Supporting information for Elenitobna-Johnson *et al.* (2003) *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.1137463100

Supporting Text

Introduction

We have investigated the variations in gene expression associated with the two stages of FL: low grade (category A) and DLBCL (category B). A notable feature of our study is that we compare gene expression data from the same group of subjects in both stages and identify genes whose expression typically either decreases or increases for every subject in the transformation from low-grade FL to DLBCL, as well as genes that distinguish all samples of one category from a sample of the other.

For the identification of genes involved in FL transformation, we performed two types of experiments. The first set of experiments (indirect) entailed pairwise analysis of the hybridization data obtained from the FLs and DLBCLs, both compared to the reference RNA samples consisting of pooled RNA from five cell lines (Jurkat, SKW-3, L428, Raji, and NCEB). The raw data pertaining to the indirect experiments for the FLs are accessible at www.path.utah.edu/labs/kojo/KA.txt. The raw data for the indirect DLBCL experiments are at www.path.utah.edu/labs/kojo/KB.txt. The second set of experiments (direct) involved direct hybridization of cDNA obtained from a FL against the cDNA obtained from the corresponding DLBCL obtained from same patient. The raw data for this analysis are available at www.path.utah.edu/labs/kojo/direct.txt. We performed a consistency analysis between both the direct and indirect methods and found a reasonable agreement between the results of both analyses (see below).

Consistency Analysis of Direct and Indirect Measurement of Expression Variation

We analyzed the consistency of the direct microarray measurements of differential expression of genes between each patient in their aggressive (DLBCL) and indolent (FL)

phases (i.e., 1B vs. 1A, 2B vs. 2A,…) and the indirect measurement of the same ratio (1B vs. reference mRNA) vs. (1A vs. reference mRNA) and found reasonable agreement.

$$l_i(g,e) = log_2\left(f(g,e)/f_r(g,e)\right) - \overline{log_2\left(f(g,e)/f_r(g,e)\right)}$$

where f / f_r is the fluorescence ratio relative to reference composite, and the bar indicates the median over all experiments, e, with respect to a particular gene, g. This is the indirect case.

When direct comparisons were made, the value represents

$$l_d(g,e) = log_2\left(f(g,e_B)/\ f(g,e_A)\right)$$

where e_B and e_A represent the DLBCL and FL stages of an experiment. The direct measurement has absolute significance, and the common normalization of the (logs of the) two indirect measurements drops out under the subtraction used to compare with the direct measurement.

In Fig. 7, we have plotted the number of gene and sample pairs for which the differential in the log of this ratio is in a given interval [x, x + 0.1) so each bar in the graph is associated with the left endpoint of the interval. The endpoints range from –3 to +2 by increments of 0.1. The overall mean was –0.94 and standard deviation was 1.31, indicating that, in general, the indirect measurements tended to give a slightly smaller ratio than the direct measurements. If the correlation was perfect, all genes and experiments would be associated with the bar whose left endpoint is 0.

Data Processing, Normalization, and Scaling

The data on which our analysis is based comes from samples of both low-grade FL (indolent) and DLBCL (aggressive) tumors from each of 12 subjects and is recorded in four microarray data files available in the directory at www.path.utah.edu/labs/kojo.

These are KA.txt, KB.txt, KC.txt, and direct.txt. These addresses contain the raw microarray expression ratios indicated from 6,912 spots from 12 subjects with the FLs labeled 1A-12A vs. reference mRNA (KA.txt), the same 12 subjects with DLBCL at the stage labeled 1B-12B vs. reference mRNA (KB.txt), various purified lymphocyte subpopulations and lymphoma cell lines vs. reference mRNA (KC.txt), and direct expression ratios of nA vs. nB, n = 1-12 (direct.txt). Each of these the values were measured with four hybridization experiments, and in all analyses except when determining reference genes, the median value of the four hybridization experiments was used. When spots representing identical clones were present, the median of their median values was determined to give distinct gene clones equal weight, leading to 6,154 distinct clones upon which the analysis proceeded. The data were $\log_2$-transformed to provide a more natural additive Euclidean geometric setting than working directly with multiplicative factors and log-normal distributions would. A different base would merely scale all values equally, and the base two setting acquires a natural significance in the context of PCR analysis of expression. However, measurements in one subject (9B) were repeatedly unsatisfactory in the indirect experiments, so valid data for both types of tumor cell existed on 11 of the 12 pairs. Nevertheless, the results of the direct experiments for this sample remained valid.

Finally, the values for each gene were centered relative to the median value for that gene. Given the equal presence of both types of tumor cells, this does not bias the results in the way a less balanced sample might. And for this reason, we median-centered with respect only to the matched lymphoma pairs only samples. Thus, values of –2, –1, 0, 1, or 2 indicate that the median expression of a certain gene in a certain tumor cell are, respectively, ¼, ½, 1, 2, or 4 times the median expression in each gene. This normalization convention has been used in many other gene expression analysis references. The only situation in which the data was not normalized was in analyzing the

consistency of the direct and indirect measurements of relative expression of genes between aggressive and indolent stages from the sample patient. Here, the direct measurement has absolute significance, and the common normalization of the (logs of the) two indirect measurements drops out under the subtraction used to compare with the direct measurement.

Although most analyses we performed took Pearson, Euclidean, and sometimes uniform/L^{∞}, metrics into consideration, we usually favored Euclidean measurements because they seemed to more accurately reflect the biological distinctions observed. In particular, in many cases, the location of the two different histologic categories with respect to various gene expression levels was distinct in Euclidean space, but less so on the unit sphere (i.e. in normalized direction alone). Although Pearson analysis was helpful when looking at large numbers of genes simultaneously to cancel uncorrelated noise, when considering few genes, we found that samples whose logarithmic expression vectors were large multiples of each other were not necessarily more biologically alike than those with common magnitude in a slightly different direction.

Selection of Differentially Expressed Genes Involved in FL Transformation

Genes and gene-tuples (i.e. pairs and triples), which were sensitive to the distinction between the two stages of lymphoma, were identified and selected on the basis of both the direct (FL vs. corresponding DLBCL) and indirect (FL vs. reference, compared with corresponding DLBCL vs. reference) experimental data set by any of several criteria.

The magnitude and direction of each gene's change from a particular sample in the indolent FL histology to the DLBCL histology can be determined from either the direct or indirect measurements. We performed such rankings with both data sets and also performed a general consistency analysis as indicated in Fig. 7.

The direct aggressive-to-indolent ratios cannot be used to determine collective separation of the indolent and aggressive classes, but we could determine this from the indirect

measurements. For example, although every DLBCL sample expressed *Alpha 1* less in comparison to its corresponding FL sample, not all DLBCL samples expressed *Alpha 1* higher than every FL sample. Conversely, while no individual gene did separate the two categories collectively, pairs and triples of genes that did so most effectively were not necessarily those that changed most uniformly with respect to individual pairs. The metrics and other ranking criteria were:

Pearson

$$\frac{\frac{1}{N}\sum_{j=1}^{N} u_j v_j}{(\frac{1}{N}\sum_{j=1}^{N} u_j^2)^{1/2}(\frac{1}{N}\sum_{j=1}^{N} v_j^2)^{1/2}}$$

where j is an index for the N genes (dimensions) in which two experiments both have valid expression values.

Euclidean

$$(\frac{1}{N}\sum_{j=1}^{N}(u_j - v_j)^2)^{1/2}$$

where j is an index for the N genes (dimensions) in which two experiments both have valid expression values.

Uniform / L^{∞}

$$\max_j |u_j - v_j|$$

where j is an index for the N genes (dimensions) in which two experiments both have valid expression values.

In the case of nonparametric rankings, where there were necessarily several genes ranked equally, for example having the same number of experiments in which expression increased, these genes were ordered by other criteria described here, or arbitrarily if such criteria were irrelevant.

In general, the rankings were based on the three metrics above and pairwise or typewise comparison. In pairwise comparison, the expression values for each particular patient in FL and DLBCL stages were compared. In typewise comparison, the collective values of all FL samples and all DLBCL samples were compared, for instance, in uniform analysis, by their maximum and minimum values. If the minimum of one type was greater than the maximum of the other, then the values of all samples would be separated by that gene.

Along with biological relevance, other selection criteria used were rankings of:

Greatest number of increasing or decreasing experiments from FL to DLBCL.

$$\left|\{e\}\right| \quad \left| \; l_d(g, e_B) \rangle\, l_d(g, e_A) \right.$$

Greatest number or above or below median expressing genes in indolent (FL) experiments

$$\left|\{e\}\right| \quad \left| \; l_d(\mathrm{g}, \mathrm{e_A}) \rangle\, 0 \right.$$

Greatest magnitude of uniform increase or decrease from FL to DLBCL.

$$\min_{g} \min_{e} \; \boldsymbol{l_d}(\boldsymbol{g}, \boldsymbol{e_B}) - \boldsymbol{l_d}(\boldsymbol{g}, \boldsymbol{e_A})$$

for maximum uniform increase and analogous expression for decrease, etc.

The smallest ratios of the Euclidean distance to the *k*th nearest distinct neighbor of a common type over the distance to the nearest neighbor of the opposite type, based on expression levels of genes, gene pairs, and gene triples.

$$\min_{g} \min_{e} \frac{\left\| e_X - e'_X \right\|_g}{\left\| e_X - e'_{\overline{X}} \right\|_g}$$

where $\| . \|_g$ represents the Euclidean norm restricted to a particular subset of genes *g* and e'_X is the *k*th nearest neighbor of the experiment e_x of type X = A, B of the same type, and $e'_{\overline{X}}$ is the nearest neighbor of the opposite type.

As above, we also used several variations on this ranking, with different values of *k* and rankings of the fewest FL-DLBCL experiment pairs (unmatched) violating linear separability in some direction, and their parametric analogues.

Instructions for Interactive 3D Animation

Interactive 3D animation is available at www.path.utah.edu/labs/kojo/lymph.htm. Click on START to view 3D sample distribution of FLs of DLBCLs. The blue spheres represent the FLs and the red spheres represent the DLBCLs. The black spheres represent the axes. To animate, position the cursor on or within the circumference of the circle. Left click on or within the circumference while holding mouse down in the clicked position and move the mouse around slowly in any desired axis. To magnify, press the ↑ key, and repeat until the desired size is achieved. Press the ↓ key to return to original size or to diminish image size.

RT-PCR Conversion and Reference Gene Analysis Selection

To compare the RT-PCR measurements of expression levels of the test set with microarray measurements of the learning set, we used a control gene whose expression

level is highly stable independent of a particular gene to adjust for small variations in total mRNA present in the tissue samples. Thus, the control gene serves to measure how much mRNA is present, and the difference in C_t, the critical cycle number to maximum-second-derivative between the control gene and other genes of interest provides reasonable estimate of relative expression per fixed amount of sample mRNA. We have taken cycle number difference to be an adequate indicator of relative initial copy number. The results are rescaled to the microarray data empirically and collectively by using a linear function fit by least squares to benchmark samples 3A, 3B; 6A, 6B from the learning set, which were measured in both ways to account for variations in efficiency and saturation level that can affect the relationship between C_t and estimated copy number. The estimated copy number can also be adjusted for each sample directly by incorporating the effects of efficiency and saturation level into a logistic model of the curves and using a nonlinear least-squares fit.

$$y = M \Big/ \left(1 + Ce^{-at}\right)$$

The reference gene erythrocyte membrane protein 4.1 like-1 (*EMP 4.1-L1*) was chosen based on its minimal spread in both least squares and uniform senses (L^2 and L^∞). These two quantities were ranked with respect to the four-sample median and all four samples for the learning set, and the learning set plus the additional cell lines, for a total of eight different rankings. EMP 4.1-L1 was favored in multiple rankings, and it varied on the order of 100-fold less in RT-PCR comparisons with GAPDH.

Bayesian Classification

To classify the test set we used the simplest Bayesian classifier obtained by comparing the multivariate distributions (1).

$$f_k(x) = (2\pi)^{-N/2} \left|\boldsymbol{R_k}\right|^{-1/2} exp\left[-\frac{1}{2}(x - \boldsymbol{m_k})^{\boldsymbol{T}} \boldsymbol{R_k}^{-1}(x - \boldsymbol{m_k})\right]$$

where $\boldsymbol{m}_k$ and $\boldsymbol{R}_k$ are the mean vector and covariance matrices of each class, based on the expression values of the learning set in the classifier genes, assuming equal prior probabilities and misclassification cost for the two types of lymphoma. $\boldsymbol{R}_k^{-1}$ is assumed to exist and $|\boldsymbol{R}_k|$ is the determinant of $\boldsymbol{R}_k$ and T donates the transpose.

We estimated the means and covariance matrices of each class based on the learning set, and assuming equal prior probabilities of both classes, and equal misclassification cost, assigned estimated relative likelihoods for each test sample to belong to the indolent (FL) or aggressive (DLBCL) classes based on the relative magnitudes of the estimated distributions corresponding to the (converted) measured values.

The RT-PCR data were converted to correspond to the microarray scale by fitting data from the benchmark samples for which both values were known by using the linear least-squares model.

1. Young, T. & Calvert, T. (1974) *Classification, Estimation, and Pattern Recognition* (Elsevier, New York).