

Classification of cancers by gene expression profiles from peripheral blood

Andrey Loboda, Michael Nebozhyn, Steven W. Johnson*, Peter J. O'Dwyer#, Calen Nichols, Linda Alila, Louise C. Showe, and Michael K. Showe

The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104 *Dept. of Pharmacology, University of Pennsylvania #Dept of Hematology/Oncology, University of Pennsylvania

ABSTRACT

We have analyzed gene expression in peripheral blood lymphocytes from patients at early stages of solid tumors (not blood related) and normal controls. The samples from patients with solid tumors can be easily distinguished from controls by hierarchical clustering or supervised procedures including shrunken centroids or penalized discriminant analysis (PDA). In most cases, a single sample randomly picked from each group is enough to provide correct classification for the rest of the samples. In spite of the consistency of the cancer group, it has a substantially greater variance than the control group. Part of this variance stems from a variable degree of progressive failure of the immune response in cancer patients. We have estimated the disease progression by projecting each sample on the AverageControl-AverageCancer vector, by the shrunken distance to control centroid, and by crossvalidation with PDA. To identify the most significant genes we have used either the median distance from the control centroid, or discriminant loadings obtained from PDA performed on Control-Cancer groups after the outliers have been removed. Among the genes changed the most with disease progression we identified two main groups. The first group, which contains genes associated with cell growth, actin remodeling, energy production, and mitosis, was identified by the highest scores obtained with both methods described above. The second group was identified by selecting genes associated with NK cells and cytotoxic T-cells, and demonstrating that all of these genes are significantly downregulated.

Materials and Methods

Solid tumor patients comprised 10 non-small cell lung carcinoma, 2 sarcoma, 2 pancreatic carcinoma, and 1 each esophageal, ovarian, small cell, adrenal, and mesothelioma, together with 9 normal controls. Approximately 8000 genes had a complete set of expression values across all patients and controls and these values were used for further analysis.

Results

Analysis of the distance matrix generated using Manhattan or Euclidian distances shows that for most of the samples the most distant member of sample's own group is closer than the closest member of the opposite group (Fig. 1). The only outlier is a lung cancer sample that is positioned exactly halfway between the cancer centroid and control centroid. But this sample is an outlier also relative to the controls: its distance to the control centroid is larger than for any control sample. We confirmed our findings about the distinction between cancer and control groups using two discriminant methods. First, all samples were analyzed with a multiclass shrunken centroids algorithm (1). In this analysis all genes and four samples from each of the two solid tumor groups and control group were used for training and the remaining samples were used as a test set. There is no misclassification of any solid tumor as a control or vice versa. Second, crossvalidation using PDA between control and cancer groups is 100% accurate. To estimate the extent of changes in individual patient samples compared to the control group, we have used the following three metrics. 1) Normalized and shrunken distance to control centroid; 2) Predictive scores for each patient obtained with PDA crossvalidation; 3) Projection of each sample on the Average Control-Average Cancer vector (Fig.2). Metrics 2, and 3 gave similar results.

We have used several methods to pick out the most informative genes. Although simple t-test between patients and controls identified a substantial number of differentially expressed genes, we did not use it as a primary metric because of the substantial variance in the patient group. Indeed, we observed that the average distance between cancer samples is much greater than the average distance between normal controls. The distribution of variance in gene expression is the lowest in the two sets of healthy controls, and the largest in lung adenocarcinomas and a group of "mixed cancers". Therefore we used only the variance in the control group to normalize the changes in gene expression and sorted genes by median Z score in the patient group, relative to the control centroid. Alternatively we have used discriminant loadings from PDA classification of Controls vs Patients. To reduce the variance in the most informative genes, we have removed the patient samples that were least advanced in the disease progression as described above. Finally, we have selected the genes known to be expressed exclusively in cytotoxic T-cells and NK cells and found them significantly downregulated.

Discussion

We observed drastic changes in gene expression from PBMC of cancer patients compared to normal controls. These changes are 1) much greater than both the noise level and biological variability of normal samples, and 2) consistent in all cancer patients, so that a single sample is enough to classify all the patients. The changes associated with disease progression were variable among patients and presented the main source of variance in the patient group. We have used several methods to estimate the extent of disease progression. All of the methods we used show that lung carcinoma patients have less change and are more variable as a group than the “mixed cancer” group which contains many different kinds of cancers. We will follow up the clinical progression of the patients to determine if there is a correlation between outcome and observed changes in gene expression. The changes of gene expression that we expect to see in the cancer patients can be associated with either changes in the cell composition or changes in gene expression in the existing PBMC subpopulations. Our data suggest that both are taking place and that changes in cell composition are not the main driving factor.

Table 1. Left: NK and Cytotoxic T-cells genes are significantly downregulated. Right: most significantly downregulated genes according to median Z test.

Symbol	t-test	Median Z score	Symbol	t-test	Median Z score
FCGR3A	6.59E-06	-15.2611	PTMA	7.36E-08	-75.4871
GNLY	4.83E-08	-11.7988	HNRPA1	2.78E-10	-37.839
PRF1	2.06E-05	-8.43536	RPL41	1.91E-07	-20.4659
CD8A	3.11E-09	-6.84261	SERPINB6	1.1E-07	-17.6533
CD2	3.44E-06	-6.28723	LDHA	5.38E-13	-15.2629
IL2RG	1.22E-06	-5.50725	TACC1	4.02E-09	-14.7357
GZMK	0.000137	-5.26209	ARPC3	8.56E-07	-12.9633
ITGAM	3.89E-07	-5.25155	CDC25B	2.16E-06	-12.824
PFN1	1.78E-09	-3.64608	CX3CR1	2.23E-06	-12.015
TIA1	0.003585	-3.31555	LDHB	7.67E-10	-11.4787
GZMB	1.93E-06	-2.27722	HNRPA2B1	1.34E-07	-11.3855
IL2RB	0.000158	-2.09404	CDC25A	9.08E-08	-11.2665

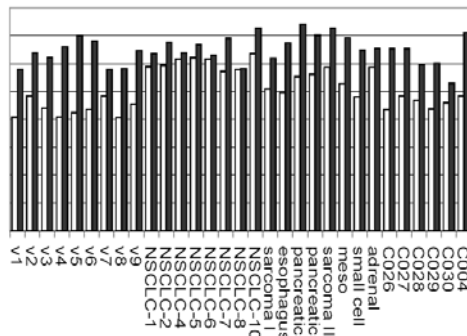


Fig. 1. Most distant samples from the same group are closer than the closest samples from different group.

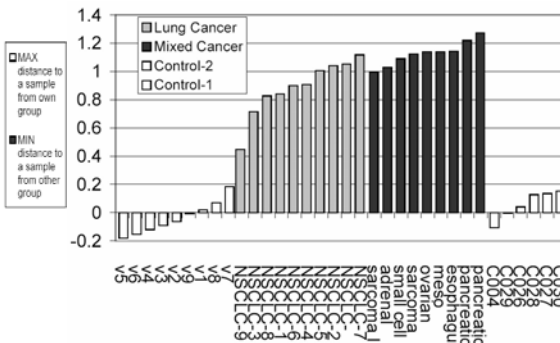


Fig. 2. Relative changes in PBMC as determined by the projection on the Control-Cancer axis.

References

1. Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* 99:6567-6572.