Detecting subtle differential expression in T cells of Melanoma patients using Microarray Experiments

Susan Holmes
Statistics Department, Stanford University
Collaboration with Peter P Lee from Stanford’s Hematology Department
Work supported by the ACS and the NSF.

Berkeley, October 23rd, 2003

Thanks to Bioconductor contributors, in particular Sandrine Dudoit, Jean Yee et al. for marrayNorm and Wolfgang Huber for vsn.
70 agilent arrays, 13,000 genes in 5 batches. Much batch to batch effect, no replication.

- five patients with melanoma, five healthy subjects.
- Stimulated and unstimulated T-cells of 7 different types.
T-cell sorting
The percentages of naïve, memory, and effector CD8+ subsets in each healthy donor or melanoma patient sample was determined by CD27 and CD45RA expression. There was no statistically significant differences ($p > 0.05$) in the relative distribution of the subsets either amongst subjects within each group or between melanoma and healthy subjects, confirming that melanoma patients do not have gross perturbations of their CD8+ T cell subsets. CD8+ T cell subsets were then isolated from PBMC samples by FACSorting based on their expression of CD27 and CD45RA, RNA was extracted, amplified, and hybridized onto Agilent microarrays.
Two channels

Mean intensities.
Channel 1: Green, reference mRNA mixture of Human cell types.
Channel 2: Red, T-cells, of 7 different types after sorting and amplification.
Array system noise determination and dye swap experiment

Self-self hybridizations. 500 ng aRNA (from sorted CD8+ cells) labeled with Cy3 or Cy5 were competitively hybridized against each other on microarrays.
The scatter plot shows that virtually all the features lie close to the diagonal line corresponding to the expected log ratio of 0. Overall the signal from the two channels is similar.

To test the difference of dye incorporation in probes, we performed dye swap experiments. The figure shows that among 12842 features, 2773 display good correlation with similar differential expression for both dyes, less than 1% features are differentially expressed, and about 97% of the genes were correlated to each other.
Combined microarray data from two microarrays representing the two halves of a dye swap experiment. In one experiment, the microarray was hybridized with cyanine-3 labeled T cell cDNA and cyanine-5 labeled reference cDNA generated from 500 ng of the corresponding cRNA. In the second experiment, the microarray was hybridized with cyanine-5 labeled T cell cDNA and cyanine-3 labeled reference cDNA. Together, the combined plot represents the two polarities in a dye swap experiment. The yellow points are genes that displayed good correlation with similar differential expression for both polarities of the dye swap experiment. Shown in blue are genes that are unchanged and in red are genes that were found to be differentially expressed in one polarity but unchanged in the other polarity. The pink points represent anti-correlated genes in the two polarities.
Data Normalization and Variance Stabilizing

- Variance Stabilizing Transformation (vsn) was used on the batches one at a time.

- Batch effect was taken out by removing the batch medians.
Discriminant Analysis between healthy and melanoma patients

Correction for cell effects allowed us to pool data from the three cell subsets, to maximize the statistical power of our dataset. In order to pool the cell types, they had to be homogenized, so cell type differences between the CD8+ cells were removed through a simple linear model on the transformed data.

We used linear discriminant analysis. Then the Westfall and Young multiple testing procedure was applied to detect the difference between patients and healthy donors. A subset of 50 genes with the smallest p-values was selected by taking the 50 features with smallest adjusted p-values. These were used as the input to a discriminant analyses with cross-validation that chose the best subset for discriminating the arrays into healthy and melanoma groups with low classification error.
Eleven features (10 genes) were found to consistently discriminate between T cells from melanoma patients versus healthy controls with adjusted p-values < 0.05.
<table>
<thead>
<tr>
<th>Abrev.</th>
<th>UniGene Name/Description</th>
<th>Adj. p-value</th>
<th>Rel. Expres.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPC1</td>
<td>Paraspeckle component 1</td>
<td>0.0002</td>
<td>↓</td>
</tr>
<tr>
<td>DDB2</td>
<td>Damage-specific DNA binding protein 2</td>
<td>0.0160</td>
<td>↑</td>
</tr>
<tr>
<td>VIL2</td>
<td>Villin 2 (ezrin)</td>
<td>0.0278</td>
<td>↑</td>
</tr>
<tr>
<td>LGALS1</td>
<td>Lectin, galactoside-bind.1 (galectin 1)</td>
<td>0.0278</td>
<td>↑</td>
</tr>
<tr>
<td>ITPKB</td>
<td>Inositol 1,4,5-trisphosphate 3-kinase B</td>
<td>0.0340</td>
<td>↑</td>
</tr>
<tr>
<td>LGALS1</td>
<td>Lectin, galactoside-bind.1 (galectin 1)</td>
<td>0.0340</td>
<td>↑</td>
</tr>
<tr>
<td>——</td>
<td>Hs, Simi.RNA binding motif prot., X</td>
<td>0.0492</td>
<td>↓</td>
</tr>
<tr>
<td>NMT2</td>
<td>N-myristoyltransferase 2</td>
<td>0.0492</td>
<td>↓</td>
</tr>
<tr>
<td>——</td>
<td>LOC154084; Hs clone mRNA</td>
<td>0.0492</td>
<td>↑</td>
</tr>
<tr>
<td>NME2</td>
<td>Non-metastatic cells 2, protein in</td>
<td>0.0492</td>
<td>↑</td>
</tr>
<tr>
<td>FLJ2205</td>
<td>Hypothetical protein FLJ22059</td>
<td>0.0492</td>
<td>↑</td>
</tr>
</tbody>
</table>
A hierarchical clustering of the data was performed using the mva package in R. The results are displayed here for the 11 retained features.
Paraspeckle component 1
H. similar to RNA bind
N−myristoyltransferase 2
Galectin 1
Galectin 1
FLJ22059
inositol 1,4,5−triph.
H.clone 24889
DDB2
protein (NM23B) non−metastatic.
Villin 2
The white are the most expressed, the red the least, with yellow as the intermediary. Note that the clustering separates the two groups neatly between melanoma and healthy patients. (b) Hierarchical Clustering of the best 9 apoptosis genes, this clustering, even though done using the best possible classifier is unable to decompose the two groups distinctly.

The same analysis was carried out on a set of 86 apoptosis genes, of which the 9 most discriminant were chosen, these 9 genes were then analysed using the same hierarchical clustering as the 10 overall best genes.

The figure shows that this smaller subset of genes classifies healthy donors and melanoma patients into two groups (only two outliers were found in the cluster, and they are just borderline cases). In order to check our procedure, we have used a cross validation method that leaves the known label of each array out in turn and builds a predictor from the remaining arrays. We can then compare to the known classification. This cross validation estimate of prediction error gave a 100% well classified
score, encouraging us to think that these 10 genes are indeed good predictors of T cells from melanoma versus healthy subjects. The expression for each gene in this subset is significantly different between healthy donors and melanoma patients ($p < 0.05$). This data demonstrates that CD8+ T cells (except CD45RA-CD27-) in melanoma patients are subtly different from CD8+ T cells from healthy donors.

As the clustering done only on known apoptosis genes shows, the distinction between melanoma and healthy patients cannot be made simply on the evidence of genes involved in the classical apoptosis pathway.
Check and recalibration

Male subjects were in a 2:1 ratio between the melanoma and healthy patients in this study, creating an expected fold difference that was observed on several specific Y-chromosome genes. When we performed the multiple testing procedure with the Y-genes re-inserted into the dataset, the SMC (mouse) homolog Y gene appears tied with galectin-1, with an adjusted p-value of 0.0342. This puts the level of the differential expression between the two groups of patients at around two-fold for the galectin-1 gene.
To confirm the significance of the genes identified to be differentially expressed using microarrays, we performed quantitative real-time PCR (RQ-PCR) analysis using the original, unamplified RNA materials. CD8 subsets were not combined in these experiments. Due to limited quantities of unamplified RNA, six of ten genes were selected for RQ-PCR analysis: DDB2, FLJ10955, Galectin-1, Myristoyltransferase, Hypothetical Protein 669. A housekeeping gene, GAPDH, was also amplified for normalization of data. The threshold cycle ($C_T$) for each sample is calculated by iCycler Real Time Detection System. After normalized with GAPDH, the $C_T$ of the six genes was analyzed using a Wilcoxon test. Significant differences were found in the CD8+ T cells between melanoma patients and healthy donors, matching the microarray data. The only exception was expression of gene FLJ10955, which did not show significant differences between healthy and melanoma in any of the three subsets.
<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Name/description</th>
<th>Naive</th>
<th>Effector</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSPC1</td>
<td>Paraspeckle component 1</td>
<td>0.53</td>
<td>0.0498</td>
<td>0.54</td>
</tr>
<tr>
<td>DDB2</td>
<td>Damage-specific DNA binding protein 2</td>
<td>0.28</td>
<td>0.27</td>
<td>0.027</td>
</tr>
<tr>
<td>LGALS1</td>
<td>Lectin, galactoside-binding,1(galectin 1)</td>
<td>0.048</td>
<td>0.041</td>
<td>0.021</td>
</tr>
<tr>
<td>——</td>
<td>Hs, Similar to RNA binding motif protein, X chromosome, clone, mRNA</td>
<td>0.08</td>
<td>6.10(^{-6})</td>
<td>0.69</td>
</tr>
<tr>
<td>NMT2</td>
<td>N-myristoyltransferase 2</td>
<td>0.27</td>
<td>0.0003</td>
<td>0.45</td>
</tr>
<tr>
<td>FLJ22059</td>
<td>Hypothetical protein FLJ22..</td>
<td>0.75</td>
<td>0.62</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 2: Significance of PCR validation of most differentially expressed genes. The p-values are computed by using a Wilcoxon two sample test comparing the melanoma patients and the healthy subjects.
Histogram of M3[, 12]
Parametric Bootstrap

Generate data under various null hypotheses that have many features similar to the original data:

- Same error distribution (form and moments).
- Same components.
- Vary some underlying tuning parameters.
VSN Noise Model: Additive and Multiplicative Components

\[ Y_{k, red} = a_{red} + v_{k, red} + \beta_{red} \gamma_k \eta_{k, red} \]

\(a_{red}\) and \(\beta_{red}\) are the offset and the factor. \(\gamma_k\) is gene specific and the other terms are the noise. additive and multiplicative, we make weak assumptions of symmetry of noise around 0:

\[
\sum_k v_{ki} = 0 \quad \sum_k \eta_{k, red} = 0 \quad \sum_\ell \eta_{\ell, green} = 0
\]

\[\eta_{\ell, green} + \eta_{\ell, red} = 0\]
Finding the right distribution for the residuals

Heavy tails, extreme case was from a different data set:
Non normal residuals

For this data, the recentering-rescaling was always better with mad and median.

\[ f_Y(y) = \frac{1}{2\phi} exp[-|y - \theta|/\phi], \quad \phi > 0 \]
Useful representation:

\[ Y = \sqrt{X} \cdot Z, \quad X \sim \text{Exp}(1), \quad Z \sim \mathcal{N}(0, 1) \]

A mixture of Normals whose scale parameters vary from an exponential.

The probe sequences are of different length and so have varying hybridization precisions.
The marginals each have asymmetric Laplace distributions:

\[ Y_i \overset{d}{=} m_i X + X^{1/2} \sigma_{ii} Z_i \]

where \( Z \sim \mathcal{N}(0, 1) \) and \( X \sim \text{Exp}(1) \).
Assymmetric Multivariate Laplace Distribution

Definition (Kotz et al):
A random vector $Y$ in $R^d$ has a multivariate asymmetric Laplace distribution (AL) if its characteristic function is given by

$$
\psi(t) = \frac{1}{1 + \frac{1}{2}t'\Sigma t - im't}, \quad t \in R^d.
$$

where $m \in R^d$, $m \neq 0$ and is a $dd$ non-negative definite symmetric matrix. This distribution is $AL_d(m, \Sigma)$.

Let $Y \sim AL_d(m, \Sigma)$ and let $Z \sim N_d(0, \Sigma)$. Let $X$ be an exponentially distributed r.v. with mean 1, independent of $X$. Then, the following representation is valid:

$$
Y \overset{d}{=} mX + X^{1/2}Z.
$$
and we have

\[ \text{Cov}(Y) = \Sigma + mm' \]
Random Number Generation

An $AL_d(m, \Sigma)$ generator (Kotz et al., 2003)

1. Decompose $\Sigma = CC'$ using Choleski decomposition.

2. Generate a standard exponential variate $X$.

3. Independently of $X$, generate multivariate normal $\mathcal{N}_d(0, \Sigma)$ variate $N$:
   - Generate iid normal vectors in a matrix of dimensions $n \times p$: $W$.
   - $N = C \times W$.

4. Set $Y = m \cdot X + \sqrt{X}N$.

5. Return $Y$
Parametric Bootstrap

Very useful in situations where there are not enough replicates.

Pretend that the data come from a parametric family $F_\theta$.

Replace in the generation of ‘new’ data, the unknown parameters by $F_{\hat{\theta}}$. 
Model, after renormalization, and eventual variance stabilization.

\[ m_{jk} = m_k \epsilon_k + \sqrt{\epsilon_k} Z_{kj}, \quad Z_{kj} \sim \mathcal{N}(0, \Sigma) \]

Genes are \( k, k = 1 \ldots n \). Estimate the parameters, covariance matrix, medians, mads.

This model is added onto the variance stabilization model.

One may want to change the covariance matrix to adjust for the hypotheses being tested.

Plug these into the Asymmetric Multivariate Laplace generator.

We can compare the number of genes chosen with the parametric model with a given covariance structure as the one we get with the data.
Calibration through Simulation Experiments

If we assume that the measured signal $y_{kj}$ increases, to sufficient approximation, proportionally to the mRNA abundance $c_{kj}$ of gene $k$ on the $j$-th array, or on the $j$-th color channel:

$$y_{kj} \approx a_j + b_j b_k c_{kj} + m_{kj}.$$  

(1)
Simulation of Data with same parameters

For red channel, and \( k = 1 \ldots 600 \) consistently expressed genes

\[
    r_k = a_{k,\text{green}} + b_{k,\text{green}} \times \sinh(\text{arcsinh}(l_k) + \epsilon_{k,\text{green},j})
\]

\[
    1/l_k \sim \Gamma(1, 1), \quad \epsilon \sim \mathcal{N}(0, c^2)
\]

All the parameters were estimated from the data \( c^2 \) is estimated from \( \text{Var}(h(Y) - \text{mean}(\mu_{k,j})) \).

For consistently expressed genes:

\[
\begin{align*}
    > \ lks &\leftarrow 1/\text{rgamma}(600,1,1) \\
    > \ \text{simuldata} &\leftarrow \text{matrix}(0,600,96) \\
    > \ \text{eps} &\leftarrow \text{matrix}(\text{rnorm}(600*96,0,0.25),600,96) \\
    > \ \text{areds} &\leftarrow \text{matrix}(\text{rnorm}(600*48,-0.5,0.26),600,48) \\
    > \ \text{agreens} &\leftarrow \text{matrix}(\text{rnorm}(600*48,-1.66,0.52),600,48)
\end{align*}
\]
> breds <- matrix(rnorm(600*48,0.006,0.0038),600,48)
> bgreens <- matrix(rnorm(600*48,0.004,0.0022),600,48)
> simuldata[,refs] <- agreens+bgreens*sinh(asinh(lks)+eps[,refs])
> simuldata[,samps] <- areds+breds*sinh(asinh(mks)+eps[,samps])

For differentially expressed genes (of which we believe there are about 20), we use an additional Bernoulli random variable (at the strain level and an extra uniform random variable to say how much the new expressions differ).
Conclusions

Beyond the bootstrap: Distributional Assumptions can be useful.

- Fréchet bounds for the correlations of genes under these distributional assumptions can be found.

- The correlation coefficient may have a much narrower range than we believe.

- We can attack the Multiple Testing Problem in a Multivariate Framework.

- We will probably need a different measure of association.
References


