BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1454–1461

# Nonparametric methods for identifying differentially expressed genes in microarray data

*Olga G. Troyanskaya [1], Mitchell E. Garber [1], Patrick O. Brown [2, 3], David Botstein [1,*] and Russ B. Altman [1,*]*

*[1]Department of Genetics, Stanford University School of Medicine, [2]Department of Biochemistry, Stanford University School of Medicine and [3]Howard Hughes Medical Institute, Stanford, CA, USA*

Received on February 27, 2002; revised on April 21, 2002; accepted on May 24, 2002

## ABSTRACT

**Motivation:** Gene expression experiments provide a fast and systematic way to identify disease markers relevant to clinical care. In this study, we address the problem of robust identification of differentially expressed genes from microarray data. Differentially expressed genes, or discriminator genes, are genes with significantly different expression in two user-defined groups of microarray experiments. We compare three model-free approaches: (1) nonparametric $t$-test, (2) Wilcoxon (or Mann–Whitney) rank sum test, and (3) a heuristic method based on high Pearson correlation to a perfectly differentiating gene ('ideal discriminator method'). We systematically assess the performance of each method based on simulated and biological data under varying noise levels and $p$-value cutoffs.

**Results:** All methods exhibit very low false positive rates and identify a large fraction of the differentially expressed genes in simulated data sets with noise level similar to that of actual data. Overall, the rank sum test appears most conservative, which may be advantageous when the computationally identified genes need to be tested biologically. However, if a more inclusive list of markers is desired, a higher $p$-value cutoff or the nonparametric $t$-test may be appropriate. When applied to data from lung tumor and lymphoma data sets, the methods identify biologically relevant differentially expressed genes that allow clear separation of groups in question. Thus the methods described and evaluated here provide a convenient and robust way to identify differentially expressed genes for further biological and clinical analysis.

**Availability:** By request from the authors.

**Contact:** russ.altman@stanford.edu

## BACKGROUND

DNA microarray technology allows for the monitoring of expression levels of thousands of genes under a variety of conditions. A major question in microarray studies is how to select genes associated with specific physiological states or clinical parameters–genes whose expression in a tumor sample is related to a specific tumor subtype or patient survival. In a clinical context, such differentially expressed genes are often referred to as clinical markers. Clinical markers can form the basis for diagnostic tests, particularly if they can be assayed in reliable and inexpensive ways. Identification of clinical markers may lead to improved diagnosis and treatment guidance, early disease detection, and clinical outcomes prediction.

While routine clinical use of microarrays is still not feasible, they may provide methods for fast, accurate, and systematic identification of biomedical markers from the data generated by gene expression experiments. Clinicians can then assay the expression of one or a few such markers by immunohistochemistry or quantitative PCR (Kim, 2001). Moreover, relating specific groups of genes with specific biological correlates is a critical step toward understanding the underlying molecular mechanisms and identifying novel therapeutic targets.

The most commonly used tools for identification of differentially expressed genes include qualitative observation (usually following some form of clustering of expression patterns), heuristic rules, and model-based probabilistic analysis. The simplest heuristic is setting cutoffs for gene expression changes over a background expression level. In an early gene expression study, Iyer *et al.* (1999) sought genes whose expression changed by a factor of 2.20 or more in at least two of the experiments. DeRisi *et al.* (1997) looked for 2-fold induction of gene expression compared to baseline. Xiong *et al.* (2001) identified indicator genes based on classification errors by feature wrappers (including linear discriminant analysis, logistic regression, and support vector machines). Although this approach is not based on specific data modeling assumptions, the results are affected by assumptions behind the specific classification methods used for scoring.

*To whom correspondence should be addressed.

The probabilistic approaches applied to microarray analysis include a $t$-test based on a Bayesian estimate of variance among experiment replicates with a Gaussian model for expression measurements (Long *et al.*, 2001) and a hierarchical Bayesian modeling framework with Gaussian gene-independent models combined with a $t$-test (Baldi and Long, 2001). Newton *et al.* (2001) identified differentially expressed genes by posterior odds of change based on a hierarchical Gamma–Gamma–Bernoulli model for expression ratios. All these methods either use an arbitrarily selected cutoff or probabilistic inference based on a specific data model.

As microarray data is often noisy and not normally distributed (Hunter *et al.*, 2001), it is challenging to construct a statistical data model applicable to all microarray data sets. In this context, nonparametric methods that do not assume a specific distribution of data are attractive. Previous studies suggest that using rank-transformed data in microarray analysis is advantageous (Raychaudhuri *et al.*, 2000; Tsodikov *et al.*, 2002). Dudoit *et al.* (2002a). used a nonparametric $t$-test with family-wise error rate corrected $p$-values, while the significance analysis of microarrays (SAM) method used a statistic similar to the $t$-test and permutations of repeated measurements to estimate the false discovery rate of differentially expressed genes (Tusher *et al.*, 2001). Pan (2002) used a mixture modeling approach that estimates the distribution of $t$-statistic-type scores using normal mixture models and compared it with two parametric approaches, including a regular $t$-test. Park *et al.* (2001) scored genes based on the number of permutations of expression values required to make that gene into a perfectly discriminating marker, where all high expression values belong to one group of experiments and all low expression values belong to the other group. Significance of scores was assessed based on column permutations of the data set and comparison of the distribution of scores from permuted data to that of the original data (Park *et al.*, 2001). Other investigators used similar approaches, but looked for genes with high correlation to an idealized expression pattern that perfectly discriminates between two groups; they determined statistical significance from repeating the analysis on permuted data (Galitski *et al.*, 1999; Golub *et al.*, 1999).

Although several groups presented evaluations of classification methods (Ben-Dor *et al.*, 2000; Dudoit *et al.*, 2002b), no systematic comparative studies of model-free methods that identify differentiator genes have been published. In this paper, we compare three model-free approaches: (1) a nonparametric $t$-test, (2) a rank sum test, and (3) a heuristic method based on high Pearson correlation to a perfectly differentiating gene, which we will refer to as the ideal discriminator method. We chose these methods to compare the power of model-free probabilistic reasoning (nonparametric $t$-test and rank sum test) with heuristic-based inference. We evaluate the performance of these methods on generated expression data as well as on real biological data sets.

## METHODS

### Experimental methods

We implemented and evaluated three methods for model-free identification of differentially expressed genes in microarray analysis: a nonparametric $t$-test, a Wilcoxon rank sum test, and a heuristic idealized discriminator method. The evaluation included applications to both simulated data and real biological data. By using simulated data, we could first evaluate the methods on data sets with known differentiator genes in the context of different noise levels. The simulated data were generated to create plausible distributions of microarray expression values while not perfectly matching any particular data set. From qualitative comparisons of distribution histograms and Quantile–Quantile plots of several biological data sets (Alizadeh *et al.*, 2000; Garber *et al.*, 2001; Gasch *et al.*, 2000), we found that normally generated data with uniform noise generated from uniform distribution in the range of $\mathrm{U}(-0.01, 0.01)$ to $\mathrm{U}(-0.1, 0.1)$ approximated the true distributions reasonably well. Such an approximate fit to biological data is similar to the differences in data distributions between real microarray experiments.

To test the methods, we generated ten simulated data sets (5000 genes by 40 experiments each) at each of the six noise levels ($\mathrm{U}(-0.01, 0.01)$, $\mathrm{U}(-0.05, 0.05)$, $\mathrm{U}(-0.1, 0.1)$, $\mathrm{U}(-0.5, 0.5)$, $\mathrm{U}(-0.75, 0.75)$, $\mathrm{U}(-1.0, 1.0)$). Increasing noise levels in the data sets allowed us to test robustness of the methods on very noisy data. Each data set included twenty predictor genes (markers), whose values were generated from two different normal distributions: group 1 (20 experiments) and group 2 (20 experiments). The rest of the genes, for which all values were generated from one normal distribution per gene, were considered nonpredictors. The means of each normal distribution were generated from a random normal distribution with a mean of 0 and standard deviation of 0.25 for nonpredictors and standard deviation of 0.5 for predictors. Each of the methods was then applied to each simulated data set, and true positive rate (TPR) and false positive rate (FPR) were calculated according to the following formulae.

$$TPR = \frac{number\ of\ predictors\ identified}{total\ number\ of\ predictors}$$

$$TPR = \frac{number\ of\ nonpredictors\ identified}{total\ number\ of\ nonpredictors}$$

### Nonparametric $t$-test

The $t$-statistic is well suited to finding differentially expressed genes because it allows selection of an expression

pattern that has maximal difference in mean level of expression between the two groups and minimal variation of expression within each group. Although the $t$-test assumes normal distribution of samples within each group, we make the procedure nonparametric by estimating $p$-values from permuted data sets (random column permutations). We can calculate the $t$-statistic from the permuted data set ($t_{j_{\text{perm}}}$) for each gene $j$ and count how many times it exceeds the true t statistic for that gene ($t_{j_{\text{obs}}}$). In this study, we perform 50 000 permutations for each data set. We correct the $p$-values for multiple testing by using the Bonferroni correction.

group 1: $n_1$ samples, with average expression $\overline{X}_1$
group 2: $n_2$ samples, with average expression $\overline{X}_2$

$$t\text{-statistic: } t = \frac{(\overline{X}_1 - \overline{X}_2)}{S_{\overline{X}_1 - \overline{X}_2}} \qquad S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$P_j = \frac{count(t_{j_{\text{perm}}} > t_{j_{\text{obs}}})}{count(permutations)}$$

$$P_{j_{\text{Bonferroni}}} = \min(m \times p_j, 1),$$

where $m$ = number of genes.

### Wilcoxon rank sum test

Wilcoxon rank sum test is a nonparametric test for equality of means of two samples that are nonnormal. Since it operates on rank-transformed data, it appears to be a robust choice for microarray data, which are often nonnormal and contain outliers. RST first ranks gene expression values for each gene across all experiments, and then tests for equality of means of the two ranked samples. For data sets where both $n_1$ and $n_2$ exceed 8, normal approximation of the $p$-values can be used (Walpole and Myers, 1993). The $p$-values are corrected for multiple testing by Bonferroni correction.

group 1: $n_1$ samples  group 2: $n_2$ samples$(n_2 > n_1)$

$w_1 : \sum ranks_{sample1}$

$u_1 : w_1 - n_1 \times (n_1 + 1)/2$

$mean_{u_1} = n_1 \times n_2/2$

$\text{var}_{u_1} = n_1 \times n_2 \times (n_1 + n_2 + 1)/12$

$z = (u_1 - mean_{u_1})/\sqrt{\text{var}_{u_1}}$

$z \in N(0, 1)$ when $n_1 > 8$

### Ideal discriminator method

In this method, we define an ideal discriminator as a theoretical gene that is maximally expressed in all group 1 samples and minimally expressed in all group 2 samples. The method selects genes that have the highest Pearson correlation to the ideal discriminator and assesses significance by comparing Pearson correlation score of gene $j$ for true data to the best Pearson correlation score from randomly permuted data (50 000 column permutations).

$$p_{j_{\text{perm}}} = \frac{count(\max_i(\rho_{i_{\text{perm}}}) > \rho_{j_{\text{obs}}})}{count(permutations)}$$

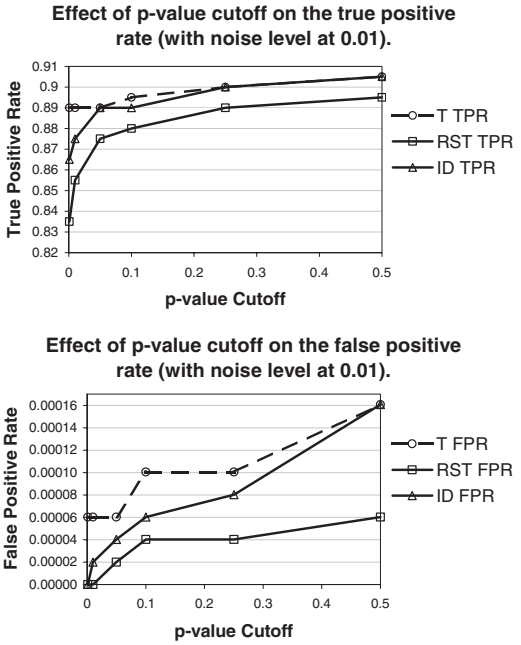



**Fig. 1.** Effect of the choice of the $p$-value cutoff on the performance of the methods. Noise level is at 0.01 (biological data sets examined showed noise level in the range 0.01–0.1). (a) Effect of the $p$-value cutoff on the true positive rate for all three tests. (b) Effect of $p$-value cutoff on false positive rate for all tests.

## RESULTS AND DISCUSSION

### Performance on simulated data

We evaluated performance of the nonparametric $t$-test (T), rank sum test (RST), and the ideal discriminator method (ID) on the simulated data sets, with controlled level of noise and with known discriminator genes, allowing us to compute the true positive rate (TPR) and false positive rate (FPR). We considered changes in TPR and FPR in response to varying amounts of noise, different $p$-value cutoffs, and decreasing sample size.

The performance of a method depends on the $p$-value cutoff below which genes are considered differentially expressed. Ideally, a method will exhibit high TPR and low FPR over a range of $p$-value cutoffs. When we examine the relationship between TPR and the $p$-value cutoff (Bonferroni) for data sets with noise level of $U(-0.01, 0.01)$, we find that for all of the methods the slope of the curves is relatively high for $p \leq 0.1$ (Figure 1a). In fact, all methods achieve TPR $\geq 0.88$

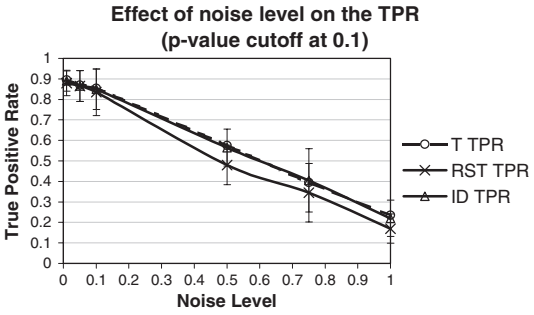



**Fig. 2.** Effect of noise level in the data set on TPR at $p$-value cutoff of 0.1. Note that biological data sets we examined showed noise level in the range 0.01–0.1. The error bars presented are one standard deviation based on 10 simulated data sets at each noise level.

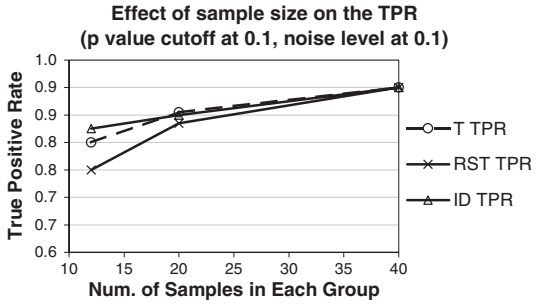



**Fig. 3.** Effect of sample size on TPR at $p$-value cutoff of 0.1 and noise level of 0.1. Note that biological data sets we examined showed noise level in the range 0.01–0.1.

for $p \geq 0.1$ at noise level of 0.01. The maximal TPR of 0.905 is achieved by the T and ID tests when a $p$-value cutoff of 0.5 is used, while the maximum TPR for RST is 0.895 at the same $p$-value cutoff. Similar results are observed for higher noise levels (for example TPR $>$ 0.83 for $p \geq 0.1$ at noise level of 0.1), although a lower maximal TPR is achieved by each test. Thus, the T and ID tests appear superior to the RST in terms of sensitivity over all $p$-value cutoffs. However, when we consider FPR, the RST exhibits the lowest FPR over all $p$-value cutoffs (Figure 1b). At $p$-value cutoff of 0.5, the RST incorrectly identifies only 2 genes as predictors, whereas T and ID mistakenly call 7 and 8 genes, respectively. All tests make little incorrect identification, and the RST appears to be most conservative.

Although the noise level definitely has a significant effect on the performance of the tests, the methods are robust to noise in the range between $U(-0.01, 0.01)$ and $U(-0.1, 0.1)$ which is most similar to real data (Figure 2). Even with extremely high noise ($U(-1.0, 1.0)$), all methods still identify some of the predictor genes (TPR $> 0.1$), and the FPR remains low for the RST and ID tests (FPR $< 4 \times 10^{-5}$). The T test shows a higher FPR. The FPR is under $8 \times 10^{-5}$ for all tests at every noise level.

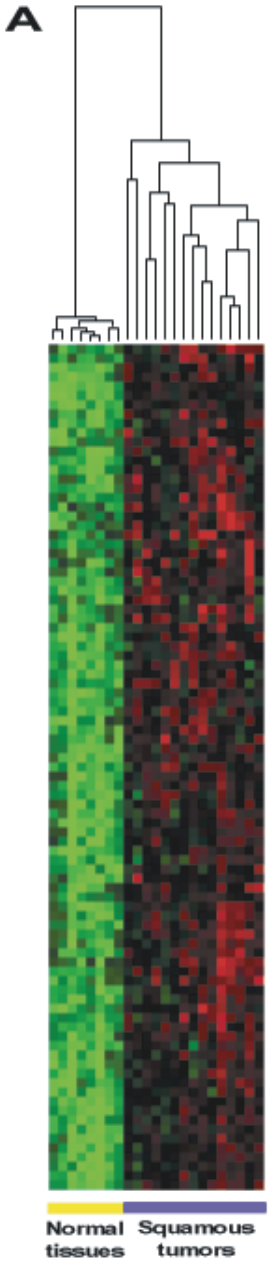
While larger sample size for both groups enhances the methods performance with respect to the TPR, the more sensitive T and ID tests correctly identify more predictors than the more conservative RST. However, the differences are not very large (a maximum of 0.7 TPR difference between RST and ID) and manifest themselves only at very small sample size (12 samples in one group; Figure 3). With noise level of 0.1 and $p$-value cutoff of 0.1, all of the methods have FPR of 0.

Overall, the RST exhibits lower true positive and false positive rates over most noise levels and $p$-value cutoffs, and is thus the more conservative or specific test (at the price of somewhat lower sensitivity). When considering the problem of marker identification, it may be beneficial to use a more conservative measure, so as to have higher confidence in genes that are then pursued biologically in a process that may require significant resources. Furthermore, the RST does not require multiple permutation iterations and thus is not computationally intensive. Thus for most data sets the rank sum test is the most appropriate first choice for marker identification. However, if a higher-sensitivity test is necessary, the nonparametric $t$-test is appropriate—or the use of a less stringent $p$-value cutoff. For both the RST and T test, a $p$-value cutoff of 0.1 appears to present the best compromise between sensitivity and false positive rate.

## Performance on biological data

A thorough comparative evaluation of marker identification methods on biological data is challenging due to the difficulty of defining a gold standard. Recently, microarray data sets with experimental validation of select markers have been reported, and the public dissemination of such data sets would provide an opportunity for reliable of marker selection methods (Gerhold *et al.*, 2001; Islam *et al.*, 2002; Mayanil *et al.*, 2001; Rajeevan *et al.*, 2001). However, we can obtain some insight into the relative utility of methods for identifying differentially expressed genes by examining their performance on simple biological data sets. For example, we can cluster data sets using only marker genes and assess their ability to separate groups of interest. In addition, we can approximately evaluate the level of method's performance by examining the biological function of differentially expressed genes selected by the method.

We applied all three methods to a data set comprised of normal lung and squamous cell lung tumor specimens. In order to concentrate on a relatively simple biological problem to assess the performance of our methods, we sought genes specific to squamous lung tumors (15 arrays) compared to normal lung (8 arrays). We used the $p$-value




B

**Proliferation related genes**

BRCA1 breast cancer 1
BUB1B budding uninhibited by benzimidazoles
C20orf1 chromosome 20 open reading frame 1
CCNE1 cyclin E1
CDC45L CDC45 homolog
CDC6 CDC6 homolog
CDKN3 cyclin-dependent kinase inhibitor 3
CENPF centromere protein F
CKAP2 cytoskeleton associated protein 2
CKS1 CDC28 protein kinase 1
CNAP1 chromosome condensation-related SMC-associated protein 1
DJ616B8.3 hypothetical protein dJ616B8.3
E2IG4 hypothetical protein, estradiol-induced
EIF4A2 **eukaryotic translation initiation factor 4A, isoform 2
EST Hs.339665 R22949
EST Hs.118338 R25481
EST Hs.118338 W85843
FEN1 flap structure-specific endonuclease 1
FLJ10540 hypothetical protein FLJ10540
FLJ14502 TRAF4 associated factor 1
FLJ22009 hypothetical protein FLJ22009
FOXM1 forkhead box M1
HELLS helicase, lymphoid-specific
Homo sapiens cDNA FLJ14365 fis
Homo sapiens, Similar to RIKEN cDNA 2610036L13 gene
KIAA0101 KIAA0101 gene product
KPNA2 karyopherin alpha 2
LOC51053 geminin
MAD2L1 MAD2-like 1
MCM2 minichromosome maintenance deficient 2
MCM4
MCM5
MCM6 minichromosome maintenance deficient 6
MSH2 mutS (E. coli) homolog 2
NUF2R hypothetical protein NUF2R
PCNA proliferating cell nuclear antigen
PLK polo (Drosophia)-like kinase
PTTG1 pituitary tumor-transforming 1
RAD51 RAD51 (S. cerevisiae) homolog
RANBP1 RAN binding protein 1
RFC4 replication factor C (activator 1) 4
RRM2 ribonucleotide reductase M2 polypeptide
SHMT2 serine hydroxymethyltransferase 2 (mitochondrial)
TMPO thymopoietin
TOP2A topoisomerase (DNA) II alpha*
TRIP13 thyroid hormone receptor interactor 13
TROAP trophinin associated protein (tastin)
UBE2C ubiquitin-conjugating enzyme E2C

**Squamous markers**

DSG2 desmoglein 2
DSP desmoplakin (DPI, DPII)*
KRT17 keratin 17*

**Genes reported overexpressed in some tumors**

ADE2H1 multifunctional polypeptide
BMP7 bone morphogenetic protein 7 (osteogenic protein 1)*
IGFBP3 insulin-like growth factor binding protein 3
MYBL2 v-myb avian myeloblastosis viral oncogene homolog-like 2
PFN2 profilin 2
PLAU plasminogen activator, urokinase
PRAME preferentially expressed antigen in melanoma
TFAP2C transcription factor AP-2 gamma
TK1 thymidine kinase 1
UMPK uridine monophosphate kinase
ZIC2 Zic family member 2

**Genes with other functions**

CCT4 chaperonin containing TCP1, subunit 4 (delta)*
CCT5 chaperonin containing TCP1, subunit 5 (epsilon)
ERO1L ERO1 (S. cerevisiae)-like
GGH gamma-glutamyl hydrolase
GMPS guanine monphosphate synthetase
GSS glutathione synthetase
LALP1 lysosomal apyrase-like protein 1
MGC4308 hypothetical protein MGC4308
MRPL3 mitochondrial ribosomal protein L3
MRPS12 mitochondrial ribosomal protein S12
NUDT1 nudix (nucleoside diphosphate linked moiety X)-type motif 1
PRKDC protein kinase, DNA-activated, catalytic polypeptide
RPA40 RNA polymerase I subunit
SLC16A1 solute carrier family 16 (monocarboxylic acid transporters), member 1
SLC2A1 solute carrier family 2 (facilitated glucose transporter), member 1
SNRPB small nuclear ribonucleoprotein polypeptides B and B1
SORD sorbitol dehydrogenase
WASF1 WAS protein family, member 1

**Fig. 4.** Markers strongly expressed in squamous lung tumors versus normal lung tissue. (a) Hierarchical clustering of squamous lung tumors and normal tissues based on 91 clones chosen by all three methods. Different lengths of branches in the hierarchical clustering tree correspond to distances between samples (longer branches mean less similarity). (b) Marker genes and their functions based on literature. Functional categories were identified based on literature references in Medline (at least 1 paper report). The symbol '∗' identifies genes for which two different clones were selected by all three methods.

cutoff of 0.1, chosen according to our simulated data experiments described above. The rank sum test identified 92 markers, the smallest number of all three methods. Of those, 91 were identified by all three methods; they corresponded to 86 distinct genes. The ID method reported 301 clones as good differentiators, whereas the T test found 202 markers. Hierarchical clustering of the data set based on the 91 differentiator clones identified by all three methods yielded clear and robust separation of normal samples from squamous tumors (Figure 4).

Of the 86 distinct genes identified as good markers for squamous tumors versus normal lung samples, 59 are known cell-cycle regulated genes (Whitfield *et al.*, 2002) or genes known to be strongly expressed in tumors. In addition, the list included desmoplakin, keratin 17, and desmoglein 2—genes known to be strongly expressed in extensively keratinized squamous lung tumors (Blobel *et al.*, 1984; Harada *et al.*, 1996; Yang *et al.*, 1995). Two other known squamous tumor markers, keratin 5 and tumor protein p63 (Blobel *et al.*, 1984; Kaufmann *et al.*, 2001), were selected by the *t*-test and ID methods, but not by the more conservative RST method, probably because both of these markers are poorly expressed in two of the 15 squamous samples studied. The RST list also included 18 genes and 6 ESTs that have no documented evidence of cell cycle regulation or role in oncogenesis. Thus, the majority of the markers for squamous tumors versus normal lung tissue identified by the RST have been experimentally identified as cellular proliferation genes or markers of squamous differentiation.

To address a more challenging biological problem, we re-examined the study of diffuse large B-cell lymphoma (DLBCL) by Alizadeh *et al.* (2000). In this work, the authors defined two molecularly distinct subtypes of DLBCL, corresponding to the originating cell type of the tumor (Alizadeh *et al.*, 2000). Germinal center B-like DLBCL expresses genes normally seen in germinal center B cells, while activated B-like DLBCL strongly expresses genes that are induced during *in vitro* activation of peripheral blood cells. We applied the RST method to the lymphoma data set[†], and identified 72 clones (70 of which are unique sequences) that are good discriminators between activated B-like DLBCL and germinal center B-like DLBCL (Figure 5). Of the genes we identified, 57 are in common with the list of over 350 clones identified by Alizadeh *et al.* (2000) from observation based on hierarchical clustering as selectively expressed in GC B-like DLBCL and activated B-like DLBCL. Among the genes identified by the RST and discussed in the Alizadeh study are known markers of germinal center differentiation, including CD10, BCL-6 and A-myb. Other genes in common with the Alizadeh study are genes that can be altered by translocations in lymphoid malignancies: BCL-7A, LMO2 (TTG-2/RBTN2), and IRF4 (Alizadeh *et al.*, 2000). Among the 13 genes not reported as differentiators in the Alizadeh *et al.* (2000) study are leukemia viral BMI-1 oncogene, MCL1 myeloid cell differentiation protein, and cyclin H (Figure 5).

When we hierarchically clustered the DLBCL samples based only on the expression profiles of the 72 clones identified by RST, the two major branches of the hierarchical tree appear identical to the GC B-like versus activated B-like lymphoma distinction presented in the Alizadeh *et al.* (2000) study (Figure 5). Thus, the smaller subset of genes identified statistically by the RST appears sufficient to distinguish the two DLBCL subclasses.

## CONCLUSIONS

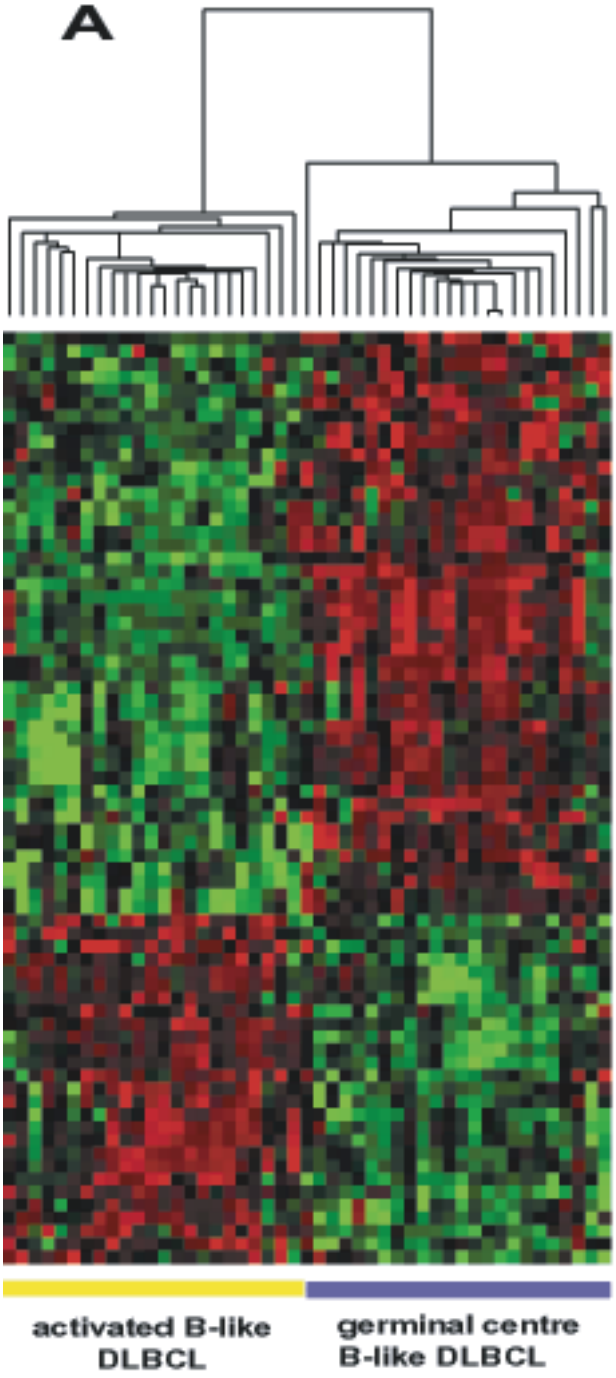
All the methods exhibit very low false positive ratse (FPR) and identify a large fraction of the discriminator genes in the simulated data sets with noise level similar to that of the real biological data. With increasing noise level, the FPR remains low, although the true positive rate (TPR) decreases. However, even in very noisy data sets some markers were correctly identified (nonzero TPR) by all methods. At noise levels close to those of real data, the *p*-value cutoff of 0.1 appears to provide the best compromise between false positive and true positive rates for all tests.

Overall, the rank sum test proved to be the most conservative method (lower FPR and TPR). In a situation where the most reliable list of markers is desirable, the best approach may be to examine the intersection of genes identified by all three methods, or by the more conservative rank sum test and *t*-test. The results of several methods may also be combined via a voting scheme with voting weights parameterized based on particular problem specifications and individual preferences. For example, a conservative voting scheme would give the greatest weight to any gene that passed all three tests but would also reward genes that passed the more conservative rank sum test with very low *p*-values more than those genes that passed only the ideal discriminator method. Our results on two real biological data sets indicate that the methods described here provide a robust way to select genes whose differential expression between groups of samples warrants further biological and clinical analysis.

## ACKNOWLEDGEMENTS

We would like to thank Max Diehn, Sandrine Dudoit, Mike Liang, Art Owen, Gavin Sherlock, Michael Whitfield, and Soumya Raychaudhari for thoughtful comments and discussions. This research was supported by grants to DB (CA77097) and POB (CA85129). O.G.T. is supported

[†] Lymphoma data set was obtained from http://llmpp.nih.gov/lymphoma/data/figure1/. DLBCL subtypes were defined as in Figure 1B of the Alizadeh *et al.* paper (Alizadeh *et al.*, 2000).




BCL-6
BCL-7A
BMI-1
CBF/PEBP2aA1/AML1
CD10
Cyclin D2
Cyclin H
Deoxycytidylate deaminase
DNA (cytosine-5-)-methyltransferase
FLICE-like inhibitory protein long form (I-FLICE)
FMR2 (Fragile X mental retardation 2)
IRF-4
JAW1 lymphoid-restricted membrane protein
JNK3 Stress-activated protein kinase
MCL1 myeloid cell differentiation protein
myb-related gene A (A-myb)
PKU-beta
Potassium voltage-gated channel shaker-related mem 3
PRK putative serine/threonine protein kinase
PTP-1B phosphotyrosyl-protein phosphatase
RPD3L1 homologue of yeast RPD3 TF
SLAP src-like adapter protein
T-cell protein-tyrosine phosphatase
TdT terminal deoxynucleotide transferase
TTG-2/Rhombotin-2
zinc finger protein 42 MZF-1

**Fig. 5.** Genes differentially expressed in germinal center B-like DLBCL versus activated B-like DLBCL based on the data from Alizadeh *et al.* (2000) study. (a) Hierarchical clustering of lymphoma samples from based on 72 clones identified by the rank sum test with 0.1 *p*-value cutoff. Different lengths of branches in the hierarchical clustering tree correspond to distances between samples (longer branches mean less similarity). (b) Named genes identified in the 72 clone set. Only named genes are listed, and genes not previously reported in Alizadeh *et al.* (2000) study as discriminators are marked in blue.

by a Howard Hughes Medical Institute Predoctoral Fellowship and a Stanford Graduate Fellowship. RBA is supported by NIH-GM61374, NIH-LM06244, NSF DBI-9600637, SUN Microsystems and a grant from the Burroughs-Wellcome Foundation.

## REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Ben-Dor,A., Bruhn,L., Friedman,N., Nachman,I., Schummer,M. and Yakhini,Z. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–583.

Blobel,G.A., Moll,R., Franke,W.W. and Vogt-Moykopf,I. (1984) Cytokeratins in normal lung and lung carcinomas. I. Adenocarcinomas, squamous cell carcinomas and cultured cell lines. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.*, **45**, 407–429.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2002a) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, in press.

Dudoit,S., Yang,Y.H., Speed,T.P. and Gallow,M.J. (2002b) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, in press.

Galitski,T., Saldanha,A.J., Styles,C.A., Lander,E.S. and Fink,G.R. (1999) Ploidy regulation of gene expression. *Science*, **285**, 251–254.

Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Genbelbach,M., Rijn,M., Rosen,G.D., Perou,C.M., Whyte,R.I. *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, **98**, 13784–13789.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gerhold,D., Lu,M., Xu,J., Austin,C., Caskey,C.T. and Rushmore,T. (2001) Monitoring expression of genes involved in drug metabolism and toxicology using DNA microarrays. *Physiol Genomics*, **5**, 161–170.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Harada,H., Iwatsuki,K., Ohtsuka,M., Han,G.W. and Kaneko,F. (1996) Abnormal desmoglein expression by squamous cell carcinoma cells. *Acta Derm Venereol*, **76**, 417–420.

Hunter,L., Taylor,R.C., Leach,S.M. and Simon,R. (2001) GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, **17** Suppl 1, S115–S122.

Islam,T.C., Lindvall,J., Wennborg,A., Branden,L.J., Rabbani,H. and Smith,C.I. (2002) Expression profiling in transformed human B cells: influence of Btk mutations and comparison to B cell lymphomas using filter and oligonucleotide arrays. *Eur. J. Immunol.*, **32**, 982–993.

Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C., Trent,J.M., Staudt,L.M., Hudson,Jr,J., Boguski,M.S. *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum.

Kaufmann,O., Fietze,E., Mengs,J. and Dietel,M. (2001) Value of p63 and cytokeratin 5/6 as immunohistochemical markers for the differential diagnosis of poorly differentiated and undifferentiated carcinomas. *Am. J. Clin. Pathol.*, **116**, 823–830.

Kim,D.W. (2001) Real time quantitative PCR. *Exp. Mol. Med.*, **33**, 101–109.

Long,A.D., Mangalam,H.J., Chan,B.Y., Tolleri,L., Hatfield,G.W. and Baldi,P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J. Biol. Chem.*, **276**, 19937–19944.

Mayanil,C.S., George,D., Freilich,L., Miljan,E.J., Mania-Farnell,B., McLone,D.G. and Bremer,E.G. (2001) Microarray analysis detects novel Pax3 downstream target genes. *J. Biol. Chem.*, **276**, 49299–49309.

Newton,M.A., Kendziorski,C.M., Richmond,C.S., Blattner,F.R. and Tsui,K.W. (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comput. Biol.*, **8**, 37–52.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Park,P.J., Pagano,M. and Bonetti,M. (2001) A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput.*, 52–63.

Rajeevan,M.S., Ranamukhaarachchi,D.G., Vernon,S.D. and Unger,E.R. (2001) Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. *Methods*, **25**, 443–451.

Raychaudhuri,S., Stuart,J.M., Liu,X., Small,P.M. and Altman,R.B. (2000) Pattern recognition of genomic features with microarrays: site typing of Mycobacterium tuberculosis strains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 286–295.

Tsodikov,A., Szabo,A. and Jones,D. (2002) Adjustments and measures of differential expression for microarray data. *Bioinformatics*, **18**, 251–260.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Walpole,R.E. and Myers,R.H. (1993) *Probability and Statistics for Engineers and Scientists*, 5th edn, Macmillan, New York.

Whitfield,M.L., Sherlock,G., Saldanha,A., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, in press.

Xiong,M., Fang,X. and Zhao,J. (2001) Biomarker identification by feature wrappers. *Genome Res.*, **11**, 1878–1887.

Yang,J., Luo,X. and Cai,L. (1995) Expression of desmoplakin II in tumors. *Zhonghua Bing Li Xue Za Zhi*, **24**, 170–172.