Exploiting identifiability and inter-gene correlation for improved detection of differential expression

J.R. Deller, Jr. and Hayder Radha
Michigan State University
Department of Electrical and Computer Engineering, 2120 EB
East Lansing, MI 48824, U.S.A.
deller@egr.msu.edu, radha@egr.msu.edu

J. Justin McCormick
Michigan State University
Department of Molecular Biology & Biochemistry
Carcinogenesis Laboratory, 341 FST
East Lansing, MI 48824, U.S.A.
mccormi1@msu.edu

Abstract
Accurate differential analysis of microarray data strongly depends on effective treatment of inter-gene correlation. Such dependence is ordinarily accounted for in terms of its effect on significance cut-offs. 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 non-differential with very high confidence], but (ii) doing so in a way that accounts for linear dependencies among identifiable and non-identifiable 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 non-identified 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.
I. 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 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, inter-gene correlation is attributable to co-expression 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], [11], [12].

Correctly detecting differentially-expressed genes — or the related task of estimating the FDR — in the presence of substantial inter-gene 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 [15] discusses a quantity called the “correlation-shared” t-statistic, and derives theoretical bounds on its performance. Hu et al. [16] examine the covariance structure of the expression data and discover benefits of linking co-expression and differential expression in a distance measure — reflecting the more recent interest in characterizing broader statistical patterns in microarray data.

Recent research into the hybrid differential-expression—co-expression problem has also yielded results and methods that could ultimately benefit the gene discovery problem. Because the differential co-expression research is often concerned with differing phenotypes, rather than with different treatment conditions, two given research efforts involving differential co-expression 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], [18], [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 co-expression analysis [22], fuzzy expression level assignments [23], expression-profile mining with decorrelation [24], 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 non-differential 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 re-ranking criterion that accounts for inter-gene 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.

II. Problem Formulation

Suppose that expression data for G genes are measured on M microarrays, resulting in a gene expression matrix, say $X \in \mathbb{R}^{G \times M}$, with $(g,m)$ element $x_{gm}$. Each of the M microarray experiments takes place under one of two conditions (indexed by $k = 1$ or 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, $X$, we seek to identify a “small” number, $G_s \ll G$, of genes that are significantly differentially expressed between the two groups. One widely-used strategy (e.g., [31], [32]) is to hypothesize that the each gene is not differentially expressed. We refer to this as the null hypothesis, denoted $H_0$. 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 $H_0$. Gene $g$ is tested against $H_0$ 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 \( \tilde{\nu} \), or, equivalently, the FDR, defined as \( \tilde{\nu} \equiv \tilde{\nu}/G_* \), to find a maximal \( G_* \) with the allowable \( \tilde{\nu} \) or \( \tilde{\nu} \) (e.g., [3], [5], [13], [33], [34], [35], [36], [37]).

As discussed in [4] and [11], for an “over powered” \( X \) matrix, there may be significantly fewer tail-area null counts than expected, whereas for an “under powered” \( X \), 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 “under powered” \( X \), 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 \( X \) matrices.

The new method uses a vector of \( t \)-statistics,
\[
\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \cdots & t_G \end{bmatrix}^T,
\]
and an estimate of the covariance matrix of the vector \( \mathbf{t} \), to output a substantially revised version of \( \mathbf{t} \), denoted \( \mathbf{\tau} \), whose entries provide an improved gene-ranking. For expediency, we will refer to the procedure that produces \( \mathbf{\tau} \) from \( \mathbf{t} \) as correlation adjusted re-ranking, or simply re-ranking. The essence of re-ranking is embodied in some fundamental data conditioning procedures to effect the two outcomes mentioned in the introduction – exploiting identifiability, and nullifying the effects of inter-gene correlation between identified and non-identified genes.

In the following sections we develop the theoretical basis for the re-ranking process. The performance of re-ranking 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.

III. Methods

A. 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 \( \mu \in \mathbb{R}^G \) and covariance matrix \( \Sigma \in \mathbb{R}^{G \times G} \). Henceforth, we write \( \mathbf{t} \sim (\mu, \Sigma) \). This is a theoretical model only, as \( \mu \) and \( \Sigma \) are generally unknown. No further distribution information is required.

Let \( \overline{\tau}_g \) denote the average differential expression level for gene \( g \), \( \overline{\tau}_g = M^{-1} \sum_{m=1}^M \tau_{g,m} \). Further, let \( \overline{\tau}_{g,k} \) be the average expression level for gene \( g \) in treatment condition \( k \). Then, the (unpaired) \( t \)-statistic for gene \( g \) is computed as
\[
t_g = \frac{\overline{\tau}_{g,2} - \overline{\tau}_{g,1}}{s_g},
\]
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 \( \nu \) 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 \( \mu_g \) and \( \sigma_g^2 \) denoting the \( g^{th} \) element of \( \mu \), and the \( g^{th} \) diagonal element of \( \Sigma \), respectively. For \( g \notin \mathbb{H}_0 \), \( \mu_g \) and \( \sigma_g^2 \) depend on the amount of up- or down-regulation of the gene expression, the number of samples in each treatment group, and the number of degrees of freedom, \( \nu \).

B. Invoking identifiability: The zero assumption

The direct use of \( t \)-scores for ranking neglects some important information that is inherent in the microarray matrix \( X \). 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 – satisfy \( \mathbb{H}_0 \). The ZA plays a central role in the literature on estimating the proportion of null genes, as in [13], [38]. The ZA is equally crucial for the two-group model approach developed in the Bayesian microarray literature, as in [39], [40], [41]. Furthermore, the assertion that the method developed by Storey [36] improves upon the well-known Benjamini-Hochberg FDR procedure [34] (in terms of statistical power) crucially relies on an adaptive version of the ZA.

The use of the ZA is justified in the re-ranking 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 \( \sim 50\% \) of the genes in a cell would not be differentially expressed in the formation of prostate cancer. Accordingly, in the initial step of the re-ranking 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 \( G_1 \equiv G - G_0 \) genes, so-named because they remain “candidates” to become “discovered” genes (i.e., to be among the \( G_* \) 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 \( \mathbb{E}\{t_g\} = 0 \) 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 re-ranking 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 co-expressed 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 in re-ranked 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

$$|t_1| \leq |t_2| \leq \cdots \leq |t_{G_0}| \leq |t_{G_0+1}| \leq \cdots \leq |t_G|,$$

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 $t_j$ 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 non-null designation (the candidate set), $\{t_{G_0+1}, t_{G_0+2}, \ldots, t_G\}$. For convenience, express the vector $t$ in terms of these two partitions,

$$t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{G_0} \\ t_{G_0+1} \\ t_{G_0+2} \\ \vdots \\ t_G \end{bmatrix}^T = \begin{bmatrix} t_0 \\ t_1 \end{bmatrix}. \quad (4)$$

The random vector $t$ has the following moments:

$$t = \begin{bmatrix} t_0 \\ t_1 \end{bmatrix} \sim \left( \mu = \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix} \right), \quad (5)$$

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

C. Theoretical estimator of $\tau$

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

To derive the desired expression for $\tilde{t}^1$, we adopt a simple approach based on well-known ideas from the theory of linear operators (e.g., [42, 43, 44, 45]). Let us view the space of random vectors in $\mathbb{R}^{G_1}$ as a Hilbert space, $\mathcal{H}$, with inner product $\langle v, w \rangle \overset{\text{def}}{=} \mathcal{E}\{v^T w\}$ for all $v, w \in \mathbb{R}^{G_1}$. The inner product induces a norm on $\mathcal{H}$ given by $\|v\|_2 \overset{\text{def}}{=} \sqrt{\mathcal{E}\{v^T v\}}$, and, in turn, a metric

$$d(v, w) \overset{\text{def}}{=} \|v - w\|_2 = \sqrt{\mathcal{E}\{(v - w)^T (v - w)\}}. \quad (6)$$

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

$$d(v, \tilde{v}) = \arg\min_{\tilde{w} \in \mathcal{S}} d(v, \tilde{w}) \quad (7)$$

1We are assuming $G_1 \geq G_0$, or $p_0 \leq 0.5$. If the converse be true, simply reverse the roles of the $\mathbb{R}^{G_1}$ and $\mathbb{R}^{G_0}$ spaces in this development.
in which \( d \) is the metric in (6). The \textit{Hilbert projection theorem} states the solution \( \hat{v} \) exists and is unique. Moreover, \( \hat{v} \) has the property that the vector difference between \( v \) and \( \hat{v} \) is orthogonal to all vectors in the subspace \( S \); that is, \( (v - \hat{v}) \in S^\perp \), in which \( S^\perp \) is the orthogonal complement to subspace \( S \) in \( \mathbb{R}^{G_1} \). In particular, random vector \( (v - \hat{v}) \) is orthogonal to \( v \in S^\perp \) which means that \( \langle (v - \hat{v}), v \rangle = 0 \) by definition. Thus, \( \mathcal{E}\{ (v - \hat{v})^T v \} = 0 \), implying that \( (v - \hat{v}) \) and \( v \) are stochastically orthogonal (uncorrelated if \( \mu_v = 0_{G_1 \times 1} \) or \( \mu_{v} = 0_{G_1 \times 1} \)).

The Hilbert-space formulation provides the structure in which achieve the desired decomposition of the random vector \( t_c^1 \) into components that are correlated, and uncorrelated, with \( t^0 \). Vector \( t_c^1 \) 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 \( t^0 \)) that is close to \( t_c^1 \). (This will be the component of \( t_c^1 \) that is correlated with \( t^0 \).) 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 “\( S \)” in which to find an optimal vector; rather we are given only a vector \( t^0 \) 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 \( t_c^1 \). 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 \( t^0 \) and its implicit initial subspace \( \mathbb{R}^{G_0} \).

Given \( t^0 \), let us conceptualize a \( G_0 \)-dimensional subspace of \( \mathbb{R}^{G_1} \), \( S \), consisting of all the (random) vectors \( \{ v : v = Ft^0, \ F : \mathbb{R}^{G_0} \rightarrow \mathbb{R}^{G_1} \} \). \( F \) 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 \( t^0 \), there is a unique vector \( \hat{t}_c^1 = \hat{F}t^0 \in S \), hence a unique linear operator, \( \hat{F} \), such that

\[
d(t_c^1, \hat{F}t^0) = \arg\min_{F \in \mathbb{R}^{G_1 \times G_0}} d(t_c^1, Ft^0) = \arg\min_{F \in \mathbb{R}^{G_1 \times G_0}} \| t_c^1 - Ft^0 \| = \arg\min_{F \in \mathbb{R}^{G_1 \times G_0}} \sqrt{(t_c^1 - Ft^0)^T(t_c^1 - Ft^0)} = \arg\min_{F \in \mathbb{R}^{G_1 \times G_0}} \mathcal{E}\{ (t_c^1 - Ft^0)^T(t_c^1 - Ft^0) \}.
\]

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

\[
d^2(t_c^1, Ft^0) = \mathcal{E}\{ (t_c^1 - Ft^0)^T(t_c^1 - Ft^0) \} = \mathcal{E}\{ (t_c^1)^T t_c^1 \} - 2\mathcal{E}\{ t_c^1 F t^0 \} + \mathcal{E}\{ t^0 (t^0)^T \}
\]

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

\[
\nabla_F d^2(t_c^1, Ft^0) = -2\mathcal{E}\{ t_c^1 (t^0)^T \} + 2F \mathcal{E}\{ t^0 (t^0)^T \} \equiv 0_{G_1 \times G_0}.
\]

Recognizing that [recall Eqn. (5)] \( \mathcal{E}\{ t_c^1 (t^0)^T \} = \Sigma^{10} \) and \( \mathcal{E}\{ t^0 (t^0)^T \} = \Sigma^{00} \), the solution becomes

\[
\hat{F} = \Sigma^{10}(\Sigma^{00})^{-1}.
\]

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

\[
\Sigma^{\hat{t}1} = \mathcal{E}\{ \hat{t}_c^1 (t_c^1)^T \} = \mathcal{E}\{ \hat{F}t^0 (t_c^1)^T \} = \hat{F} \Sigma^{01} = \Sigma^{10}(\Sigma^{00})^{-1} \Sigma^{01}.
\]

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

\[
t_c^1(t) = t_c^1 - \hat{t}_c^1 = t_c^1 - \hat{F}t^0 = t_c^1 - \Sigma^{10}(\Sigma^{00})^{-1} t^0.
\]

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

\[
\tau = t^1 - \Sigma^{10}(\Sigma^{00})^{-1} t^0.
\]
D. Sample estimator of $\hat{t}_1$

Estimates of the elements of $\Sigma$ are required. For this purpose, we make several observations that can be verified through simulation using the MATLAB script testcov.m available with the software developed for this research (see www.msu.egr.edu/~deller). Note that Eqn. (15) does not require the covariance between $t_g$ and $t_{g'}$ when $g, g' \notin H_0$.

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

$$
\gamma(u, v) \overset{\text{def}}{=} E\{[u - E(u)][v - E(v)]\}
$$

$$
\rho(u, v) \overset{\text{def}}{=} E\{uv\} = \gamma(u, v) + E(u)E(v).
$$

In these terms, we make the following observations:

Observation 1. If $g, g' \in H_0$, then [3], [4]

$$
\gamma(t_g, t_{g'}) \approx \frac{\nu}{\nu - 2} \rho(x_g, x_{g'}). \tag{18}
$$

Observation 2. If $g \in H_0$ and $g' \notin H_0$ (or conversely), then

$$
\gamma(t_g, t_{g'}) \approx \frac{\nu}{\nu - 2} \frac{M_2 \text{Cor}(x_{g|1}, x_{g'|1}) + M_1 \text{Cor}(x_{g|2}, x_{g'|2})}{M_1 + M_2}. \tag{19}
$$

Equation (19) accommodates the possibility that the correlation between a null and a non-null gene may change between treatment groups. If this does not occur, then (19) reduces to (18).

Observation 3. Furthermore, if $M_1 \approx M_2$ (true for most microarray studies), then Eqn. (19) simplifies to:

$$
\gamma(t_g, t_{g'}) \approx \frac{\nu}{\nu - 2} \frac{\text{Cor}(x_{g|1}, x_{g'|1}) + \text{Cor}(x_{g|2}, x_{g'|2})}{2}. \tag{20}
$$

Equations (18) and (20) suggest that we may use the sample correlation to estimate $\gamma(t_g, t_{g'})$:

$$
\hat{\gamma}(t_g, t_{g'}) \approx \frac{\sum_m \bar{x}_{gm} \bar{x}_{g'm}}{\sqrt{(\sum_m \bar{x}_{gm}^2)(\sum_m \bar{x}_{g'm}^2)}}, \tag{21}
$$

where $\bar{x}_{gm}$ denotes the expression level of the $g$th gene measured on the $m$th 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 ($\Sigma^{00}$)$^{-1}$ and $\Sigma^{01}$ are multiplied in Eqn. (15), so that estimating $\nu$ is not required.

In light of Eqn. (21), Eqn. (15) takes the practical form

$$
\tilde{t}^1 = t^1 - \hat{R}^{10} (\hat{R}^{00})^{-1} t^0 \overset{\text{def}}{=} t^1 - \hat{R}^{10} \hat{P} t^0, \tag{22}
$$

where $\hat{R}^{00}$ and $\hat{R}^{10}$ are the partitions [similarly to Eqn. (5)] of $\hat{R}$, the sample correlation matrix of the gene expression matrix $\hat{X}$ (after removing the treatment effects). For notation compactness, we have defined $\hat{P} \overset{\text{def}}{=} (\hat{R}^{00})^{-1}$. In principle, the set of elements in the vector $\tilde{t}^1$ of Eqn. (15) embodies the gene-expression re-ranking 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 $\tilde{t}^{-1}$ in Eqn. (22).

IV. IMPLEMENTATION

A. Gene re-ranking algorithm

A stepwise procedure for the gene re-ranking is given in Fig. 1. The process begins by re-indexing the genes based on their two-sample $t$-statistics [Eqn. (3)]. Then, based on the ZA, the first $G_0$ genes are declared to satisfy $H_0$. In the experiments reported below, $p_0$, the fraction of genes declared to be identifiable, is set to $\approx 0.5$ by default. 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, $X$ is converted to $\hat{X}$ by subtracting each gene’s average response within each treatment group. The sample correlation matrix $\hat{R}$ of $\hat{X}$ is subsequently computed. The critical step is to compute $\hat{t}^{-1}$ [Eqn. (22)]. The elements of $\hat{t}^{-1}$ determine the gene ranking: Gene $g$ is ranked more highly than gene $g'$ if $|\hat{t}_{g^1}| > |\hat{t}_{g'}^1|$ in which $\hat{t}_{g^1}$ is the $g$th element of $\hat{t}^1$. The first $G_*$ genes in the re-ranked list are reported as differentially expressed.

2The precise value is $p_0 = 6312/12625$ for the database used which has $G = 12625$ total genes. See the second paragraph of Subsection III-B for an explanation.
Fig. 1. Steps in the gene re-ranking procedure.

**Input:** \( X = \text{labeled } G \times M \text{ gene expression matrix} \)

\( G_* = \text{ Desired size of differential gene list} \)

**Steps:**

1. Calculate two-sample (unpaired) \( t \)-statistics as in Eqn. (2).
2. Re-index genes such that \(|t_1| \leq |t_2| \leq \cdots \leq |t_G|\).
3. Set \( p_0 \) (default value \( \approx 0.5 \)). Set first \( G_0 = Gp_0 \) elements of \( t \) to zero (\( t^0 \) partition).
4. Convert \( X \) to \( \hat{X} \) by subtracting each gene's average response within each treatment group.
5. Compute \( \hat{R} = \text{sample correlation matrix of } \hat{X} \) (Sec. III-D).
6. Find \( \hat{t}^1 \) as in Eqn. (22). (See discussion of product \( \hat{P} t^0 \) in Sec. IV-B.)
7. Find \( \tilde{t}^1 \) as in Eqn. (22). (See discussion of product \( \tilde{P} t^0 \) in Sec. IV-B.)
8. Create a list of re-ranked genes in the descending order of the values \( |\tilde{t}^1_g| \), in which \( \tilde{t}^1_g \) is the \( g^{th} \) element of \( \tilde{t}^1 \).†
9. Report \( G_* \) genes with the largest re-ordered \( |\tilde{t}^1_g| \) 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, \( t \), 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 & 9] is created according to *descending* \( |\tilde{t}^1_g| \) values. This is a more natural ordering for the end result as we are interested in only the \( G_* \) largest values.

---

**B. 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 \( \hat{R} \) to make it invertible. After this augmentation, the algorithm above exhibits excellent numerical stability.

If implemented in a naïve way, the matrix inversion to compute \( \hat{P} = (\hat{R}^{00})^{-1} \) in Eqn. (22) 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 (22), \( \hat{P} t^0 \), by solving the system of simultaneous linear equations \( \hat{R}^{00} s^0 = t^0 \) for the vector \( s^0 = \hat{P} t^0 \) is much faster than explicitly computing the matrix inverse \( \hat{P} \) and forming the product. In particular, we can employ the Cholesky decomposition to exploit the fact that the matrix \( \hat{R}^{00} \) is symmetric and positive definite (e.g., [47, Thm. 4.2.5]). MATLAB implementation uses the built-in function \texttt{linsolve} with appropriate settings, which, in turn, uses the highly-optimized routines of \texttt{LAPACK} (www.netlib.org/lapack/).

The prostate cancer data [48] used in the experiments of Section V 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 ease of use, not computational efficiency. Dedicated, lower-level coding of the re-ranking steps, implemented on a faster machine with more parallelism could reduce the re-ranking time significantly.

**V. Experimental Results**

**A. Technical comparisons**

The re-ranking 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], [49]). SAM adds a small exchangeability factor \( s_0 \) to the pooled sample standard deviation when computing the two-sample \( t \)-statistic:

\[
\hat{t}_g = \frac{\bar{x}_g \bar{x}_g - \tilde{x}_g s_1}{s_0 + s_1};
\]

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 [50], [51], the shrunked \( t/F \)-test [52], the empirical Bayes local FDR [41], and the a posteriori probability approach [53]. It is noteworthy that the re-ranking 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, $\tilde{\pi}$, and the corresponding FDR values, $\tilde{\pi}$, in the reported gene lists. Broadly speaking, the smaller the FDR, better the technique.

B. Results

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. [48]. 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 www-genome.wi.mit.edu/MPR/prostate.

The general purpose of the present experiments is to compare performance of the re-ranking algorithm with the published state-of-the-art methods EDGE and SAM. The software RMAExpress [54] 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 [55], each gene was represented in the final expression matrix $X$ by the log of its expression level.

Comparative algorithm performance and insight into the inner-workings of the re-ranking 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 inter-gene correlation in $X$ 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 $X$ was transformed to $X$ 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 $g$,

$$M_k^{-1} \sum_{m=1}^{M} \hat{x}_{gm} = 0 \quad \text{and} \quad M_k^{-1} \sum_{m=1}^{M} \hat{x}_{gm}^2 = 1. \tag{24}$$

in which $\hat{x}_{gm}$ is the $(g,m)$ element of matrix $X$. Each row represents one gene (and two conditions), so that with this transformation, all genes have equal energy and yet the same within group inter-gene correlation structure as the original $X$.

To generate a test data set from $X$, 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 $G_\delta = G_+ + G_- \tag{25}$.

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,

$$P_\delta \overset{\text{def}}{=} \{p_\delta, G_+, G_-, x_+, x_-\} \tag{26},$$

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,

$$P_c \overset{\text{def}}{=} \{p_0, G, G_*, M_1, M_2\} \tag{27}.$$

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 (www.genomine.org/edge), and samr.r from the SAMR package (www-stat.stanford.edu/~tibs/SAM), respectively. Both routines computed their native gene summary statistics given the matrix $X$ and corresponding column labels. These statistics, in turn, were used to determine the top $G_*$ genes. Results based on re-ranking proceed from the steps outlined in Fig. 1 with $G_*$ values corresponding to the largest $|t_f|$ scores of the re-ranked 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, $p_\delta$, 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.

---

1Normalizing within-group sample mean squares to unity is not implemented in the re-ranking algorithm. The normalization is done 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) inter-gene correlation.

2$G_\delta$ is to be contrasted with $G_*$, the number of genes determined by experimentation to be differentially expressed.
a) Case 1: Small \( p_\delta \): In the first case, \( p_\delta \) is small, \( p_\delta \sim 0.01–0.05 \), meaning that there are relatively few truly-differentially-expressed genes. The smaller \( p_\delta \) 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

\[
\mathcal{P}_1 = \{ p_\delta = 0.025, G_+ = 2G_- = 200, x_\pm = \pm 0.1 \},
\]

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 [48] 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.

Case 1.1: Small \( p_\delta \), small \( p_{s,\delta} \). In the first experiment, the parameter set is

\[
\mathcal{P}_{1.1} = \{ p_0 = 0.5, G=12000, G_\delta = 100, M_1 = 50, M_2 = 52 \}.
\]

The size of the gene-discovery list, \( G_* = 100 \), is significantly smaller than \( G_\delta = 300 \), or \( p_{s,\delta} = G_* / G_\delta = 1/3 \). 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. [48] build a classifier around only 16 of 12625 features, they discuss the need to include as many reliable features as possible.

Figure 2(a) presents results for the test pair \( \mathcal{P}_{1.1,1} \) of Case 1.1. Remarkably, for 36 of 40 \( X \) matrices, the re-ranking procedure reports gene lists with no false discoveries at all, while the other techniques fail to obtain a single gene list with \( \overline{\delta} < 0.5 \). This result is typical of many “small \( p_\delta \)” experiments carried out with an array of parameter sets. In particular, the quality of the results notwithstanding (as measured by \( \overline{\delta} \), see below), the re-ranking strategy uniformly outperformed EDGE and SAM in every scenario. In any rational detection algorithm built around a parameterized stochastic framework, it is possible to find regions of the parameter space in which performance deteriorates. In the “small \( p_\delta \)” gene identification problem, for a fixed \( G \) and \( M \), increasing \( G_* \) (more specifically, increasing the ratio \( p_{s,\delta} \)), or decreasing the “signal” magnitudes of either up- (\( x_+ \)) or down- (\( x_- \)) regulation, all create increasing probabilistic risk of false discoveries, \( \overline{\delta} \). As \( G_* \) was allowed to approach \( G_\delta \) in the Case 1 experiment above, the performance of all methods, EDGE, SAM, and re-ranking, all deteriorated as measured by \( \overline{\delta} \), yet the re-ranking approach remained consistently superior to the others according to this measure. “Better,” however, does not always mean “good.” For illustration, we report a second Case 1 experiment.

Case 1.2: Small \( p_\delta \), large \( p_{s,\delta} \). In this variation of the Case 1 experiment, we take \( G_* = G_\delta = 300 \). This is still a “small \( p_\delta \)” situation, but with a “greedy” approach to gene discovery, an attempt to identify “all” genes (estimated to be) differentially expressed, \( p_{s,\delta} = 1 \). 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 2(b) shows plots of the number of false positives, $\bar{F}$, over 40 data sets for the Case 1.2 experiment. The corresponding FDR, $FDR = \bar{F}/G_+$, 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 list is not an easy task, as evident from the results. Among all methods only re-ranking achieved sufficiently low values of $\bar{F}$ to rescue a few $X$ matrices, but, clearly, even re-ranking would not provide scientifically-useful or reliable gene lists in this high-risk environment.

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

$$P_\delta^2 = \{p_\delta = 0.1, G_+ = G_- = 600, x_\pm = \pm 0.02\}.$$  

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 light of the “increased prior,” the up / down-regulation 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.

Case 2.1: Large $p_\delta$, small $p_{\delta}$. In the first Case 2 test, the experiment parameter set is given by

$$P_\delta^4 = \{p_0 = 0.5, G = 12000, G_e = 300, M_1 = 50, M_2 = 52\}.$$ 

The test pair $(P_\delta^2, P_\delta^4)$ represents a large $p_\delta$ proportion, but relatively small $p_{\delta} = 300/1200 = 0.25$. The experimental results for this case over 40 trials are shown in Fig. 3(a). In spite of the decreased signal strength, the re-ranking procedure produces no false discoveries in a vast majority of trials, similarly to the small $p_\delta$, small $p_{\delta}$. experiment of Case 1.1. EDGE and SAM consistently report a very large proportion of false discoveries (typically 250, or 90%).

Case 2.2: Large $p_\delta$, 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 $(P_\delta^2, P_\delta^4)$ remain identical to those in Case 2.1 experiment, except that $G_+ = 1200$, so $p_{\delta} = 1$. Results are shown in Fig. 3(b). Like the large $p_{\delta}$ experiment in Case 1.2, the re-ranking approach significantly outperforms EDGE and SAM, with typically $\bar{F} \sim 0.5$ for re-ranking and $\bar{F} \sim 0.9$ for the standard methods. The sample variances for $\bar{F}$ and $\bar{F}$ are notably smaller in the Case 2.2 trials relative to similar trials in Case 1.2. Also as in Case 1.2, re-ranking consistently outperforms the standard methods for large $p_\delta$ throughout the range $0 < p_{\delta} \leq 1$. At some application-dependent point in this range, re-ranking, although “better,” is not sufficiently “good” at producing reliable discoveries. Clearly, in the present experiment in which $p_{\delta} = 1$, the re-ranking result of typically $\bar{F} \sim 0.5$ is not indicative of reliable gene discoveries. For a few trials, the rate drops as low as $\bar{F} \sim 0.45$, but even this best-case rate implies $\sim 50$ false discoveries in the reported list of 1200 genes.

c) 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 the Case 2 (large $p_\delta$) 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 the Case 1 values of $x_\pm = \pm 0.1$. The simulation parameters used in Case 3, $P_\delta^3$, are identical to $P_\delta^2$ of Eqn. (30). As in the previous cases, we ran two experiments, the first (Case 3.1) with a “conservative” gene discovery list, $G_+ = 300$ or $p_{\delta} = 0.25$, the second (Case 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 at random from the original prostate cancer expression matrix $X$. The data generation process (including row standardization) detailed in Section V-B1 was then applied to the selected columns. Some compensation for the reduction in the number of samples is is potentially present in the increased differential signal. The $\bar{F}$ and $\bar{F}$ values over 40 trials for the three methods are shown in Fig. 4(a) for Case 3.1, and in Fig. 4(b) for Case 3.2. Even with the significantly reduced sample size, the re-ranking 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 re-ranking procedure increases power in the analysis of small sample data sets.

2) Simulated data: Before devising the test data setup of Section V-B1, the re-ranking 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 $x_m$ the $m^{th}$ column of a simulated expression matrix $X$. We assume that the random vector $x_m$ is multivariate Gaussian with (stationary with index $m$) mean $\mu G_{x,1}$ and covariance matrix $\Lambda$. Each such column represents $G=3226$ genes, which, in their null expressions, are modeled by a covariance matrix $\Lambda$ that introduces roughly the same amount of linear dependence as found in the BRCA data of [56]. We chose simulation parameters

$$P_\delta = \{p_\delta = 0.031, G_+ = G_- = 50, x_\pm = \pm 1\}.$$ 

(32)
DELLER, ET AL., EXPLOITING IDENTIFIABILITY AND INTER-GENE CORRELATION FOR IMPROVED DETECTION OF DIFFERENTIAL EXPRESSION

Fig. 3. Case 2. $\delta$ (left ordinate) and $\overline{\delta}$ (right ordinate) values over 40 trials with larger $p_0$ data. The number of truly differential genes, $G_\delta = 1200$. The "signal" is very weak ($x_+ = \pm 0.02$). (a) Case 2.1 results. $G_* = 300$, small relative to $G_\delta$ ($p_* = 0.25$). The plot shows the number of false positives, $\delta$, over 40 data sets. The corresponding $\overline{\delta} = \delta/G_*$ is shown on the secondary ordinate axis. In spite of the relatively weak signal, re-ranking 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) Case 2.2 results. $\delta$ and $\overline{\delta} = 1200$ ($p_* = 1$).

Fig. 4. $\delta$ and $\overline{\delta}$ values over 40 trials for Case 3 in which $G_\delta = 1200$. (a) Case 3.1, $G_* = 300$ ($p_* = 0.25$); (b) Case 3.2, $G_* = 1200$ ($p_* = 1$).

and the experiments were run with parameters

$$P_e = \{p_0 = 0.5, G = 3226, G_*, M_1 = M_2 = 10\},$$

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

Figures 5(a) and (b) show plots of the $\delta$ and $\overline{\delta}$ values over 40 trials for $G_* = 50$ and $G_* = 100$, respectively. With a smaller $M$, preeminence of the re-ranking method scales down. Nevertheless, for 26 out of 40 simulated $X$ realizations, re-ranking achieves a FDR $\overline{\delta} < 0.15$, and for 30 of 40 trials, $\delta < 0.25$.

Table I shows results for some of the $X$ realizations for $G_* = 100$ [Fig. 5(b)]. Shown are the largest 100 values of $|\tilde{t}_g|$ and each corresponding original $t_g$ with concomitant rank. Even in this challenging case, the results indicate favorable aspects of
the re-ranking procedure. First, it is noteworthy that re-ranking results in $\bar{\delta} = 22$ false discoveries in the list of 100 genes, whereas $\bar{\delta} = 68$ when raw $t$ statistics are used. Further, all but two of the false discoveries reported by re-ranking received an even higher ranking by $t$-statistics. On the other hand, 46 of the correct discoveries by re-ranking 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 $p_{\alpha \delta} \rightarrow 1$, yet, re-ranking persistently outperforms the other methods.

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 $s_\theta$ to adjust the effective pooled variance [31].

VI. 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) inter-gene correlation itself. Analysis in papers by Efron [57], [58] suggests this view of the rich information structure inherent in the data. In this light, re-ranking 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 Figs. 2–5. 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 re-ranking 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 non-differential – 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 $G_\alpha$.

The method was tested on data constructed from the the prostate cancer database of Singh et al. [48] and some simulated data. Compared with SAM [31], EDGE [14], and the raw $t$-statistic approach itself, re-ranking shows substantial improvement in statistical power. As is the case with all published techniques, the re-ranking 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.
TABLE I

Top 100 $t_g$ 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 Fig. 5(b). Data representing false discoveries are printed in boldface.

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

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.

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).