Ensemble classification of DLBCLs

The prediction of several (n=14) classifiers was combined to assign each of the cell lines to one of the consensus clusters (Table S2). The cluster membership of each cell line was determined by a simple (unweighted) majority vote, with cell line *L* assigned to cluster *C* if a majority of the classifiers classified *L* as belonging to *C*. Classifiers predicting a sample with less than probability *p* were excluded from the count for that sample. Similar assignments were obtained by setting *p* to different values (0.5, 0.6, 0.9).

Voting classifiers

The classifiers used were: naïve-Bayes [NB] [2]; K-Nearest-Neighbor [KNN] [3]; *Linear* and *Quadratic Discriminant Analysis* [LDA/QDA] [4]; *shrunken centroid classifier* [pamr] [5]; a classifier based on *hierarchical clustering* [6]; a simple centroid-based classifier; a multi-class generalization of the classifier based on *linear predictive scores* adopted in [7]; a *random forest* classifier [8]; and a *support vector classifier* [9,10]. Below are further details about some of the classifiers used and detailed descriptions can be found in the included references.

- The NB classifier models the class-dependent probability of the predictors/genes as univariate Gaussian distributions each with independent mean and standard deviation:

$$P(g_i \mid C = c) \sim N(\mu_{ic}, \sigma_{ic}^2), \quad c = \{Oxphos, BCR, HR\}$$

- The LDA and the QDA classifiers are generalizations of the NB classifier where the probabilities of the predictor genes are modeled as multivariate Gaussian distributions with class-dependent means, and a common class-independent covariance matrix in LDA, and a class-dependent covariance matrix in QDA.

$$P(\{g_1, g_2, ..., g_n\} \mid C = c) \sim N(\boldsymbol{\mu}_c, \Sigma), \quad LDA$$

$$P(\{g_1, g_2, ..., g_n\} \mid C = c) \sim N(\boldsymbol{\mu}_c, \Sigma_c), \quad QDA$$

Since both LDA and QDA work best with a limited number of predictors, the dimensionality of the gene expression data was first reduced by principal component analysis (PCA). The number of components necessary to explain at least 85% of the variance was selected for prediction. It should be emphasized that PCA was applied to the train and test data combined (since PCA does not use the class labels).

- Two KNN classifiers were used, using K=3 and K=5 nearest neighbors and a distance-weighted voting scheme.
- The simple centroid-based classifier defines a class *centroid* as the average expression of each gene within that class. It then assigns a new (unknown) sample to the class whose centroid is closest as measured by rank correlation between the sample profile and the class centroid.
- The classifier based on hierarchical clustering clusters a new sample together with the samples in the training set. The new sample is then assigned to the class that has (relative) majority representation within the cluster in which it is contained.

Data preprocessing

LOOCV on the training set was also used as the criterion to choose the type of data pre-processing to adopt. We tested four different pre-processing schemes: i) RMA signal extraction followed by 2^x transformation (RMA2, for short); ii) RMA2 signal extraction followed by rank transformation (i.e., by replacement of the expression values with their within-chip ranks); iii) MAS5 signal extraction; and iv) MAS5 signal extraction followed by rank transformation. As shown in Figure 1, based on the LOOCV error rates, RMA2+rank transformation (rma2.rank)

performed best, and it was thus adopted to process both the training set and the test set. Additionally, the test set was scaled so that the distribution of each gene was the same as in the training set. That is, if $\mu_{g,trn}$ and $\sigma_{g,trn}$ are the mean and standard deviation of gene *g* in the training set, and $\mu_{g,tst}$ and $\sigma_{g,tst}$ are the mean and standard deviation of the same gene in the test set, each value *x* of *g* in the test set was transformed as follows:

$$x' = \frac{x - \mu_{g,tst}}{\sigma_{g,tst}} \times \sigma_{g,trn} + \mu_{g,trn} .$$

REFERENCES

1. Monti S, Savage KJ, Kutok JL, et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*. 2005;105:1851-1861.
2. Duda RO, Hart PE. *Pattern Classification and Scene Analysis*. Wiley, New York. 1973.
3. Mitchell TM. *Machine Learning* McGraw Hill, New York. 1997.
4. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Springer, New York. 2001.
5. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*. 2002;99:6567-6572.
6. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*. 1998;95:14863-14868.
7. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. *N Engl J Med*. 2002;346:1937-1947.
8. Breiman L. *Machine Learning*. McGraw Hill, New York. 2001.
9. Burges CJC. *Data Mining Knowledge Discovery*. 1998.
10. Vapnik VN. *Nature of Statistical Learning Theory*. Springer, New York. 1995.