Hindawi Publishing Corporation
ISRN Bioinformatics
Volume 2013, Article ID 404717, 15 pages
http://dx.doi.org/10.1155/2013/404717

Research Article

Exploiting Identifiability and Intergene Correlation for Improved Detection of Differential Expression

J. R. Deller Jr.,[1] Hayder Radha,[1] and J. Justin McCormick[2]

[1] *Department of Electrical and Computer Engineering, Michigan State University, 2120 EB, East Lansing, MI 48824, USA*
[2] *Department of Molecular Biology & Biochemistry, Carcinogenesis Laboratory, Michigan State University,
341 FST, East Lansing, MI 48824, USA*

Correspondence should be addressed to J. R. Deller Jr.; deller@egr.msu.edu

Received 13 October 2012; Accepted 19 November 2012

Academic Editors: H. Ma, K. Mizuguchi, and H.-C. Yang

Accurate differential analysis of microarray data strongly depends on effective treatment of intergene correlation. Such dependence is ordinarily accounted for in terms of its effect on significance cutoffs. In this paper, it is shown that correlation can, in fact, be exploited to share information across tests and reorder expression differentials for increased statistical power, regardless of the threshold. Significantly improved differential analysis is the result of two simple measures: (i) adjusting test statistics to exploit information from identifiable genes (the large subset of genes represented on a microarray that can be classified *a priori* as nondifferential with very high confidence], but (ii) doing so in a way that accounts for linear dependencies among identifiable and nonidentifiable genes. A method is developed that builds upon the widely used two-sample *t*-statistic approach and uses analysis in Hilbert space to decompose the nonidentified gene vector into two components that are correlated and uncorrelated with the identified set. In the application to data derived from a widely studied prostate cancer database, the proposed method outperforms some of the most highly regarded approaches published to date. Algorithms in MATLAB and in R are available for public download.

1. Preamble

In certain ways, this paper represents a departure from current trends in scientific publishing. The Worldwide Web has made available extraordinary resources in the form of databases for comparative analysis of methods in bioinformatics and numerous other disciplines. The benefits of using *common* sets of *real* data to compare and contrast new algorithms are obvious. In some fields of investigation, especially, perhaps, research in early states of knowledge (e.g., genomics), there is an equally obvious drawback in using real data—that the "correct answers are not known," making it difficult to ultimately interpret differences in performance as anything but differences.

Lest the reader be preparing for an argument promoting classic simulation studies, we hasten to state at the outset that this argument is not forthcoming. Before the age of the internet, simulation studies using reasonably justified data models (Gaussian errors, etc.) were a time-honored standard in all areas of math, science, and engineering. The ready availability of rich data resources makes it irrational to advocate to a return to "pure simulation" using models that are untested against these existing data sets. The authors of this paper in no way promote a return to such methods and appeal to the reader to recognize that "models detached from reality" are not used in any way in this paper.

This research centers on some rather straightforward adjustments to classic hypothesis-testing procedures for use in differential analysis of microarray expression data. The salient result is a "reranking" of the order in which gene expression data are considered for "truly differential" status. In an effort to objectively compare these modified methods with the performance of established algorithms, it was decided to create a database of microarray expression simulations in which the data retained as much second-order statistical character as possible relative to a widely used prostate cancer database. The up- and downregulations of gene expressions were indeed synthetically modulated onto a carefully constructed baseline. The consequence is a set of comparative performance results that are objectively based on data that were "guided by nature" but which were, quite openly stated, synthesized with this natural guidance. Rather than dismissing these results as "simulations," the reader is urged to consider whether there is merit in moving the uncertainty to the data generation side (if that uncertainty can be intelligently controlled), if it permits objective results on the data analysis side. The current *modus operandi* is to accept uncertainty in the performance results with the benefit of authenticity in the data generation. The authors hope that this question might engender some debate and research.

Using the testing approach employing "nature-guided simulation," the results for the reranking method presented here are remarkably good relative to two established methods developed by respected authorities. Many people have reviewed these results, including some very eminent statisticians and bioinformaticians. Reactions have ranged from encouragement and amazement to deep skepticism. The comment "too good to be true" has been used at least twice, including, once, in a constructive criticism, by a distinguished editor of this journal. We understand this response: unfortunately (or fortunately, depending on one's perspective), no one has been able to find a flaw in the methods. We suggest two possibilities. (1) Some aspect of the simulation procedure is creating a bias toward the developed detection method. (2) Authors with a somewhat different perspective (two signal processing engineers and a cancer researcher, with over a century of combined research experience, but without extensive work in bioinformatics) were able to see some relatively simple algorithmic adjustments that eluded researchers focusing on deeper issues.

Rather than viewing the results of this paper as a claim of superiority of the new method over respected algorithms, the authors appeal to the reader to accept the report in the spirit it is offered: interesting, potentially useful, results that raise many questions, and possibilities for further research. The "intelligent stimulation" approach in itself may offer some grounds for innovation in the field. Attempts to verify that the present results are, indeed, "too good to be true" may reveal technical information benefiting differential expression detection methods. This is the classical way in which research moves forward. We are grateful to the *ISRN Journal of Bioinformatics* for the opportunity to bring these ideas to the attention of the research community.

2. Introduction

The DNA microarray was initially touted as a tool that would revolutionize the understanding of complex diseases and usher in an era of personalized medicine. This optimism is on display in Lander's 1999 *Nature Genetics* article entitled "Array of hope" [1]. It is not unusual, however, for near-term impacts of emerging technologies to be overestimated when first deployed, then to have the expectations moderated as the technologies reveal new complexities in the problems they are designed to solve. Over the past decade, early optimism about the microarray has given way to a pragmatic understanding of challenges and the need for further research and development. This normal course of events led to Frantz's 2005 article in *National Review of Drug Discovery* entitled "An array of problems" [2]. The study of microarray data has shown the need for exceeding care in the design and regularization of experiments and in the data collection and preprocessing, but the biggest hindrance to progress has been the lack of definitive methods for *interpretation* of microarray results.

One of the main challenges to proper analysis is the presence of significant correlation among gene expressions manifest in the microarray results [3, 4]. One measurable indication of the uncertainty caused by correlated differential-expression tests is the resultant increase in the variance of the *false discovery rate* (FDR) [3]. Linear statistical dependence among gene-expression correlations, therefore, can be quantifiably linked to higher-risk detection algorithms for the discovery of active genes. Among many causes, intergene correlation is attributable to coexpression of genes [5] and to unmodeled factors that introduce systematic effects across genes [6, 7]. As a result, for most real data, the assumption of independence or weak dependence among gene expressions is unfounded, and methods treating correlation are necessary [8, 9]. In fact, a few investigators have even questioned the adequacy of accounting for correlation alone and have examined the implications of nonlinear dependence on the discovery of genes [10–12].

Correctly detecting differentially expressed genes—or the related task of estimating the FDR—in the presence of substantial intergene correlation is a challenging problem that has received much recent attention since reported in papers by Owen, Efron, and others (e.g., [3, 4, 9, 13]). For example, Storey et al. [14] present an approach to the notion of sharing information across t scores, which they describe as "borrowing strength across the tests" for a potential increase in statistical power. Tibshirani and Wasserman [15] discuss a quantity called the "correlation-shared" t-statistic and derive theoretical bounds on its performance. Hu et al. [16] examine the covariance structure of the expression data and discover benefits of linking coexpression and differential expression in a distance measure—reflecting the more recent interest in characterizing broader statistical patterns in microarray data.

Recent research that is jointly concerned with differential expression and coexpression has also yielded results and methods that could ultimately benefit the gene discovery problem. Because the differential coexpression research is often concerned with differing phenotypes, rather than with different treatment conditions, two given research efforts involving differential coexpression might seek answers to different sets of genetic questions through expression data. Like the "treatment conditions researchers," however, the "phenotype" researchers have encountered their own forms of confounding dependencies, notably the relative gene locations, the expression time sequencing, and phase information (e.g., [17–19]). Papers have been published addressing these issues, including the exposition of new statistical approaches—for example, "CorScor" [20], the "ECF-statistic" [21], gene-set coexpression analysis [22], fuzzy expression level assignments [23], expression-profile mining with decorrelation [24], and detection of microarray outliers [25]—as well as new clustering methods—for example, a web-based expression analyzer [26], high-order preclustering methods [27], and the "BioSym" distance measure [28]. A recent review of clustering methods in genomics appears in the paper by Dalton et al. [29]. A more general examination of the performance of classifiers of microarray expressions appears in the paper by Ancona et al. [30].

This paper is focused exclusively on the differential expression problem. Research in this area has largely focused on understanding harmful correlation effects on the choice of the threshold demarcating the statistical boundary between differential and nondifferential expression. In fact, however, the nominally confounding correlation can be used to advantage in increasing statistical power of microarray studies. This paper presents a differential analysis method that exploits identifiability and uses a gene expression reranking criterion that accounts for intergene correlation. The framework is readily generalizable for use in studies with multiple or continuous covariates, as well as to other multiple comparison applications. An example method presented here builds upon the widely used two-sample t-statistic approach with a decomposition of the expression vectors into subspaces of correlated and uncorrelated components.

3. Problem Formulation

Suppose that expression data for G genes are measured on M microarrays, resulting in a gene expression matrix, say $\mathbf{X} \in \mathbb{R}^{G\times M}$, with (g, m) element x_{gm}. Each of the M microarray experiments takes place under one of two conditions (indexed by $k = 1 \quad \text{or} \quad 2$) such as control and treatment. These two subsets of the data are called *treatment groups* in the paper.

Based on the gene expression matrix, $\mathbf{X}$, we seek to identify a “small” number, $G_* \ll G$, of genes that are significantly differentially expressed between the two groups. One widely used strategy (e.g., [31, 32]) is to hypothesize that each gene is *not* differentially expressed. We refer to this as the *null hypothesis*, denoted $\mathbb{H}_0$. For convenience, we use the shorthand notation $g \in \mathbb{H}_0$ to indicate that the null hypothesis is known to be true for gene g, and, conversely, $g \notin \mathbb{H}_0$ indicates that gene g does not satisfy $\mathbb{H}_0$. Gene g is tested against $\mathbb{H}_0$ using a two-sample t-statistic, say, t_g. The magnitudes of the statistics $t_1, t_2, \ldots, t_G$, establish a gene ranking and the G_* genes with the largest t-scores are reported as statistically significant discoveries. The investigator can either supply a value for G_* or rely on an estimation of the number of false discoveries (type I errors, false positives), say $\mathfrak{F}$, or, equivalently, the FDR, defined as $\overline{\mathfrak{F}} \overset{\text{def}}{=} \mathfrak{F}/G_*$, to find a maximal G_* with the allowable $\mathfrak{F}$ or $\overline{\mathfrak{F}}$ (e.g., [3, 5, 9, 13, 33–36]).

As discussed in [4, 11], for an “overpowered” $\mathbf{X}$ matrix, there may be significantly fewer tail-area null counts than expected, whereas for an “underpowered” $\mathbf{X}$, the situation can worsen with an excessive number of tail-area null counts. It is important to note that techniques for estimating the FDR change the numbers of genes in reported lists, not the gene *rankings*. The present research was motivated by the hypothesis that, for an “underpowered” $\mathbf{X}$, it would be possible to exploit correlation across t scores to establish a gene ranking with more statistical power than the raw t-based ranking. The method that resulted from an exploration of this question indeed seems to improve the statistical power of *all* $\mathbf{X}$ matrices.

The new method uses a vector of t-statistics,

$$\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \cdots & t_G \end{bmatrix}^T, \tag{1}$$

and an estimate of the covariance matrix of the vector $\mathbf{t}$, to output a substantially revised version of $\mathbf{t}$, denoted $\boldsymbol{\tau}$, the entries of which provide an improved gene ranking. For expediency, we will refer to the procedure that produces $\boldsymbol{\tau}$ from $\mathbf{t}$ as *correlation adjusted reranking*, or simply *reranking*. The essence of reranking is embodied in some fundamental data conditioning procedures to effect the two outcomes mentioned in the introduction—exploiting identifiability, and nullifying the effects of intergene correlation between identified and non-identified genes.

In the following sections, we develop the theoretical basis for the reranking process. The performance of reranking is then compared with that of state-of-the-art methods on data derived from real expression experiments. Results of some judiciously developed simulation studies are also reported for their value in understanding certain aspects of the performance.

4. Methods

4.1. Per Gene Summary Statistic . Let us first supply the details surrounding the vector of t statistics introduced above. $\mathbf{t}$ will be viewed as a random vector with mean vector $\boldsymbol{\mu} \in \mathbb{R}^G$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{G\times G}$. Henceforth, we write $\mathbf{t} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This is a theoretical model only, as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are generally unknown. No further distribution information is required.

Let $\overline{x}_g$ denote the average differential expression level for gene g, $\overline{x}_g = M^{-1} \sum_{m=1}^{M} x_{gm}$. Further, let $\overline{x}_{g|k}$ be the average expression level for gene g in treatment condition k. Then, the (unpaired) t-statistic for gene g is computed as

$$t_g = \frac{\overline{x}_{g|2} - \overline{x}_{g|1}}{s_g}, \tag{2}$$

where s_g is the pooled within-group sample standard deviation of gene g. If $g \in \mathbb{H}_0$, then we expect $t_g \sim (0, \nu[\nu - 2]^{-1})$, where ν is the number of degrees of freedom, obtained from either the unpaired t-test theory or the permutation null calculations [4]. Otherwise, we expect $t_g \sim (\mu_g, \sigma_g^2)$ with μ_g and σ_g^2 denoting the gth element of $\boldsymbol{\mu}$ and the gth diagonal element of $\boldsymbol{\Sigma}$, respectively. For $g \notin \mathbb{H}_0$, μ_g and σ_g^2 depend on the amount of up- or downregulation of the gene expression, the number of samples in each treatment group, and the number of degrees of freedom, ν.

4.2. Invoking Identifiability: The Zero Assumption. The direct use of t-scores for ranking neglects some important information that is inherent in the microarray matrix $\mathbf{X}$. Fundamentally, the use of raw t-scores does not exploit *identifiability*, the strongly justified assumption that certain genes almost surely satisfy $\mathbb{H}_0$. Formal backing for the creation of an identifiable set is found in Efron's *zero assumption* (ZA) [5], which states that a fraction, say p_0, of the genes—those with the smallest t_g statistics—satisfies $\mathbb{H}_0$. The ZA plays a central role in the literature on estimating the proportion of null genes, as in [13, 37]. The ZA is equally crucial for the two-group model approach developed in the Bayesian microarray literature, as in [38–40]. Furthermore, the assertion that the method developed by Storey [35] improves upon the well-known Benjamini-Hochberg FDR procedure [33] (in terms of statistical power) crucially relies on an adaptive version of the ZA.

The use of the ZA is justified in the reranking procedure as long as the parameter p_0 is sufficiently small. For example, we set $p_0 \approx 0.5$ in experiments below based on the almost certain knowledge that ~50% of the genes in a cell would not be differentially expressed in the formation of prostate cancer [5]. Accordingly, in the initial step of the reranking process, we invoke the ZA to partition the G genes into an *identified set* of G_0 elements assumed to satisfy $\mathbb{H}_0$ and a *candidate set* of $G_1 \overset{\text{def}}{=} (G-G_0)$ genes, so-named because they remain "candidates" to become "discovered" genes (i.e., to be among the $G_* \le G_1 < G$ genes declared to be differentially expressed). For convenience and without any significant loss of generality (especially for large G), we will assume that if the fraction, p_0, rather than the cardinality, G_0, is used to specify the size of the identified set, then p_0 is selected so that p_0G is an integer. That integer is therefore G_0, and $p_0 = G_0/G$.

A simple corollary to the ZA is that the t_g value for each $g \in \mathbb{H}_0$ represents a "noise" value in the t-score for that gene. This is because the expected value of this statistic is $\mathscr{E}\{t_g\} = 0$ whenever $g \in \mathbb{H}_0$. Accordingly, we can view the t_g value for $g \in \mathbb{H}_0$ as a random variation around the nominal value of zero differential expression. The reranking procedure exploits this information to adjust the values of the candidate gene statistics. This "adjustment" is a consequence of decorrelating candidate expression values from those in the identified set. That these two subsets would be correlated may, at first, seem counterintuitive because the "interesting" differentially expressed genes must, by definition, come from the candidate set. Would it not be the case, therefore, that the significant intergene correlation would occur among candidate expression values that similarly respond to the change in treatment condition? Whereas two coexpressed genes would likely have correlated expression differentials, it is not this dependence that potentially causes false discoveries. To the extent that the correlation between genes $g, g' \notin \mathbb{H}_0$ is a reflection of response to treatment conditions, the correlation is expected and informative. Correlation between truly expressed and unexpressed genes (possibly identified) reflects normal variations in expression unrelated to treatment condition. Correcting for such correlation (whether the effect is to increase or decrease the expression level) is important to the proper assessment of candidate t-scores. Moreover, the existence of the identified set, with its "noise only" interpretation of t-scores, makes possible this correction; ultimately resulting is reranked expression statistics.

To proceed, we must add some formality to the descriptions of the identified and candidate sets. Without loss of generality, we may assume that the complete set of genes is indexed so that

$$|t_1| \le |t_2| \le \cdots \le |t_{G_0}| \le |t_{G_0+1}| \le \cdots \le |t_G|, \qquad (3)$$

in which G_0 is the number of genes declared null under the ZA. Then, genes with indices $1, 2, \ldots, G_0$ are assumed null and therefore comprise the identified set as defined above. Let us partition the set of t_g statistics into those corresponding to genes declared null under the ZA, $\{t_1, t_2, \ldots, t_{G_0}\}$ and those for the remaining $G_1 \overset{\text{def}}{=} (G - G_0)$ genes that continue to compete for the nonnull designation (the candidate set), $\{t_{G_0+1}, t_{G_0+2}, \ldots, t_G\}$. For convenience, express the vector $\mathbf{t}$ in terms of these two partitions:

$$\mathbf{t} = \underbrace{[\, t_1 \;\; t_2 \;\; \cdots \;\; t_{G_0}}_{(\mathbf{t}^0)^T} \; \underbrace{t_{G_0+1} \;\; t_{G_0+2} \;\; \cdots \;\; t_G \,]^T}_{(\mathbf{t}^1)^T} = \begin{bmatrix} \mathbf{t}^0 \\ \mathbf{t}^1 \end{bmatrix}. \qquad (4)$$

The random vector $\mathbf{t}$ has the following moments:

$$\mathbf{t} = \begin{bmatrix} \mathbf{t}^0 \\ \mathbf{t}^1 \end{bmatrix} \sim \left(\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^0 \\ \boldsymbol{\mu}^1 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{00} & \boldsymbol{\Sigma}^{01} \\ \boldsymbol{\Sigma}^{10} & \boldsymbol{\Sigma}^{11} \end{bmatrix} \right), \qquad (5)$$

in which $\boldsymbol{\mu}^i \overset{\text{def}}{=} \mathscr{E}\{\mathbf{t}^i\}$ and $\boldsymbol{\Sigma}^{ij} \overset{\text{def}}{=} \mathscr{E}\{(\mathbf{t}^i - \boldsymbol{\mu}^i)(\mathbf{t}^j - \boldsymbol{\mu}^j)^T\}$, for $i, j = 0, 1$. Recall that the goal of the reranking process is to find a vector $\boldsymbol{\tau}$ the elements of which represent a reordering of the elements of $\mathbf{t}$, such that gene ranking represented by $\boldsymbol{\tau}$ has better statistical power for detecting nonnull genes than that based on $\mathbf{t}$ itself. In the light of the newly defined notation in (5), we can more specifically say that $\boldsymbol{\tau}$ represents a reevaluation and reordering of the G_1 elements in $\mathbf{t}^1$. The remaining elements of the revised vector, comprising the vector partition, say "$\tilde{\mathbf{t}}^0$," are effectively set to zero since these genes are assumed to represent null genes. Recall that the t-scores in the $\mathbf{t}^0$ vector (before processing) are assumed to represent noise variations around the nominal zero differential value for a null gene.

4.3. Theoretical Estimator of $\boldsymbol{\tau}$. Conditioned upon the random vector $\mathbf{t}$, we seek a revised vector, $\tilde{\mathbf{t}}^1(\mathbf{t}) \equiv \boldsymbol{\tau}$, in which the expression statistics for the candidate gene set (originally $\mathbf{t}^1$) are uncorrelated with those of the identified set (originally $\mathbf{t}^0$). There are many ways to derive the desired result, each with its own interpretation, but all, of course, ultimately equivalent. Whatever the approach is, it is expedient to remove the mean from the vector $\mathbf{t}^1$ and work with the centered vector $\mathbf{t}^1_c \overset{\text{def}}{=} \mathbf{t}^1 - \boldsymbol{\mu}^1$. The centered counterpart to the reranking vector, $\boldsymbol{\tau}$, will be denoted $\tilde{\mathbf{t}}^1_c \overset{\text{def}}{=} \boldsymbol{\tau} - \boldsymbol{\mu}^1$. The constant vector $\boldsymbol{\mu}^1$ will be returned to the result at the end. Recall that the mean of the vector $\mathbf{t}^0$ is $\boldsymbol{\mu}^0 = \mathbf{0}_{G_0\times 1}$, so that "centering" is unnecessary for $\mathbf{t}^0$ (throughout the paper, the notation $\mathbf{0}_{I\times J}$ denotes the zero matrix (vector) in the Cartesian space $\mathbb{R}^{I\times J}$.)

To derive the desired expression for $\tilde{\mathbf{t}}^1$, we adopt a simple approach based on well-known ideas from the theory of linear operators (e.g., [41–44]). Let us view the space of random vectors in (we are assuming $G_1 \ge G_0$, or $p_0 \le 0.5$.

If the converse is true, simply reverse the roles of the $\mathbb{R}^{G_1}$ and $\mathbb{R}^{G_0}$ spaces in this development.) $\mathbb{R}^{G_1}$ as a Hilbert space, $\mathscr{H}$, with inner product $\langle \mathbf{v}, \mathbf{w} \rangle \overset{\text{def}}{=} \mathscr{E}\{\mathbf{v}^T\mathbf{w}\}$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{G_1}$. The inner product induces a norm on $\mathscr{H}$ given by $\|\mathbf{v}\|_2 \overset{\text{def}}{=} \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\mathscr{E}\{\mathbf{v}^T\mathbf{v}\}}$, and, in turn, a metric

$$\begin{aligned} d(\mathbf{v}, \mathbf{w}) &\overset{\text{def}}{=} \|\mathbf{v} - \mathbf{w}\|_2 = \sqrt{\langle \mathbf{v} - \mathbf{w}, \mathbf{v} - \mathbf{w} \rangle} \\ &= \sqrt{\mathscr{E}\left\{(\mathbf{v} - \mathbf{w})^T (\mathbf{v} - \mathbf{w})\right\}}. \end{aligned} \tag{6}$$

Now let $\mathscr{S}$ be a closed subset of $\mathbb{R}^{G_1}$. Given some $\mathbf{v} \in \mathbb{R}^{G_1}$, we wish to find closest element of $\mathscr{S}$, say $\hat{\mathbf{v}}$, to $\mathbf{v}$, in the sense that

$$\hat{\mathbf{v}} = \underset{\mathbf{w} \in \mathscr{S}}{\operatorname{argmin}}\, d(\mathbf{v}, \mathbf{w}), \tag{7}$$

in which d is the metric in (6). The *Hilbert projection theorem* states the solution $\hat{\mathbf{v}}$ exists and is unique. Moreover, $\hat{\mathbf{v}}$ has the property that the vector difference between $\mathbf{v}$ and $\hat{\mathbf{v}}$ is orthogonal to all vectors in the subspace $\mathscr{S}$; that is, $(\mathbf{v} - \hat{\mathbf{v}}) \in \mathscr{S}^\perp$, in which $\mathscr{S}^\perp$ is the orthogonal complement to subspace $\mathscr{S}$ in $\mathbb{R}^{G_1}$. In particular, random vector $(\mathbf{v} - \hat{\mathbf{v}})$ is orthogonal to $\mathbf{v} \in \mathscr{S}^\perp$ which means that $\langle (\mathbf{v} - \hat{\mathbf{v}}), \mathbf{v} \rangle = 0$ by definition. Thus, $\mathscr{E}\{(\mathbf{v} - \hat{\mathbf{v}})^T\mathbf{v}\} = 0$, implying that $(\mathbf{v} - \hat{\mathbf{v}})$ and $\mathbf{v}$ are stochastically orthogonal (uncorrelated if $\boldsymbol{\mu}_{\mathbf{v}} = \mathbf{0}_{G_1 \times 1}$).

The Hilbert-space formulation provides the structure in which to achieve the desired decomposition of the random vector $\mathbf{t}_c^1$ into components that are correlated, and uncorrelated, with $\mathbf{t}^0$. Vector $\mathbf{t}_c^1$ resides in the space $\mathbb{R}^{G_1}$. We seek a vector in a G_0-dimensional subspace of $\mathbb{R}^{G_1}$ (because G_0 is the dimension of $\mathbf{t}^0$) that is close to $\mathbf{t}_c^1$ (this will be the component of $\mathbf{t}_c^1$ that is correlated with $\mathbf{t}^0$). The problem at hand is only subtly different from one described in the generalities above. The difference is that we are not given a subspace "$\mathscr{S}$" in which to find an optimal vector; rather we are given only a *vector* $\mathbf{t}^0$ which may reside in an uncountable number of subspaces of dimension G_0. The goal will be to perform a linear operation on the given vector to "make it close" to $\mathbf{t}_c^1$. In the process, we will inherently construct the subspace in which the optimal result resides. The subspace and resulting vector will be of dimension G_0 because they are merely different representations of $\mathbf{t}^0$ and its implicit initial subspace $\mathbb{R}^{G_0}$.

Given $\mathbf{t}^0$, let us conceptualize a G_0-dimensional subspace of $\mathbb{R}^{G_1}$, $\mathscr{S}$, consisting of all the (random) vectors $\{\mathbf{v} : \mathbf{v} = \mathbf{F}\mathbf{t}^0,\ \mathbf{F} : \mathbb{R}^{G_0} \mapsto \mathbb{R}^{G_1}\}$. $\mathbf{F}$ represents some (yet unknown) linear operator of dimensions $G_1 \times G_0$ and of rank G_0. According to the Hilbert projection theorem, for a fixed $\mathbf{t}^0$, there is a unique vector $\hat{\mathbf{t}}_c^1 = \hat{\mathbf{F}}\mathbf{t}^0 \in \mathscr{S}$, hence a unique linear operator, $\hat{\mathbf{F}}$, such that

$$\begin{aligned} d\left(\mathbf{t}_c^1, \hat{\mathbf{F}}\mathbf{t}^0\right) &= \underset{\substack{\mathbf{F} \in \mathbb{R}^{G_1 \times G_0} \\ \text{rank}(\mathbf{F}) = G_0}}{\operatorname{argmin}}\, d\left(\mathbf{t}_c^1, \mathbf{F}\mathbf{t}^0\right) = \underset{\substack{\mathbf{F} \in \mathbb{R}^{G_1 \times G_0} \\ \text{rank}(\mathbf{F}) = G_0}}{\operatorname{argmin}} \left\|\mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0\right\| \\ &= \underset{\substack{\mathbf{F} \in \mathbb{R}^{G_1 \times G_0} \\ \text{rank}(\mathbf{F}) = G_0}}{\operatorname{argmin}} \sqrt{\left\langle \mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0, \mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0 \right\rangle} \\ &= \underset{\substack{\mathbf{F} \in \mathbb{R}^{G_1 \times G_0} \\ \text{rank}(\mathbf{F}) = G_0}}{\operatorname{argmin}} \sqrt{\mathscr{E}\left\{(\mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0)^T (\mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0)\right\}}. \end{aligned} \tag{8}$$

Since the metric is nonnegative, we may equivalently seek $\hat{\mathbf{F}}$ that minimizes

$$\begin{aligned} d^2\left(\mathbf{t}_c^1, \mathbf{F}\mathbf{t}^0\right) &= \mathscr{E}\left\{\left(\mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0\right)^T \left(\mathbf{t}_c^1 - \mathbf{F}\mathbf{t}^0\right)\right\} = \mathscr{E}\left\{\left(\mathbf{t}_c^1\right)^T \mathbf{t}_c^1\right\} \\ &\quad - 2\mathscr{E}\left\{\left(\mathbf{t}_c^1\right)^T \mathbf{F}\mathbf{t}^0\right\} + \mathscr{E}\left\{\left(\mathbf{t}^0\right)^T \mathbf{F}^T\mathbf{F}\mathbf{t}^0\right\}. \end{aligned} \tag{9}$$

Now, interpreting the expectation to be a mean-square average, we compute the gradient of $d^2(\mathbf{t}_c^1, \mathbf{F}\mathbf{t}^0)$ with respect to $\mathbf{F}$ (e.g., see [45]),

$$\nabla_{\mathbf{F}} d^2\left(\mathbf{t}_c^1, \mathbf{F}\mathbf{t}^0\right) = -2\mathscr{E}\left\{\mathbf{t}_c^1\left(\mathbf{t}^0\right)^T\right\} + 2\mathbf{F}\mathscr{E}\left\{\mathbf{t}^0\left(\mathbf{t}^0\right)^T\right\}, \tag{10}$$

and set the result to $\mathbf{0}_{G_1 \times G_0}$, the solution to which is $\hat{\mathbf{F}}$:

$$-\mathscr{E}\left\{\mathbf{t}_c^1\left(\mathbf{t}^0\right)^T\right\} + \hat{\mathbf{F}}\mathscr{E}\left\{\mathbf{t}^0\left(\mathbf{t}^0\right)^T\right\} \equiv \mathbf{0}_{G_1 \times G_0}. \tag{11}$$

Recognizing that (recall (5)) $\mathscr{E}\{\mathbf{t}_c^1(\mathbf{t}^0)^T\} = \boldsymbol{\Sigma}^{10}$ and $\mathscr{E}\{\mathbf{t}^0(\mathbf{t}^0)^T\} = \boldsymbol{\Sigma}^{00}$, the solution becomes

$$\hat{\mathbf{F}} = \boldsymbol{\Sigma}^{10}\left(\boldsymbol{\Sigma}^{00}\right)^{-1}. \tag{12}$$

Now the random vector $\hat{\mathbf{t}}_c^1 \overset{\text{def}}{=} \hat{\mathbf{F}}\mathbf{t}^0$ is as strongly linearly related to (correlated with) $\mathbf{t}_c^1$ as is possible in the G_0-dimensional subspace of $\mathbb{R}^{G_1}$ that is spanned by the G_0 columns of operator $\hat{\mathbf{F}}$. The correlation matrix relating $\hat{\mathbf{t}}_c^1$ and $\mathbf{t}_c^1$ is, say,

$$\begin{aligned} \boldsymbol{\Sigma}^{\hat{1}1} &= \mathscr{E}\left\{\hat{\mathbf{t}}_c^1\left(\mathbf{t}_c^1\right)^T\right\} = \mathscr{E}\left\{\hat{\mathbf{F}}\mathbf{t}^0\left(\mathbf{t}_c^1\right)^T\right\} \\ &= \hat{\mathbf{F}}\boldsymbol{\Sigma}^{01} = \boldsymbol{\Sigma}^{10}\left(\boldsymbol{\Sigma}^{00}\right)^{-1}\boldsymbol{\Sigma}^{01}. \end{aligned} \tag{13}$$

Although $\boldsymbol{\Sigma}^{\hat{1}1}$ is of dimension $G_1 \times G_1$, it is clearly singular (rank G_0) reflecting the inability of $\hat{\mathbf{t}}_c^1$ to be linearly related to any component of $\mathbf{t}_c^1$ in the subspace $\mathscr{S}^\perp$. In fact $\hat{\mathbf{t}}_c^1$ is the component of $\mathbf{t}_1$ that embodies the potentially destructive correlation of the identified genes with the candidate genes. The decorrelated version of $\mathbf{t}_c^1$ that we seek is therefore

$$\tilde{\mathbf{t}}_c^1(\mathbf{t}) = \mathbf{t}_c^1 - \hat{\mathbf{t}}_c^1 = \mathbf{t}_c^1 - \hat{\mathbf{F}}\mathbf{t}^0 = \mathbf{t}_c^1 - \boldsymbol{\Sigma}^{10}\left(\boldsymbol{\Sigma}^{00}\right)^{-1}\mathbf{t}^0. \tag{14}$$

Note that $\tilde{\mathbf{t}}_c^1$ is precisely the difference vector $\mathbf{t}_c^1 - \hat{\mathbf{F}}\mathbf{t}^0$ that is guaranteed by the Hilbert space theory to reside in $\mathscr{S}^\perp$ and therefore to be orthogonal, hence uncorrelated, with $\mathbf{t}^0$. Finally, reinserting the mean value, $\boldsymbol{\mu}^1$, of the candidate vector, we have

$$\boldsymbol{\tau} = \mathbf{t}^1 - \boldsymbol{\Sigma}^{10}\left(\boldsymbol{\Sigma}^{00}\right)^{-1}\mathbf{t}^0. \tag{15}$$

4.4. Sample Estimator of $\tilde{\mathbf{t}}^1$. Estimates of the elements of $\boldsymbol{\Sigma}$ are required. For this purpose, we make several observations that are easily verified through simulation. Note that (15) does not require the covariance between t_g and $t_{g'}$ when $g, g' \notin \mathbb{H}_0$.

Let $\gamma(u, \mathrm{v})$ and $\rho(u, \mathrm{v})$ denote the scalar covariance and correlation, respectively, between real, scalar random variables, u and v:

$$\gamma(u, \mathrm{v}) \overset{\mathrm{def}}{=} \mathscr{E}\{[u - \mathscr{E}(u)][\mathrm{v} - \mathscr{E}(\mathrm{v})]\}$$
$$\rho(u, \mathrm{v}) \overset{\mathrm{def}}{=} \mathscr{E}\{u\mathrm{v}\} = \gamma(u, \mathrm{v}) + \mathscr{E}(u)\mathscr{E}(\mathrm{v}). \tag{16}$$

In these terms, we make the following observations.

Observation 1. If $g, g' \in \mathbb{H}_0$, then [3, 4]

$$\gamma\left(t_g, t_{g'}\right) \approx \frac{\nu}{\nu - 2}\, \rho\left(x_g, x_{g'}\right). \tag{17}$$

Observation 2. If $g \in \mathbb{H}_0$ and $g' \notin \mathbb{H}_0$ (or conversely), then

$$\gamma\left(t_g, t_{g'}\right) \approx \frac{\nu}{\nu - 2} \times \frac{M_2 \rho\left(x_{g|1}, x_{g'|1}\right) + M_1 \rho\left(x_{g|2}, x_{g'|2}\right)}{M_1 + M_2} \tag{18}$$

Equation (18) accommodates the possibility that the correlation between a null and a nonnull gene may change between treatment groups. If this does not occur, then (18) reduces to (17).

Observation 3. Furthermore, if $M_1 \approx M_2$ (true for most microarray studies), then (18) simplifies to

$$\gamma\left(t_g, t_{g'}\right) \approx \frac{\nu}{\nu - 2} \frac{\rho\left(x_{g|1}, x_{g'|1}\right) + \rho\left(x_{g|2}, x_{g'|2}\right)}{2}. \tag{19}$$

Equations (17) and (19) suggest that we may use the sample covariance to estimate $\gamma(t_g, t_{g'})$:

$$\widehat{\gamma}\left(t_g, t_{g'}\right) \propto \frac{\sum_m \check{x}_{gm} \check{x}_{g'm}}{\sqrt{\left(\sum_m \check{x}_{gm}^2\right)\left(\sum_m \check{x}_{g'm}^2\right)}}, \tag{20}$$

where $\check{x}_{gm}$ denotes the expression level of the gth gene measured on the mth microarray after subtracting the average response within the treatment group to which x_{gm} belongs ($k = 1$ or 2). The scale factor $\nu/(\nu - 2)$ cancels when the terms $(\boldsymbol{\Sigma}^{00})^{-1}$ and $\boldsymbol{\Sigma}^{01}$ are multiplied in (15), so that estimating ν is not required.

In the light of (20), (15) takes the practical form

$$\tilde{\mathbf{t}}^1 = \mathbf{t}^1 - \widehat{\mathbf{R}}^{10}\left(\widehat{\mathbf{R}}^{00}\right)^{-1}\mathbf{t}^0 \overset{\mathrm{def}}{=} \mathbf{t}^1 - \widehat{\mathbf{R}}^{10}\,\widehat{\mathbf{P}}\mathbf{t}^0, \tag{21}$$

where $\widehat{\mathbf{R}}^{00}$ and $\widehat{\mathbf{R}}^{10}$ are the partitions (similarly to (5)) of $\widehat{\mathbf{R}}$, the sample correlation matrix of the gene expression matrix $\check{\mathbf{X}}$ (after removing the treatment effects). For notation compactness, we have defined $\widehat{\mathbf{P}} \overset{\mathrm{def}}{=} (\widehat{\mathbf{R}}^{00})^{-1}$. In principle, the set of elements in the vector $\tilde{\mathbf{t}}^1$ of (15) embodies the gene-expression reranking in light of the compensatory measures taken to incorporate identifiability and to remove the effects of correlation. In practice, we rely on the *estimate* of $\tilde{\mathbf{t}}^1$ in (21).

5. Implementation

5.1. Gene Reranking Algorithm. A stepwise procedure for the gene reranking is given in Algorithm 1. The process begins by reindexing the genes based on their two-sample t-statistics (Equation (3)). Then, based on the ZA, the first G_0 genes are declared to satisfy $\mathbb{H}_0$. In the experiments reported below, p_0, the fraction of genes declared to be identifiable is set to $\approx$ 0.5 by default (the precise value is $p_0 = 6312/12625$ for the database used which has $G = 12625$ total genes. See the second paragraph of Section 4.2 for an explanation). Although the choice 0.5 is somewhat arbitrary, this fraction is clearly justifiable and it has worked well empirically in the data sets tested.

In order to nullify any genuine treatment differences, $\mathbf{X}$ is converted to $\check{\mathbf{X}}$ by subtracting each gene's average response within each treatment group. The sample correlation matrix $\widehat{\mathbf{R}}$ of $\check{\mathbf{X}}$ is subsequently computed. The critical step is to compute $\tilde{\mathbf{t}}^1$ (Equation (21)). The elements of $\tilde{\mathbf{t}}^1$ determine the gene ranking: gene g is ranked more highly than gene g' if $|\tilde{t}_g^1| > |\tilde{t}_{g'}^1|$ in which $\tilde{t}_g^1$ is the gth element of $\tilde{\mathbf{t}}^1$. The first G_* genes in the reranked list are reported as differentially expressed.

5.2. Numerical Stability and Computational Complexity. Ordinarily, $M \ll G$, so that the sample correlation matrix is severely rank deficient. A small quantity (typically 10^{-10}) is added to the diagonal entries of $\widehat{\mathbf{R}}$ to make it invertible. After this augmentation, the algorithm above exhibits excellent numerical stability.

If implemented in a näive way, the matrix inversion to compute $\widehat{\mathbf{P}} = (\widehat{\mathbf{R}}^{00})^{-1}$ in (21) would be a prohibitive operation in most computing environments, since microarray data sets may have several tens of thousand genes. Determining the rightmost product in (21), $\widehat{\mathbf{P}}\mathbf{t}^0$, by solving the system of simultaneous linear equations $\widehat{\mathbf{R}}^{00}\mathbf{s}^0 = \mathbf{t}^0$ for the vector $\mathbf{s}^0 = \mathbf{P}\mathbf{t}^0$ is much faster than explicitly computing the matrix inverse $\widehat{\mathbf{P}}$ and forming the product. In particular, we can employ the Cholesky decomposition to exploit the fact that the matrix $\widehat{\mathbf{R}}^{00}$ is symmetric and positive definite (e.g., [46, Theorem 4.2.5]). MATLAB implementation uses the built-in function `linsolve` with appropriate settings, which, in turn, uses the highly optimized routines of `LAPACK` (http://www.netlib.org/lapack/).

The prostate cancer data [47] used in the experiments of Section 6 includes $G = 12625$ genes and $M = 102$ samples. For these data, the algorithm above implemented using MATLAB version R2006b on a computer with a 2.2 GHz dual-core AMD Opteron processor and 8 GB of RAM required ~40 seconds to report the final gene list. Similar implementation with explicit matrix inversions requires ~10 minutes. These times clearly indicate the *relative* benefit of avoiding the explicit matrix inversion, but the faster reporting time of ~40 sec should certainly not be interpreted as a lower bound for a problem of this scale. Indeed, workstations with 32 GB or more of RAM and with faster processors with eight or more processing cores are commercially available at modest costs. Of course, MATLAB is designed for modularity and

```
Input:   $\mathbf{X}$ = labeled $G \times M$ gene expression matrix
         $G_*$ = Desired size of differential gene list
Steps:
    (1) Calculate two-sample (unpaired) $t$-statistics as in equation (2).
    (2) Re-index genes such that $|t_1| \leq |t_2| \leq \cdots \leq |t_G|$.
    (3) Create $\mathbf{t}$ vector with elements ($t_g$ scores) in order of ascending magnitude.
    (4) Set $p_0$ (default value $\approx 0.5$). Set first $G_0 = Gp_0$ elements of $\mathbf{t}$ to zero ($\mathbf{t}^0$ partition)†.
    (5) Convert $\mathbf{X}$ to $\check{\mathbf{X}}$ by subtracting each gene's average response within each treatment group.
    (6) Compute $\widehat{\mathbf{R}}$ = sample correlation matrix of $\check{\mathbf{X}}$ (Section 4.4).
    (7) Find $\tilde{\mathbf{t}}^1$ as in (21). (See discussion of product $\widehat{\mathbf{P}}\mathbf{t}^0$ in Section 5.2)
    (8) Create a list of re-ranked genes in the descending order of the values $|\tilde{t}_g^1|$, in which $\tilde{t}_g^1$ is the $g$th
        element of $\tilde{\mathbf{t}}^1$.‡
    (9) Report $G_*$ genes with the largest re-ordered $|\tilde{t}_g^1|$ scores as statistical discoveries.
Output:  List of $G_*$ (experimentally-determined) most differentially-expressed genes following
         the decorrelation processing.
```

†For convenience, it is assumed that p_0 is chosen so that $p_0 G$ is an integer.

‡The original vector, $\mathbf{t}$, of raw t-scores (Step (2)) is ordered by *ascending* values of t_g. This ordering simplifies the definitions of certain quantities in the formal developments. Note: however, that the output list (Steps (8) and (9)) is created according to *descending* $|\tilde{t}_g^1|$ values. This is a more natural ordering for the end result as we are interested in only the G_* largest values.

Algorithm 1: Steps in the gene re-ranking procedure.

ease of use, not computational efficiency. Dedicated, lower-level coding of the reranking steps, implemented on a faster machine with more parallelism could reduce the reranking time significantly.

6. Experimental Results

6.1. Technical Comparisons. The reranking method developed above is compared with two leading techniques, SAM (Significance Analysis of Microarrays [31]) and EDGE (Extraction and Analysis of Differential Gene Expression [14, 48]). SAM adds a small exchangeability factor s_0 to the pooled sample standard deviation when computing the two-sample t-statistic:

$$t'_g = \frac{\overline{x}_{g|2} - \overline{x}_{g|1}}{s_g + s_0}, \tag{22}$$

whereas EDGE is based on a general framework for sharing information across tests. EDGE is reported to show substantial improvement in statistical power over five prominent techniques including SAM [14], the t/F-test [49, 50], the shrunken t/F-test [51], the empirical Bayes local FDR [40], and the *a posteriori* probability approach [52]. It is noteworthy that the reranking procedure developed here shows a significant performance improvement over EDGE in the experiments conducted. To determine the performance quality of various techniques, we focus primarily on the numbers of false positives, $\mathfrak{F}$, and the corresponding FDR values, $\overline{\mathfrak{F}}$, in the reported gene lists. Broadly speaking, the smaller the FDR, the better the technique.

6.2. Results

6.2.1. Prostate Cancer Data. The primary experiments reported in this paper are based on the prostate cancer data from the work of Singh et al. [47]. This database includes expression data for $G = 12625$ genes on $M = 102$ oligonucleotide microarrays, comparing $M_1 = 50$ healthy males with $M_2 = 52$ prostate cancer patients. The purpose of the Singh study is to identify genes that might anticipate the clinical behavior of prostate cancer. The .CEL files for the prostate study are publicly available at http://www-genome.wi.mit.edu/MPR/prostate. The general purpose of the present experiments is to compare performance of the reranking algorithm with the published state-of-the-art methods EDGE and SAM. The software RMAExpress [53] was used to obtain high-quality gene expressions from the posted data files. RMAExpress applied its in-built background adjustment; however, the quantile normalization was not used. To increase normality and stabilize across-group variances [54], each gene was represented in the final expression matrix $\mathbf{X}$ by the log of its expression level.

Comparative algorithm performance and insight into the inner workings of the reranking method required expression matrices for which truly differentially expressed genes were known *a priori*. Of course, in the nascent field of genomics, such knowledge is not available, and it is the very purpose of techniques like those discussed in this paper to seek such information. We approached this circular problem by using the prostate database to create test expression data with intergene correlation in $\mathbf{X}$ resembling that in the real microarray data. This was accomplished by first row standardizing the expression matrix from the prostate database.

In particular, the true prostate matrix $\mathbf{X}$ was transformed to $\check{\mathbf{X}}$ by subtracting each gene's average response within each treatment group and by normalizing within group sample mean squares. That is, for the individual treatment groups (for $k = 1$, then for $k = 2$) and for each g:

$$M_k^{-1} \sum_{\substack{m=1 \\ \text{Group } k}}^{M} \check{x}_{gm} = 0, \qquad M_k^{-1} \sum_{\substack{m=1 \\ \text{Group } k}}^{M} \check{x}_{gm}^2 = 1. \tag{23}$$

in which $\check{x}_{gm}$ is the (g, m) element of matrix $\check{\mathbf{X}}$. Each row represents one gene (and two conditions), so that with this transformation, all genes have equal energy and yet the same within group intergene correlation structure as the original $\mathbf{X}$. Normalizing within-group sample mean squares to unity is not implemented in the reranking algorithm. The normalization is done here prior to any processing as a first step in creating an expression matrix with known differentiation of expression across groups for each gene, but with realistic (derived from real data) intergene correlation.

To generate a test data set from $\check{\mathbf{X}}$, its 102 columns were randomly divided into groups of $M_1 = 50$ and $M_2 = 52$. Next, G_+ (G_-) genes were randomly chosen for up- (down-) regulation by adding a positive (negative) offset x_+ (x_-) to the corresponding entries in group 2. The total number of truly differentially expressed genes is denoted (G_δ is to be contrasted with G_*, the number of genes determined by experimentation to be differentially expressed):

$$G_\delta = G_+ + G_-. \tag{24}$$

We also denote by $p_\delta \overset{\text{def}}{=} G_\delta / G$ the proportion of *truly* differentially expressed genes, and, for future purposes, the ratio of the size of the desired gene-discovery list to the number of truly-differential genes, $p_{*\delta} \overset{\text{def}}{=} G_* / G_\delta$. In the experiments, various choices of the simulation parameters,

$$\mathscr{P}_\delta \overset{\text{def}}{=} \{p_\delta, G_+, G_-, x_+, x_-\}, \tag{25}$$

were tested to represent a range of data scenarios encountered in practice. Also associated with each trial is a set of parameters characterizing the gene-discovery experiments, say,

$$\mathscr{P}_e \overset{\text{def}}{=} \{p_0, G, G_*, M_1, M_2\}. \tag{26}$$

These parameter sets are detailed below.

In all experiments, results for the existing EDGE and SAM methods were obtained using the subroutines `statex.r` from the EDGE software package (http://www.genomine.org/edge/) and `samr.r` from the SAMR package (http://www-stat.stanford.edu/~tibs/SAM/), respectively. Both routines computed their native gene summary statistics given the matrix $\mathbf{X}$ and corresponding column labels. These statistics, in turn, were used to determine the top G_* genes. Results based on reranking proceed from the steps outlined in Algorithm 1 with G_* values corresponding to the largest $|\tilde{t}_g^1|$ scores of the reranked list.

Three experiments (cases) involving the prostate data are reported here. The first two cases were designed, by choice of the proportion of truly differential genes, p_δ, to represent typical conditions of two general classes of gene-discovery problems. The third case was carried out to test robustness of the technique to small sample size.

Case 1 (Small p_δ). In the first case, p_δ is small, $p_\delta \sim 0.01$–0.05, meaning that there are relatively few truly differentially expressed genes. The smaller p_δ is consistent with microarray investigations seeking genes that distinguish subtypes of cancer or diabetes, for example. The complete simulation parameter set for Case 1 is

$$\mathscr{P}_\delta^1 = \{p_\delta = 0.025, G_+ = 2G_- = 200, x_\pm = \pm 0.1\}, \tag{27}$$

where $x_\pm = \pm 0.1$ means that $x_+ = +0.1$ and $x_- = -0.1$. For numerical simplicity, we based the experiments on $G = 12,000$ of Singh's [47] gene expressions, so that $G_\delta = p_\delta G = 300$. Two sets of experiment parameters are used in Case 1, differing only in the size of the list of discovered genes.

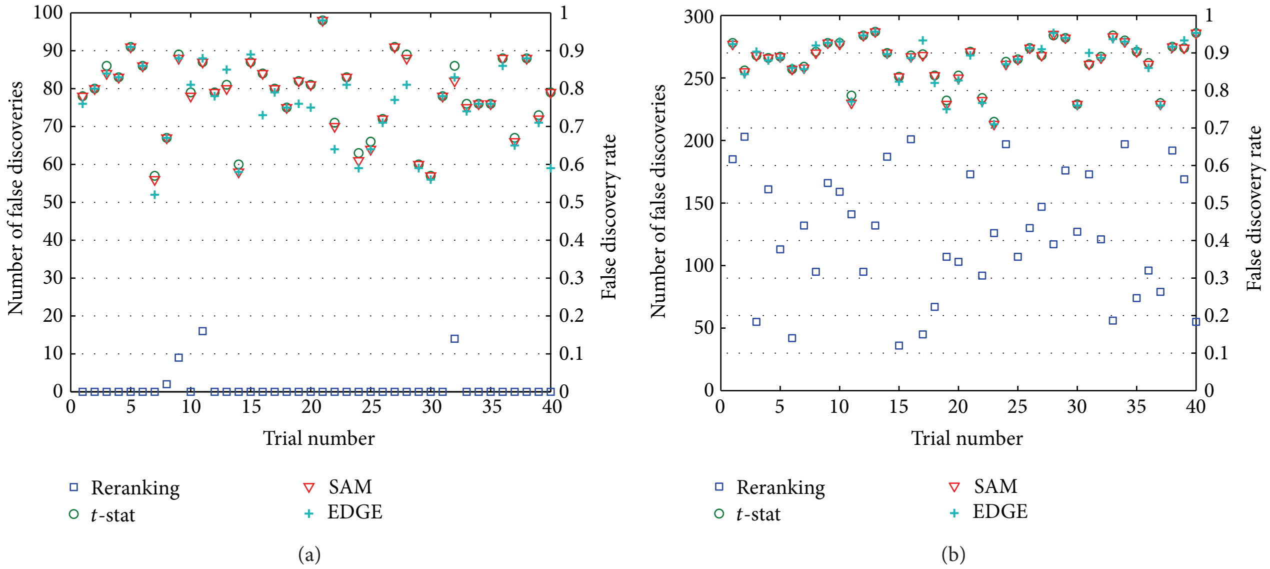
Subcase 1.1 (Small p_δ, Small $p_{*\delta}$). In the first experiment, the parameter set is

$$\mathscr{P}_e^1 = \{p_0 = 0.5, G = 12000, G_* = 100, M_1 = 50, M_2 = 52\}. \tag{28}$$

The size of the gene-discovery list, $G_* = 100$, is significantly smaller than $G_\delta = 300$, or $p_{*\delta} = G_* / G_\delta = 1/3$. In practice, a relatively small G_* would be chosen to identify high-quality, class-distinguishing features for expression-profiling-based clinical diagnosis and prognosis, in which the goal is to build accurate classifiers and predictors. Whereas Singh et al. [47] build a classifier around only 16 of 12625 features, they discuss the need to include as many reliable features as possible.

Figure 1(a) presents results for the test pair $(\mathscr{P}_\delta^1, \mathscr{P}_e^1)$ of Subcase 1.1. Remarkably, for 36 of 40 $\mathbf{X}$ matrices, the reranking procedure reports gene lists with no false discoveries at all, while the other techniques fail to obtain a single gene list with $\overline{\mathfrak{F}} < 0.5$. This result is typical of many "small p_δ" experiments carried out with an array of parameter sets. In particular, the *quality* of the results notwithstanding (as measured by $\overline{\mathfrak{F}}$, see below) the reranking strategy uniformly outperformed EDGE and SAM in every scenario.

In any rational detection algorithm built around a parametrized stochastic framework, it is possible to find regions of the parameter space in which performance deteriorates. In the "small p_δ" gene identification problem, for a fixed G and M, increasing G_* (more specifically, increasing the ratio $p_{*\delta}$) or decreasing the "signal" magnitudes of either up-(x_+) or down-(x_-) regulation, all create increasing probabilistic risk of false discoveries, $\mathfrak{F}$. As G_* was allowed to approach G_δ in Case 1 experiment above, the performance of all methods, EDGE, SAM, and reranking, all deteriorated as measured by $\mathfrak{F}$, yet the reranking approach remained consistently superior to the others according to this measure.



(a)

(b)

Figure 1: Results for two variations of Case 1 experiments. In both variations, the number of truly differential genes, $G_\delta = 300$, and the “signal strength” (amount of up- or downregulation represented by x_+ and x_-) is weak. (a) Results for Subcase 1.1 in which the size of the gene-discovery list, $G_* = 100$, is small relative to G_δ ($p_{*\delta} = 1/3$). The plot shows $\mathfrak{F}$ (left ordinate) and $\overline{\mathfrak{F}}$ (right ordinate) values over 40 trials. (b) $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ results, plotted similarly to those of part (a) of the figure, for Subcase 1.2, in which $G_* = 300$, large relatively to G_δ ($p_{*\delta} = 1$).

“Better,” however, does not always mean “good.” For illustration, we report a second Case 1 experiment.

Subcase 1.2 (Small p_δ, Large $p_{*\delta}$). In this variation of Case 1 experiment, we take $G_* = G_\delta = 300$. This is still a “small p_δ” situation, but with a “greedy” approach to gene discovery, an attempt to identify “all” genes (estimated to be) differentially expressed, $p_{*\delta} = 1$. Such a strategy would be employed in a microarray study designed to liberally identify a set of genes to be explored further—experimentally or computationally—to gain better understanding of underlying gene networks.

Figure 1(b) shows plots of the number of false positives, $\mathfrak{F}$, over 40 data sets for Subcase 1.2 experiment. The corresponding FDR, $\overline{\mathfrak{F}} = \mathfrak{F}/G_*$, is shown on the secondary ordinate axis. With $p_{*\delta} = 1$ and with relatively weak differential expression “signals” ($x_\pm = \pm 0.1$), identifying a good gene lists is not an easy task, as evident from the results. Among all methods only reranking achieved sufficiently low values of $\overline{\mathfrak{F}}$ to rescue a few $\mathbf{X}$ matrices, but, clearly, even reranking would not provide scientifically useful or reliable gene lists in this high-risk environment.

Case 2 (Large p_δ). The second case employs a larger $p_\delta \sim 0.1$, typical of studies comparing healthy versus diseased cell activities. Simulation parameters for this case are

$$\mathscr{P}_\delta^2 = \{p_\delta = 0.1, G_+ = G_- = 600, x_\pm = \pm 0.02\}. \quad (29)$$

Relative to Case 1, there are many more truly differentially expressed genes in Case 2 (increased by factor 4), thus decreasing the risk of false discoveries, especially for a small ratio $p_{*\delta}$. This is akin to an increased prior probability of a differentially expressed gene in a Bayesian detection strategy. To further challenge the algorithms in the light of the “increased prior,” the up/downregulation of expression was made considerably weaker in Case 2 (reduced by factor five relative to Case 1), as in Case 1, two sets of experiment parameters were used in Case 2, differing only in the size of the list of discovered genes.

Subcase 2.1 (Large p_δ, Small $p_{*\delta}$). In the first Case 2 test, the experiment parameter set is given by

$$\mathscr{P}_e^2 = \{p_0 = 0.5, G = 12000, G_* = 300, M_1 = 50, M_2 = 52\}. \quad (30)$$

The test pair $(\mathscr{P}_\delta^2, \mathscr{P}_e^2)$ represents a large p_δ proportion, but relatively small $p_{*\delta} = 300/1200 = 0.25$. The experimental results for this case over 40 trials are shown in Figure 2(a). In spite of the decreased signal strength, the reranking procedure produces no false discoveries in a vast majority of trials, similarly to the small p_δ, small $p_{*\delta}$, experiment of Subcase 1.1. EDGE and SAM consistently report a very large proportion of false discoveries (typically 250, or 90%).

Subcase 2.2 (Large p_δ, Large $p_{*\delta}$). A second Case 2 experiment was run to show the effects of “greedy” discovery lists, or large $p_{*\delta}$ ratios. The parameters $(\mathscr{P}_\delta^2, \mathscr{P}_e^2)$ remain identical to those in Subcase 2.1 experiment, except that $G_* = 1200$, so $p_{*\delta} = 1$. Results are shown in Figure 2(b). Like the large $p_{*\delta}$ experiment in Subcase 1.2, the reranking approach significantly outperforms EDGE and SAM, with typically $\overline{\mathfrak{F}} \sim 0.5$ for reranking and $\overline{\mathfrak{F}} \sim 0.9$ for the standard methods. The sample variances for $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ are notably smaller in Subcase 2.2 trials relative to similar trials in Subcase 1.2. Also as in

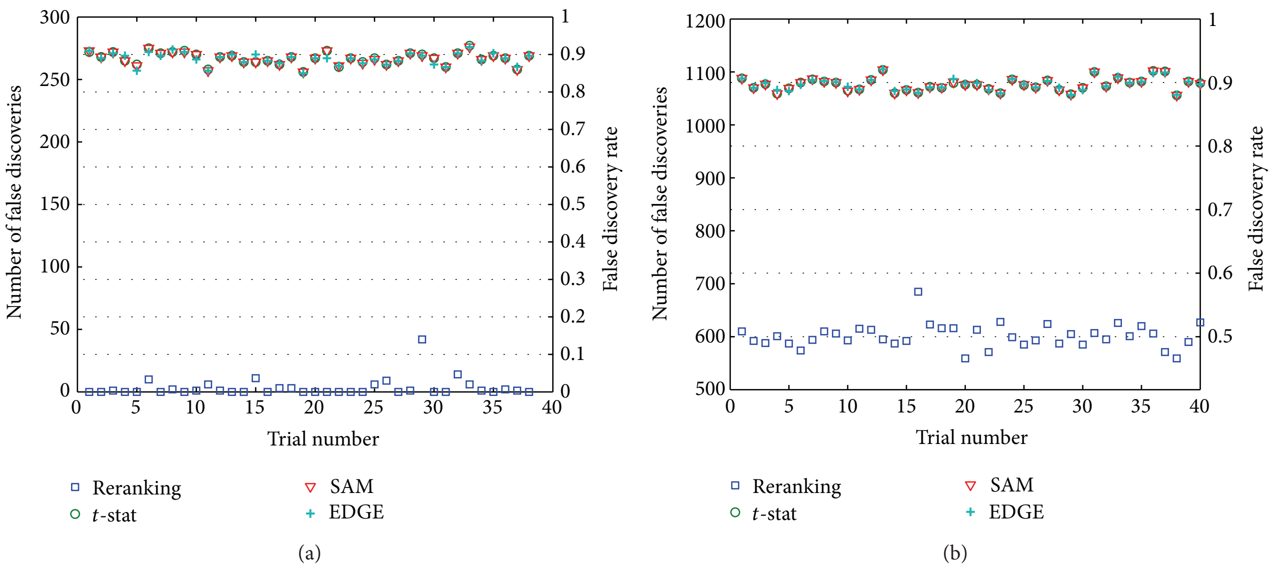


Figure 2: Case 2 $\mathfrak{F}$ (left ordinate) and $\overline{\mathfrak{F}}$ (right ordinate) values over 40 trials with larger p_δ data. The number of truly differential genes, $G_\delta = 1200$. The "signal" is very weak ($x_\pm = \pm 0.02$). (a) Subcase 2.1 results. $G_* = 300$, small relative to G_δ ($p_{*\delta} = 0.25$). The plot shows the number of false positives, $\mathfrak{F}$, over 40 data sets. The corresponding $\overline{\mathfrak{F}} = \mathfrak{F}/G_*$ is shown on the secondary ordinate axis. In spite of the relatively weak signal, reranking results in remarkably better performance than the standard approaches, and it produces the discovery list with no false positives in a majority of the 40 trials. (b) Subcase 2.2 results. $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ $G_* = 1200$ ($p_{*\delta} = 1$).

Subcase 1.2, reranking consistently outperforms the standard methods for large p_δ throughout the range $0 < p_{*\delta} \le 1$. At some application-dependent point in this range, reranking, although "better," is not sufficiently "good" at producing reliable discoveries. Clearly, in the present experiment in which $p_{*\delta} = 1$, the reranking result of typically $\overline{\mathfrak{F}} \sim 0.5$ is not indicative of reliable gene discoveries. For a few trials, the rate drops as low as $\overline{\mathfrak{F}} \sim 0.45$, but even this best-case rate implies ~540 false discoveries in the reported list of 1200 genes.

Case 3 (Tests of Robustness to Small Sample Size). The number of microarray chips available to a study, M, is ordinarily quite small compared with the number of genes investigated, G. The gene discovery operation is therefore required to draw conclusions from a sparse sampling of the gene expressions. To gain some insight into the robustness of the gene-discovery methods to small sample sizes, variations on Case 2 (large p_δ) experiments were repeated with the further stressor of a significant reduction in the sample space. M_1 and M_2 were each reduced to 20. So that observations could be more attributable to the sample size, M, the signal levels were increased back to Case 1 values of $x_\pm = \pm 0.1$. The simulation parameters used in Case 3, $\mathscr{P}_\delta^3$, are identical to $\mathscr{P}_\delta^2$ of (29). As in the previous cases, we ran two experiments, the first (Subcase 3.1) with a "conservative" gene discovery list, $G_* = 300$ or $p_{*\delta} = 0.25$ and the second (Subcase 3.2) with a "greedy" gene discovery list of size $G_* = 1200$ or $p_{*\delta} = 1$.

To create the data set for Case 3, 20 columns per treatment group were chosen randomly from the original prostate cancer expression matrix $\mathbf{X}$. The data generation process (including row standardization) detailed in Section 6.2.1 was then applied to the selected columns. Some compensation for the reduction in the number of samples is potentially present in the increased differential signal. The $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ values over 40 trials for the three methods are shown in Figure 3(a) for Subcase 3.1 and in Figure 3(b) for Subcase 3.2. Even with the significantly reduced sample size, the reranking process provides consistently superior performance with respect to existing methods. As in previous experiments, however, the better performance in the large $p_{*\delta}$ experiment of panel (b) does not mean that the results are necessarily reliable or useful. Nevertheless, these results suggest that the reranking procedure increases power in the analysis of small sample data sets.

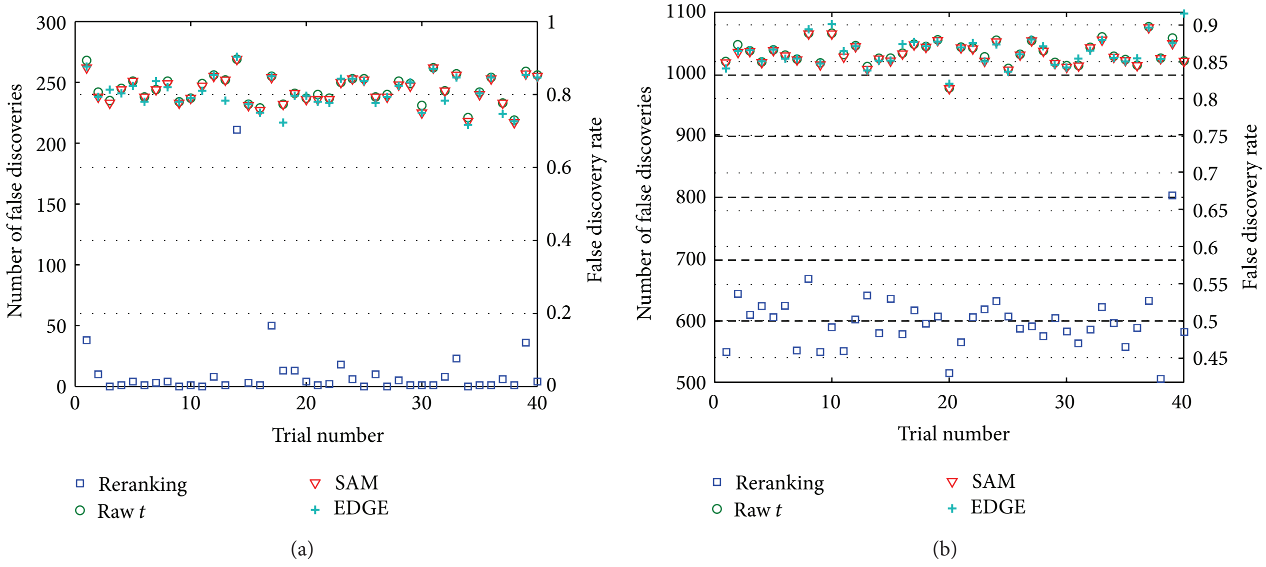
6.2.2. Simulated Data. Before devising the test data setup of Section 6.2.1, the reranking method was tested on several simulated data sets. We discuss some of these simulation results that shed further light on the small sample behavior of the method.

Let us denote by $\mathbf{x}_m$ the mth column of a simulated expression matrix $\mathbf{X}$. We assume that the random vector $\mathbf{x}_m$ is multivariate Gaussian with (stationary with index m) mean $\mathbf{0}_{G\times 1}$ and covariance matrix $\mathbf{\Lambda}$. Each such column represents $G = 3226$ genes, which, in their null expressions, are modeled by a covariance matrix $\mathbf{\Lambda}$ that introduces roughly the same amount of linear dependence as found in the BRCA data of [55]. We chose simulation parameters

$$\mathscr{P}_\delta = \{p_\delta = 0.031, G_+ = G_- = 50, x_\pm = \pm 1\}, \qquad (31)$$

and the experiments were run with parameters

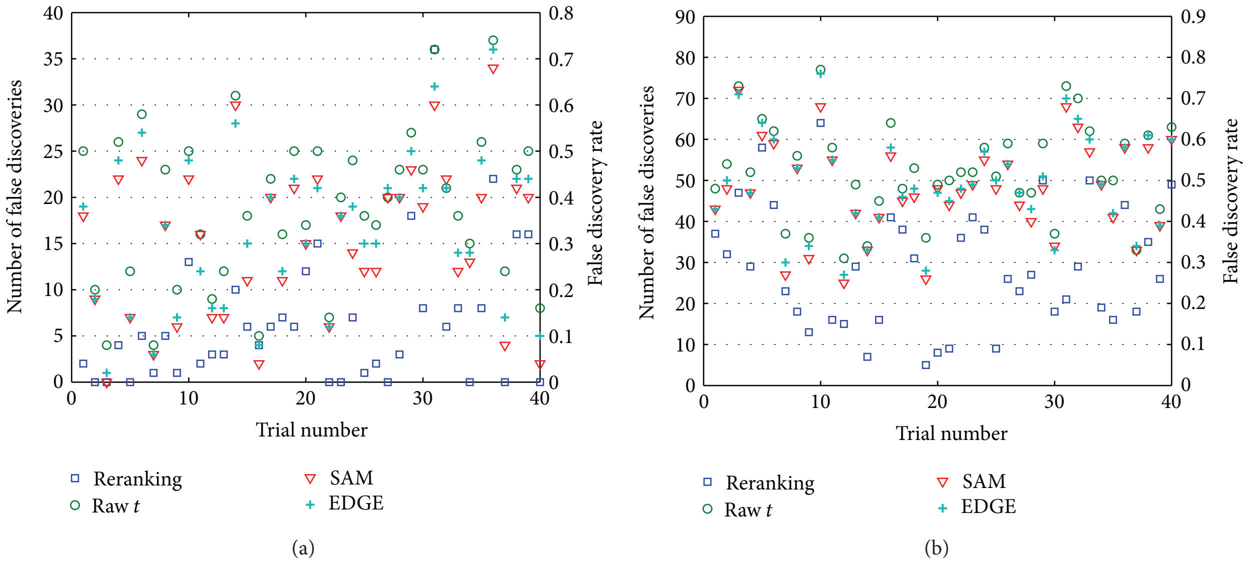
$$\mathscr{P}_e = \{p_0 = 0.5, G = 3226, G_*, M_1 = M_2 = 10\}, \qquad (32)$$



(a)

(b)

FIGURE 3: $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ values over 40 trials for Case 3 in which $G_\delta = 1200$. (a) Subcase 3.1, $G_* = 300$ ($p_{*\delta} = 0.25$); (b) Subcase 3.2, $G_* = 1200$ ($p_{*\delta} = 1$).



(a)

(b)

FIGURE 4: $\overline{\mathfrak{F}}$ values over 40 trials using simulated expression data. Sample sizes are very small, $M_1 = 10$, $M_2 = 10$. (a) $G_* = 50$ ($p_{*\delta} = 0.5$); (b) $G_* = 100$ ($p_{*\delta} = 1$).

for the two list sizes $G_* = 50$ ($p_{*\delta} = 0.5$) and $G_* = 100$ ($p_{*\delta} = 1$).

Figures 4(a) and 4(b) show plots of the $\mathfrak{F}$ and $\overline{\mathfrak{F}}$ values over 40 trials for $G_* = 50$ and $G_* = 100$, respectively. With a smaller M, preeminence of the reranking method scales down. Nevertheless, for 26 out of 40 simulated $\mathbf{X}$ realizations, reranking achieves an FDR $\overline{\mathfrak{F}} < 0.15$ and for 30 of 40 trials, $\overline{\mathfrak{F}} < 0.25$.

Table 1 shows results for some of the $\mathbf{X}$ realizations for $G_* = 100$ (Figure 4(b)). Shown are the largest 100 values of $|\tilde{t}_g^1|$ and each corresponding original t_g with concomitant rank. Even in this challenging case, the results indicate favorable aspects of the reranking procedure. First, it is noteworthy that reranking results in $\mathfrak{F} = 22$ false discoveries in the list of 100 genes, whereas $\mathfrak{F} = 68$ when raw t statistics are used. Further, all but two of the false discoveries reported by reranking received an even higher ranking by t-statistics. On the other hand, 46 of the correct discoveries by reranking would not have appeared in the list of 100 genes reported by t-statistics. Results like these were observed repeatedly in our data analysis. Consistent with the results of the prostate data studies, the reliability and utility of all techniques lessen as $p_{*\delta} \to 1$, yet, reranking persistently outperforms the other methods.

Table 1: Top 100 $\tilde{t}_g^1$ scores determined by re-ranking. Corresponding t_g scores and their ranks are also shown. The results are for some of the **X** realizations from Figure 4(b). Data representing false discoveries are printed in boldface. $\#_{\tilde{t}_g^1} \overset{\text{def}}{=}$ rank based on $\tilde{t}_g^1$ score, $\#_{t_g} \overset{\text{def}}{=}$ rank based on raw t_g score.

$\#_{\tilde{t}_g^1}$	$\tilde{t}_g^1$	t_g	$\#_{t_g}$
1–50			
1	4.22	5.87	1
2	−4.17	−5.55	2
3	−3.93	−4.26	5
4	−3.74	−4.12	7
5	−3.58	−4.49	4
6	−3.49	−3.34	28
7	−3.45	−4.25	6
8	3.35	3.87	10
9	−3.33	−3.20	35
10	3.33	3.77	13
11	3.25	3.42	25
12	−3.16	−2.18	260
13	3.14	4.54	3
14	3.10	2.87	65
15	−3.08	−3.54	17
16	−3.07	−2.80	80
17	3.06	3.49	20
18	3.02	2.29	213
19	−2.99	−3.34	27
20	−2.93	−3.13	38
21	−2.92	−2.92	57
22	2.86	3.26	31
23	−2.83	−2.82	74
24	2.82	2.37	180
25	−2.81	−2.13	276
26	**2.81**	**3.48**	**21**
27	−2.79	−3.01	47
28	2.70	2.87	64
29	−2.66	−3.15	37
30	**−2.58**	**−3.85**	**11**
31	−2.56	−2.84	71
32	−2.55	−1.72	524
33	−2.54	−2.63	106
34	−2.54	−2.69	98
35	2.53	2.30	209
36	2.48	2.45	148
37	−2.47	−2.29	212
38	−2.46	−3.21	33
39	−2.43	−2.44	154
40	2.43	2.71	94
41	2.40	2.86	66
42	−2.34	−2.60	115
43	−2.34	−2.98	50
44	**−2.33**	**−3.80**	**12**
45	−2.32	−2.06	306
46	2.29	1.81	444
47	−2.27	−1.17	1110
48	2.26	1.97	347
49	**2.24**	**3.75**	**14**
50	**2.20**	**3.88**	**9**
51–100			
51	**2.18**	**3.45**	**23**
52	**2.17**	**3.05**	**42**
53	2.16	2.80	82
54	−2.15	−2.57	122
55	2.15	1.96	357
56	−2.14	−1.47	751
57	2.13	2.25	229
58	2.13	1.77	486
59	2.13	1.44	785
60	2.12	2.14	273
61	−2.11	−1.48	744
62	**2.10**	**1.80**	**453**
63	2.09	2.60	114
64	−2.09	−2.05	312
65	2.09	2.70	96
66	2.09	2.23	237
67	2.08	2.34	188
68	−2.08	−2.24	232
69	−2.06	−2.53	130
70	−2.04	−2.11	283
71	**−2.04**	**−2.95**	**54**
72	**−2.03**	**−3.08**	**40**
73	−2.02	−2.30	210
74	**−2.01**	**−3.67**	**15**
75	2.00	2.62	109
76	−1.98	−2.38	171
77	1.98	1.43	795
78	1.96	1.69	549
79	−1.95	−1.47	746
80	1.95	1.95	361
81	**1.95**	**2.81**	**77**
82	−1.94	−1.41	813
83	**1.94**	**3.40**	**26**
84	1.94	1.30	948
85	**−1.94**	**−3.27**	**30**
86	−1.93	−1.11	1190
87	−1.93	−1.37	872
88	**−1.93**	**−3.44**	**24**
89	**−1.92**	**−3.07**	**41**
90	−1.92	−1.50	726
91	**1.90**	**3.62**	**16**
92	**−1.90**	**−2.82**	**75**
93	1.89	1.25	1007
94	**1.87**	**3.89**	**8**
95	**−1.86**	**−3.49**	**19**
96	**−1.86**	**−2.08**	**300**
97	1.85	1.20	1074
98	**−1.83**	**−2.90**	**60**
99	1.83	1.39	833
100	−1.82	−1.95	367

It is notable that, with a smaller M, SAM outperforms EDGE and the use of raw t-scores. This is not entirely surprising as a smaller M can make the noise in the per gene pooled variance s_i (and possibly the equivalent quantity in the EDGE algorithm) more prominent. SAM mitigates this issue in some measure by using the exchangeability factor s_0 to adjust the effective pooled variance [31].

7. Discussion and Conclusions

In most microarray data, there are at least three resources that can be used to advantage: (i) identifiability, (ii) parallel structure, and (iii) intergene correlation itself. Analysis in papers by Efron [56, 57] suggests this view of the rich information structure inherent in the data. In this light, reranking can be viewed as exploiting more than correlation as a means of sharing information across tests, as it also involves identifiability.

Limited time and resources often require biomedical researchers to work on only a small number of "hot (gene) prospects." Even under such highly conservative conditions, however, misleading results can occur, as is evident in the results of Figures 1–4. For all their expert development and statistical power, even state-of-the-art tools like SAM and EDGE can report spurious gene lists. The extra statistical power of reranking promises to further guard against anomalous results that can have serious consequences for the study of gene function, causation, and interaction.

In summary, this paper has reported the development and testing of a novel framework for the detection of differential gene expression. The framework exploits identifiability—the fact that in most microarray data sets, a large proportion of genes can be identified *a priori* as nondifferential—to reduce the correlation in the expression data for the remaining gene candidates. When applied to the widely used two-sample t-statistic approach, this viewpoint yielded a simple differential analysis technique which requires as inputs only a gene expression matrix, related two-sample labels, and the size of desired output gene-list G_*. The method was tested on data constructed from the prostate cancer database of Singh et al. [47] and some simulated data. Compared with SAM [31], EDGE [14], and the raw t-statistic approach itself, reranking shows substantial improvement in statistical power. As is the case with all published techniques, the reranking process' power tends to increase considerably with an increase in the number of microarray samples. However, even for small sample sizes, performance was significantly better than the alternatives in the experiments conducted here.

Protection of Human Subjects

Human subject data used in this study are publicly available and anonymous and are therefore exempt from continuing Internal Review Board scrutiny according to U.S. Health and Human Services Policy 45 CFR 36, Subpart A, §46.101 (2.b.4).

Acknowledgments

This work was supported in part by the Quantitative Biology Initiative at Michigan State University. The authors are grateful to the personnel in the MSU High Performance Computing Center for assistance in implementing the extensive computer simulations, and to Dr. K. H. Desai for many contributions to this work.

References

[1] E. Lander, "Array of hope," *Nature Genetics*, vol. 21, pp. 3–4, 1999.

[2] S. Frantz, "An array of problems," *National Review of Drug Discovery*, vol. 4, pp. 362–363, 2005.

[3] A. B. Owen, "Variance of the number of false discoveries," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 3, pp. 411–426, 2005.

[4] B. Efron, "Correlation and large-scale simultaneous significance testing," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 93–103, 2007.

[5] B. Efron, "Size, power, and false discovery rates," *Annals of Statistics*, vol. 35, no. 4, pp. 1351–1377, 2007.

[6] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, article e161, 2007.

[7] S. Degrelle et al., "Amplification biases: Possible differences among deviating gene expressions," *BMC Genomics*, vol. 9, article 46, 2008.

[8] X. Qiu, L. Klebanov, and A. Yakovlev, "Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes," *Statistics Applications in Genetics and Molecular Biology*, vol. 4, article 34, 2005.

[9] X. Qiu and A. Yakovlev, "Some comments of instability of false discovery rate estimation," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 5, pp. 1057–1068, 2006.

[10] T. Yu, H. Peng, and W. Sun, "Incorporating nonlinear relationships in microarray missing value imputation," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 3, pp. 723–731, 2011.

[11] K. Desai, J. R. Deller Jr., and J. J. McCormick, "The distribution of the number of false discoveries in DNA microarray data," in *Proceedings of the IEEE Statistical Signal Processing Workshop*, pp. 205–209, Madison, Wis, USA, August 2007.

[12] J. Deller Jr., H. Radha, J. McCormick, and H. Wang, "Nonlinear dependence in the discovery of differentially-expressed genes," *ISRN Bioinformatics*, vol. 2012, Article ID 564715, 18 pages, 2012.

[13] Y. Pawitan, K. R. K. Murthy, S. Michiels, and A. Ploner, "Bias in the estimation of false discovery rate in microarray studies," *Bioinformatics*, vol. 21, no. 20, pp. 3865–3872, 2005.

[14] J. D. Storey, J. Y. Dai, and J. T. Leek, "The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments," *Biostatistics*, vol. 8, no. 2, pp. 414–432, 2007.

[15] R. Tibshirani and L. Wasserman, "Correlation-sharing for detection of differential gene expression," http://arxiv.org/pdf/math/0608061.pdf.

[16] R. Hu, X. Qiu, and G. Glazko, "A new gene selection procedure based on the covariance distance," *Bioinformatics*, vol. 26, no. 3, pp. 348–354, 2010.

[17] Q. Cui, B. Liu, T. Jiang, and S. Ma, "Characterizing the dynamic connectivity between genes by variable parameter regression and Kalman filtering based on temporal gene expression data," *Bioinformatics*, vol. 21, no. 8, pp. 1538–1541, 2005.

[18] V. Martyanov and R. H. Gross, "Identifying functional relationships within sets of co-expressed genes by combining upstream regulatory motif analysis and gene expression information," *BMC Genomics*, vol. 11, supplement 2, article S8, 2010.

[19] R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork, "The importance of phase information for human genomics," *Nature Reviews Genetics*, vol. 12, no. 3, pp. 215–223, 2011.

[20] M. Dettling, E. Gabrielson, and G. Parmigiani, "Searching for differentially expressed gene combinations," *Genome Biology*, vol. 6, no. 10, article R88, 2005.

[21] Y. Lai, B. Wu, L. Chen, and H. Zhao, "A statistical method for identifying gene-gene co-expression dynamics," *Bioinformatics*, vol. 20, no. 17, pp. 3146–3155, 2004.

[22] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinformatics*, vol. 25, no. 21, pp. 2780–2786, 2009.

[23] E. Huerta, B. Duval, and J. K. Hao, "Fuzzy logic for elimination of redundant information of microarray data," *Genomics, Proteomics and Bioinformatics*, vol. 6, no. 2, pp. 61–73, 2008.

[24] F. Reverter, E. Vegas, and P. Sánchez, "Mining gene expression profiles: An integrated implementation of Kernel principal component analysis and singular value decomposition," *Genomics, Proteomics and Bioinformatics*, vol. 8, no. 3, pp. 200–210, 2010.

[25] S. Yang, X. Guo, and H. Hu, "MOF: An R function to detect outlier microarray," *Genomics, Proteomics and Bioinformatics*, vol. 6, no. 3-4, pp. 186–189, 2008.

[26] Z. Xiang, Z. S. Qin, and Y. He, "CRCView: A web server for analyzing and visualizing microarray gene expression data using model-based clustering," *Bioinformatics*, vol. 23, no. 14, pp. 1843–1845, 2007.

[27] A. K. C. Wong, W.-H. Au, and K. C. C. Chan, "Discovering high-order patterns of gene expression levels," *Journal of Computational Biology*, vol. 15, no. 6, pp. 625–637, 2008.

[28] S. Bandyopadhyay and M. Bhattacharyya, "A biologically inspired measure for co-expression analysis," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 8, no. 4, pp. 929–942, 2011.

[29] L. Dalton, V. Ballarin, and M. Brun, "Clustering algorithms: On learning, validation, performance, and applications to genomics," *Current Genomics*, vol. 10, no. 6, pp. 430–445, 2009.

[30] N. Ancona, R. Maglietta, A. Piepoli et al., "On the statistical assessment of classifiers using DNA microarray data," *BMC Bioinformatics*, vol. 7, article 387, 2006.

[31] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[32] B. Efron, "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.

[33] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[34] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[35] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B*, vol. 64, no. 3, pp. 479–498, 2002.

[36] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach," *Journal of the Royal Statistical Society, Series B*, vol. 66, no. 1, pp. 187–205, 2004.

[37] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 4, pp. 555–572, 2005.

[38] M. L. T. Lee, F. C. Kuo, G. A. Whitmore, and J. Sklar, "Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 18, pp. 9834–9839, 2000.

[39] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui, "On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data," *Journal of Computational Biology*, vol. 8, no. 1, pp. 37–52, 2001.

[40] B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1151–1160, 2001.

[41] N. Akhiezer and I. Glazman, *Theory of Linear Operators in Hilbert Space*, Dover, New York, NY, USA, 1993.

[42] A. Friedman, *Foundations of Modern Analysis*, chapter 6, Dover, New York, NY, USA, 1982.

[43] S. Lang, *Real and Functional Analysis*, chapter 5, Springer, New York, NY, USA, 3rd edition, 1993.

[44] S. Roman, *Advanced Linear Algebra*, chapter 13, Springer, New York, NY, USA, 1992.

[45] K. Petersen and M. Pedersen, *The Matrix Cookbook*, 2008, http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf.

[46] G. Golub and C. van Loan, *Matrix Computations*, The Johns Hopkins University Press, 3rd edition, 1996.

[47] D. Singh, P. G. Febbo, K. Ross et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.

[48] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: Extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, no. 4, pp. 507–508, 2006.

[49] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of Computational Biology*, vol. 7, no. 6, pp. 819–837, 2001.

[50] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[51] X. Gui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, vol. 6, no. 1, pp. 59–75, 2005.

[52] I. Lonnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, no. 1, pp. 31–46, 2002.

[53] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of Affymetrix GeneChip probe level data," *Nucleic Acids Research*, vol. 31, no. 4, article e15, 2003.

[54] C. A. Tsai, Y. J. Chen, and J. J. Chen, "Testing for differentially expressed genes with microarray data," *Nucleic Acids Research*, vol. 31, no. 9, article e52, 2003.

[55] I. Hedenfalk, D. Duggan, Y. Chen et al., "Gene-expression profiles in hereditary breast cancer," *The New England Journal of Medicine*, vol. 344, no. 8, pp. 539–548, 2001.

[56] B. Efron, "Bayesians, frequentists, and scientists," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 1–5, 2005.

[57] B. Efron, "R.A. Fisher in the 21st century," in *Statistics for the 21st Century: Methodologies for Applications of the Future*, vol. 1, p. 9, 2000.