appendix 1

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Dave SS, Wright G, Tan B, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N Engl J Med* 2004;351:2159-69.

Supplementary Appendix. Statistical Methods.

Measuring Gene Expression

Samples were normalized using Affymetrix MAS 5.0 software to a median expression level of 500 (in MAS arbitrary units). All the data were \log_2 -transformed for display and analysis. Array elements with median signal intensities of less than 6 \log_2 units across the samples (approximately 25 percent of array elements) were removed from the analysis entirely. This was in order to exclude poorly measured genes and genes not appreciably expressed in a sample. Primary gene expression and clinical data for all the analyses in this study are available at <http://llmpp.nih.gov/follicularlymphoma>.

Creating the Gene-Expression–Based Survival Model

1. General Analytical Approach

The survival analysis included death from any cause. Of the available 191 samples, 4 samples lacked associated survival data. The remaining samples were divided into a training set of 95 samples and a test set of 96 samples, each of which included 2 samples without clinical data. All aspects of model development were done using only data from 95 samples in the training set. Once the model was fully developed, it was tested on the independent test set in order to verify its reproducibility.

2. Dividing the Patients into Training and Test Sets

The patients were ranked by institution and then by length of survival by institution as measured by the Cox residual under the null model. The patients were then alternately assigned to the training or test sets. The four patients for whom survival data were not available were assigned randomly to either the training or test sets. No other molecular or clinical features of the available data were used in determining membership in the training or test sets. There were no subset analyses of the test set. Apart from using the length of survival to assign membership to the training or test set, the test set was not previously analyzed for survival in any way.

3. Survival Signature Analysis

To identify genes associated with survival in the training set, the Cox proportional hazards model was used with survival data only from the training set. A total of 3299 genes were identified by a Cox model as being associated with survival in the training set at a significance level of $P < 0.1$ using a Wald test. We deliberately chose a somewhat liberal cutoff for statistical significance, recognizing that the individual genes were potentially noisy surrogates of underlying biologic processes.

The genes associated with good prognosis (1568 genes) and poor prognosis (1731 genes) in the training set were hierarchically clustered separately. Each gene signature was defined using Treeview software as a cluster that met the following criteria: a centroid-correlation cutoff of at least 0.5 as calculated by the software, and the presence of at least 25 genes and no more than 50 genes in the cluster. The cutoff for minimum correlation was chosen arbitrarily in order to identify clusters of genes with highly correlated expression patterns. Using these criteria, we noted five gene signatures among the genes predicting good prognosis and five signatures among the genes predicting poor prognosis. Within each signature, the expression levels of the component genes were averaged, resulting in 10 signature averages associated with each sample. The 10 gene-expression signature averages were used to create multivariate Cox proportional hazards models of survival in the training set. We first tested all two-variable models and noted that a combination of two signatures had a strikingly higher log likelihood ratio than that of either signature alone. The signatures were later named immune-response 1 and immune-response 2 based on the known function of the constituent genes and further molecular characterization. The two-variable model formed from the immune-response 1 and immune-response 2 signatures was the statistically best two-variable model that could be created from the signatures. The striking synergism between two signatures, immune-response 1 and immune-response 2 made it clear that these two signatures would be part of any final model. For this reason, we decided to start with the two signatures as the base model and then used a “step-up” procedure²⁰ to select additional signatures that would add to the predictive power of the model in a statistically significant fashion. Of the remaining eight signatures, only one signature added significantly to the combination of immune-response 1 and immune-response 2 ($P < 0.001$). No other signatures added significantly to the model.

Thus three signatures were linearly combined to create a multivariate Cox model of survival in the training set. The multivariate model proved to be strongly associated with survival in both the training set ($P < 0.001$) and test set ($P = 0.003$). However, within the test set, only the immune-response 1 and immune-response 2 signatures contributed to the model in a statistically significant fashion. The remaining signature was then dropped from the combination and the final model consisted only of the two immune response signatures. The coefficients in the final model were derived from the Cox model applied to the training set. For each sample, the final model generated a survival-predictor score according to the formula: survival-predictor score = $(2.71 \times \text{immune-response 2 signature average}) - (2.36 \times \text{immune-response 1 signature average})$. A high survival-predictor score was associated with poor outcome. The survival-predictor scores were calculated for the test set cases using the same formula, without reoptimizing the coefficients of the Cox model. The survival-predictor score was highly associated with survival in both the training and test sets ($P < 0.001$). The highly significant results show that the survival data on the test set shared the same relationship to the clustered genes as it did on the training set, thus verifying the reproducibility of the model.

All P values were two-sided. The P values shown in Figure 2C are the result of Wald tests applied to the model score as a continuous variable. The model fitted to the training set was treated as a single variable. International Prognostic Indicator (IPI) risk group was treated as a categorical variable with three levels. To determine the significance of its relation to survival, the log-rank test was employed to calculate the P value shown in Figure 2B.

The likelihood-ratio test was used to compare the degree of association with survival of the signature averages individually and in combination. The likelihood ratios were computed by comparing the likelihood ratio under the Cox model including the variables, to the likelihood ratio under the null model. The synergy between the immune-response 1 and immune-response 2 signatures was also a feature of the test set, as well as the overall group.

The P values, relative risks and confidence intervals shown in Tables 1 and 2 were computed using a Wald test. In column 5 and 6 of Table 1, this Wald test was applied to most of the clinical variables in a univariate fashion. The exceptions were grade and IPI. For these clinical variables, the relative risk is presented with respect to grade = 1 or IPI score = 0 or 1, P values reported for grade and IPI reflect the effects of these clinical variables taken in their entirety. In columns 7 and 8 of Table 1, the statistics were from a multivariate test that included both the

survival-predictor score and the clinical variable, with the relative risk and P value of the survival-predictor score being reported. In Table 2 the Wald P values and relative risks were calculated within a binary model that included the immune response-1 and immune response-2 signatures. The relative risk and P values reported are for the effect of each signature individually within the binary model.

Survival and the International Prognostic Index

Thirty-four samples lacked associated data on one or more IPI components. For eight of these, the missing IPI component would not have changed the risk group assignment and thus an unambiguous assignment to an IPI risk group was still possible. The remaining 26 samples were excluded from all analyses involving the IPI.