Gene expression

An efficient method to identify differentially expressed genes in microarray experiments

Huaizhen Qin[1], Tao Feng[1,3], Scott A. Harding[2], Chung-Jui Tsai[2] and Shuanglin Zhang[1,3*]

[1]Department of Mathematical Sciences and [2]Biotechnology Research Center, School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI, 49931, USA and [3]Department of Mathematics, Heilongjiang University, Harbin 150080, China

ABSTRACT

Motivation: Microarray experiments typically analyze thousands to tens of thousands of genes from small numbers of biological replicates. Classical Bonferroni and recently developed false discovery rate (FDR) corrections appear conservative when applied to such analyses. The fact that genes are normally expressed in functionally relevant patterns suggests that gene expression data can be stratified and clustered into relatively homogenous groups. Cluster-wise dimensionality reduction should make it feasible to improve screening power while minimizing information loss.

Results: We propose a powerful and computationally simple method for finding differentially expressed genes. The method incorporates a novel stratification-based tight clustering algorithm and principle component analysis. Comprehensive simulations show that our method is substantially more powerful than the standard Benjamini and Hochberg FDR approach, especially for a large number of tests. The mean relative power gain over the standard FDR method approaches 177% when testing 10000 genes. We applied the method to four real microarray datasets: one from a *Populus* nitrogen deficiency experiment with 3 biological replicates; and three from public microarray datasets of human cancers with 10 to 40 biological replicates. In all four analyses, our method proved more robust than the standard FDR method for identification of differentially expressed genes.

Availability: The C++ code to implement the proposed method is available upon request for academic use.

Contact: shuzhang@mtu.edu

Supplementary information: Excel datasets of the tables and diagrams in this article can be viewed at http://www.math.mtu.edu/~shuzhang/

1 INTRODUCTION

Analysis of high-throughput microarray data is becoming commonplace with the increase of sequenced genomes and genome-wide investigations of gene expression (Brem *et al*., 2002; Yvert *et al*., 2003; Morley *et al*., 2004; Chesler *et al*., 2005; Hubner *et al*., 2005; Mehrabian *et al*., 2005; Scheetz *et al*., 2006; Tsai *et al*., 2006). Low-replication experiments are common in microarray studies (Lee *et al*., 1999; Gadbury *et al*., 2003) and testing for differential expression of many genes with small samples is problematic (Sima and Dougherty, 2006; Yang and Churchill, 2006). One central challenge is to identify, reliably and economically, as many biologically and statistically significant genes as possible while controlling false positives.

Research in multiple testing has passed several milestones. Naive application of standard hypothesis tests with no adjustment for multiplicity results in a large number of reproducible false discoveries (Manly *et al*., 2004). One solution is to control family-wise type-I error rate by the Bonferroni correction (Holm, 1979) or the step-up sequential Bonferroni correction (Hochberg, 1988), which use stringent significance criteria to prevent false positives. The tradeoff is substantially reduced power for detecting false null hypotheses. Benjamini and Hochberg (1995) pioneered BH95 (a step-up approach) for independent tests that controls FDR, the portion of significant results that are erroneous. Numerous theoretical studies and practical applications of FDR control have followed (e.g., Benjamini and Yekutieli, 2001), and more recently, FDR control has been widely adopted for identification of differentially expressed genes in microarray experiments (Reiner *et al*., 2003; Li *et al*., 2005; Pawitan, 2005; Pounds and Cheng, 2005). However, the inherent multiplicity and complex dependence structures of such experiments are a challenge for massive multiple hypothesis tests. As commented by Verhoeven *et al*. (2005), BH95 can be conservative in that it controls FDR below a nominal level no matter how many null hypotheses are true. For comprehensive reviews on significance measures and approaches of multiple testing in microarray experiments, see Dudoit (2003) and Pounds (2006).

It is not surprising that a procedure loses power as the number of tests increases. We believe that suitable dimension reduction techniques based on clustering can be applied to effectively reduce the number of tests, thereby conserving testing power. Standard clustering analysis forces all data points into groups at the expense of cluster tightness. For microarray experiments, Tseng and Wong (2005) have proposed a method to identify informative, tight, and stable clusters to enable statistically valid biological inferences from microarray data. By integrating K-means clustering with resampling, this "tight-clustering" method embodies a novel concept that does not necessitate the estimation of the number of clusters and the assignment of all genes into clusters. Although promising, it is computationally intensive and requires large sample sizes like other resampling-based techniques.

* To whom correspondence should be addressed at Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, 49931, USA

In this article, we propose an efficient method to identify differentially expressed genes. We term the method FCPC, since it is based on forward search using gene-to-gene correlation and principal component analyses. In FCPC, we first divide genes into co-expression strata using the information conveyed by gene expression. This is analogous to the post-stratification technique commonly used in large scale survey sampling (Cochran, 1977; Holt and Smith, 1979; Feng and Shi, 1996) to improve inference precision. Second, we design a tight clustering method to search each stratum for tight gene clusters in each of which the minimum gene-to-gene correlation exceeds a pre-determined level. The proposed tight clustering approach should be especially suitable for low-replicate experiments. Next, we represent tight clusters by their first principal components (PCs). In terms of mean-square error, principal component analysis is a suitable linear dimension reduction technique for defining a new dimensional space that captures the maximum information in the original dataset. Finally, we screen for significant differential expression among PCs and scattered genes, simultaneously controlling FDR. Because the PCs and scattered genes are largely uncorrelated, BH95 applies. If a PC is found significant, all genes in the cluster are declared to be significant. Simulations show that FCPC controls FDR and is more powerful than BH95. Applications to real microarray datasets also show that our method yields more noteworthy candidate genes for follow-up studies than BH95 does.

2 METHODS

Briefly, FCPC is composed of four steps: creating co-expression strata, finding tight clusters, assigning a representative value for each tight cluster by principal component analysis, and identifying differentially expressed clusters and/or scattered genes. Details of the four steps are given below.

2.1 Generation of co-expression strata

All genes (probe sets) are initially divided into two strata based on mean expression differences between control and treatment: $d_i = \bar{x}_{1i} - \bar{x}_{0i}$ for $i = 1,...,M_t$, where $\bar{x}_{0i}$ and $\bar{x}_{1i}$ are mean expression indices of the i^{th} gene in the control and the treatment, respectively, and M_t is the number of all genes under consideration. We call $\mathbb{S}_+ = \{i:\ d_i > 0\}$ and $\mathbb{S}_- = \{i:\ d_i < 0\}$ the up-regulated and down-regulated strata, respectively, and we screen for up- and down-regulated genes from $\mathbb{S}_+$ and $\mathbb{S}_-$. We further stratify $\mathbb{S}_+$ using relative expression ratios $r_i = \bar{x}_{1i}/\bar{x}_{0i}$ for $i \in \mathbb{S}_+$. For an integer $k \geq 1$, we assign all of the genes with relative expression ratios in $[k-0.5, k+0.5)$ into substratum k. Similarly, we further stratify $\mathbb{S}_-$ using $r_i = \bar{x}_{0i}/\bar{x}_{1i}$ for $i \in \mathbb{S}_-$.

2.2 Identification of tight clusters

If the smallest gene-to-gene correlation within a cluster is chosen to exceed ρ_0 (=0.8), we identify the cluster as having a tightness ρ_0. We search for tight clusters separately within each stratum. For a given tightness ρ_0, we find tight clusters recursively by a four-step algorithm:

(1) Find a cluster core. Search for the gene pair $C = (i, j)$ with the largest sample correlation $\rho_{\max} = \rho_{ij}$ in the stratum. If $\rho_{\max} \geq \rho_0$, then take C as the core of a potential cluster. Otherwise, go to step (4).

(2) Extend the core to a cluster. For gene $g \notin C$, if $\min\{\rho_{gi}: i \in C\} > \rho_0$, then add the gene to C and denote as the current cluster; otherwise, search for the next gene $\notin C$. Repeat this step until no additional genes can be added. Retain current C as a tight cluster and go to the next step.

(3) Remove the tight cluster from the stratum and repeat steps (1) and (2) for the remaining genes in the stratum to find another cluster of tightness ρ_0. Repeat this step until no additional clusters of the same tightness can be found.

(4) Reduce the value of ρ_0 and repeat steps (1) to (3) until ρ_0 falls to a pre-determined value. In our simulation studies and real database analyses, we begin with $\rho_0 = 0.8$, and then reduce it to 0.7, 0.6, and finally 0.5. See Sections 3.3 and 5 for the rationale of choosing these values.

2.3 Principal component analysis

For a tight cluster of size $m \geq 2$, denote by $x^{(j)} = (x_{1j},...,x_{mj})^\tau$ the expression indices of m genes in the j^{th} biological individual. Calculate $\Sigma = \sum_{j=1}^{n} (x^{(j)} - \bar{x})(x^{(j)} - \bar{x})^\tau$, where $n = n_0 + n_1$, n_0 is the sample size of controls, n_1 is the sample size of treatments, and $\bar{x} = n^{-1}\sum_{j=1}^{n} x^{(j)}$. All positive eigenvalues of Σ are denoted by $\lambda_1 \geq \cdots \geq \lambda_p$. The first PC of the j^{th} biological individual is given by $x_j^* = e_1^\tau x^{(j)}$, where e_1 is the eigenvector associating with λ_1. We propose the use of $x^* = (x_1^*, \cdots, x_n^*)^\tau$ to represent this tight cluster. How well the first PC represents this cluster can be measured by the ratio $\gamma = \lambda_1 \Big/ \sum_{k=1}^{p} \lambda_k$, the proportion of total variance explained by the first PC. Fig. S1 illustrates the representativeness of the first PC.

2.4 Identification of differentially expressed genes

Suppose there are M_1 tight clusters and M_2 scattered genes. We calculate the p-value of the two-sample t-test for each of the M_1 tight clusters by using its first PC and, the p-value of the two-sample t-test for each of the M_2 scattered gene by using its gene expression index. We sort the $\bar{M} = M_1 + M_2$ p-values as $p_{(1)} < \cdots < p_{(\bar{M})}$. Denote $\delta = \max\{p_{(i)}: p_{(i)} \leq i\alpha/\bar{M}\}$ for a preset rate α. All clusters and scattered genes with p-values smaller than δ are significant. To find δ, we start at $p_{(\bar{M})}$, proceed to smaller p-values as long as $p_{(i)} > i\alpha/\bar{M}$, and stop the procedure as $p_{(i)} \leq i\alpha/\bar{M}$ with $\delta = p_{(i)}$. All significant scattered genes and the genes in the significant clusters are extracted as differentially expressed candidate genes.

2.5 Method comparison

We compare the proposed method with the BH95. Denote by $\tilde{p}_i$ the p-value of the two-sample t test of the i^{th} gene, $i = 1,...,M$, where M is the number of genes in the up-regulated stratum. Denote by $\tilde{p}_{(1)} < \tilde{p}_{(2)} < \cdots < \tilde{p}_{(M)}$ the sorted p-values of all these t tests. Denote $\tilde{\delta} = \max\{\tilde{p}_{(i)} : \tilde{p}_{(i)} \le i\alpha/M\}$ for the preset rate α. The BH95 identifies all genes with p-values smaller than $\tilde{\delta}$.

3 SIMULATION STUDIES

3.1 Simulation design

We conducted simulations in a two-condition experiment with 3 controls and 3 treatments to compare FCPC with BH95 approach. The expression indices are generated in numerous independent blocks. The first four blocks contain up-regulated genes and are called up-regulated blocks, and the other blocks contain stably expressed genes and are referred to as stable blocks. The sizes of the four up-regulated blocks are fixed at 10, 40, 40 and 10 for a total of $M_d = 100$ up-regulated genes. The size of each stable block is a random number between $\underline{m}$ and M_e, where $\underline{m}$ is a preset minimum size for all simulation replications, and M_e is the number of all stably expressed genes. For one block of size m_b, we generate the gene expression indices of controls and treatments from $N(\mu_c 1_b, \Re_b)$ and $N(\mu_t 1_b, \Re_b)$ respectively, where $N(\mu_c 1_b, \Re_b)$ stands for the multivariate normal distribution with mean $\mu_c 1_b$ and a variance-covariance matrix $\Re_b = (1-\rho_b)I_b + \rho_b 1_b 1_b'$, I_b is the identity matrix of order m_b, $1_b = (1,...,1)'$, ρ_b is a random number between $\underline{\rho}$ and 1, and $\underline{\rho}$ is a preset minimum correlation across all blocks. Across all stable blocks $\mu_c = \mu_t = 5$, and across the four up-regulated blocks, $\mu_c = 5$, and $\mu_t = 15$, 12, 10, and 8, respectively.

3.2 Simulation results

To compare FCPC with BH95, we adopt FDR and power as performance criteria. As conventionally defined in the literature, we define $\text{FDR} = 1 - R^{-1}\sum_{r=1}^{R} I(M_r > 0) M_{dr}/M_r$ for a given method, where M_{dr} is the number of true discoveries and M_r is the number of all discoveries by the method at the r^{th} simulation replication, and R is the number of all simulation replications. We define $\text{power} = (RM_d)^{-1}\sum_{r=1}^{R} M_{dr}$ as the average probability of rejecting the false null hypotheses, referred to as the average power in Dudoit *et al.* (2003).

We investigated 16 scenarios described by $\underline{m} \in \{1,20,50,100\}$ and $\underline{\rho} \in \{0.25, 0.5, 0.75, 0.9\}$. Under each scenario, we set nominal FDR at 0.05, and set $R = 1000$ to evaluate the powers and true FDRs for different numbers of tests M_t. Table S1 shows the corresponding FDRs as M_t increases from 1000 to 10000. The standard BH95 is uniformly conservative in all scenarios, regardless of gene number. By contrast, FCPC appears to be conservative when M_t is small, and the conservativeness declines as M_t increases. Table S2 shows BH95 and FCPC power across the 16 scenarios. As shown by two representative scenarios in Fig. 1, BH95 power consistently declines from about 0.8 to near 0.3, whereas FCPC power is sustained. The gain of FCPC to standard BH95 becomes more pronounced as M_t increases. We calculated the relative gain of FCPC to standard BH95 as $R_G(i|M_t) = G_{\text{FCPC}}(i|M_t)/G_{\text{BH95}}(i|M_t) - 1$, where $G_{\text{FCPC}}(i|M_t) = \text{Power}_{\text{FCPC}}(i|M_t) - \text{FDR}_{\text{FCPC}}(i|M_t)$ for each given value of M_t in scenario $i \in \{1,...,16\}$. We then calculated the mean relative gain across all 16 scenarios $\bar{R}_G(M_t) = \frac{1}{16}\sum_{i=1}^{16} R_G(i|M_t)$ and the corresponding coefficient of variance for each given value of

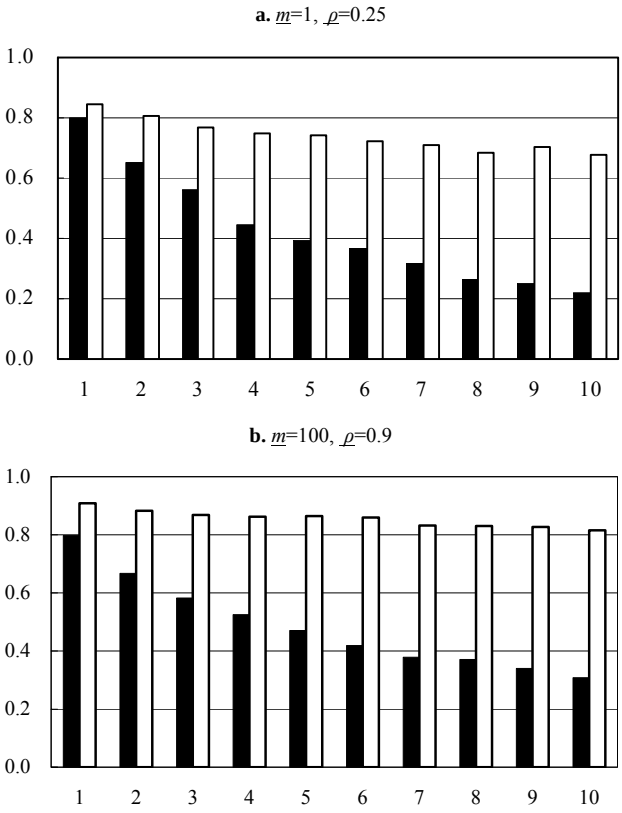

Fig. 1. Powers of BH95 (black bars) and FCPC (white bars) under two scenarios described by ($\underline{m}, \underline{\rho}$). Horizontal axis: Number of genes M_t (in thousands). Vertical axis: Power.

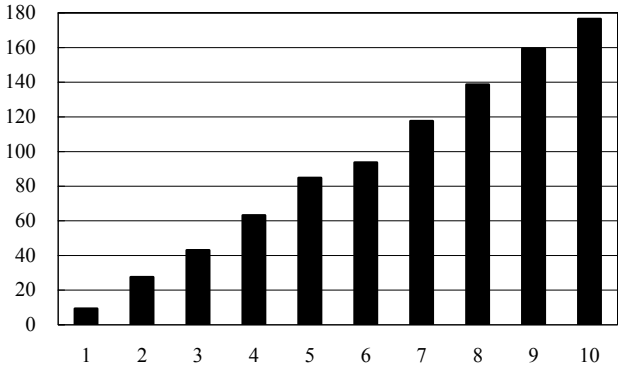

Fig. 2. Mean relative gain of FCPC to BH95 across 16 scenarios. Horizontal axis: Number of genes M_t (in thousands). Vertical axis: Mean relative gain $\bar{R}_G$ (in %). As M_t increases from 1000 to 10000, the mean relative gain increases from 9% to 177%.

M_t. As shown in Fig. 2, $\bar{R}_G(M_t)$ increases linearly from 9% to 177% as M_t increases from 1000 to 10000. The coefficient of variance of the relative gains corresponding to each value of M_t is about 0.1 except for M_t=1000, where the coefficient of variance is 0.3. Such small coefficients of variance indicate that the linear trend of mean relative gains is steady.

Although BH95 was originally developed for independent tests, Benjamini and Yekutieli (2001) showed that it also controls FDR for tests exhibiting some types of positive dependence. They proposed a modified FDR approach, referred to as BY01, to handle data with other forms of dependency. In our simulation study; however, BY01 performed rather conservatively and was much less powerful than BH95 under the dependency structure described in Section 3.1 (data not shown). Our results support the comments of Reiner *et al*. (2003) and Verhoeven *et al*. (2005) that BY01 may be too conservative for microarray experiments.

The representativeness of the first PC for each tight cluster was evaluated according to $\gamma = \lambda_1 \Big/ \sum_{k=1}^{p} \lambda_k$. Table 1 shows the distribution and certain mathematical characteristics of the γ-values of 104332 tight clusters produced by 1000 simulation replications under a scenario described in Section 3.1, where $\underline{m}$ =1, $\underline{\rho}$ =0.5, and $M_t = 5000$. There are 100221 clusters of tightness ≥ 0.8, and 2386 clusters of tightness 0.7 to 0.8, and so on. The number of clusters decreases rapidly as the level of tightness declines. For the tight clusters at each level, the median and mean are large, and the coefficient of variance is very small. In general, the first PC of a tight cluster represents the cluster well. On average, the first PC explained 82% of the total variance (Table 1, see the smallest mean and median). Additional details for first PCs of tightness ≥ 0.8 are shown in Fig. S1.

Table 1. The representativeness of the first PC

ρ_0	n_c	$\gamma_{\min}$	$\gamma_{\max}$	Mean	Median	cv
0.8	100221	0.8741	0.9999	0.9321	0.9290	0.0207
0.7	2386	0.8139	0.9927	0.8984	0.8953	0.0315
0.6	1113	0.7567	0.9967	0.8619	0.8506	0.0463
0.5	612	0.6901	0.9887	0.8209	0.8079	0.0618

ρ_0 : level of tightness; n_c : number of clusters; $\gamma_{\min}$: minimum of observed γ values; $\gamma_{\max}$: maximum of observed γ values; cv : coefficient of variance.

3.3 The basis for the forward search tight clustering

The correlations between genes of different expression patterns have distinct properties. Let $\hat{\rho}$ be a generic notation of the correlation between two genes. We distinguish three possibilities of $\hat{\rho}$: i) between two differentially expressed genes, ii) between a differentially expressed gene and a stably expressed gene, and iii) between two stably expressed genes. The proposed tight clustering is based on our simulation studies on gene-to-gene correlations.

In our method the probability $\Pr(\hat{\rho} > \rho_0)$ reflects the likelihood of assigning two genes of interest into a cluster of tightness ρ_0. To illustrate the distribution properties of $\hat{\rho}$, we sampled the expres sion indices of two genes exhibiting bivariate normal distribution in a control group with mean zero, variance 1 and correlation ρ, and bivariate normal distribution in a treatment group with mean $(\mu,\nu)'$, variance 1 and correlation ρ. For given differential ex-

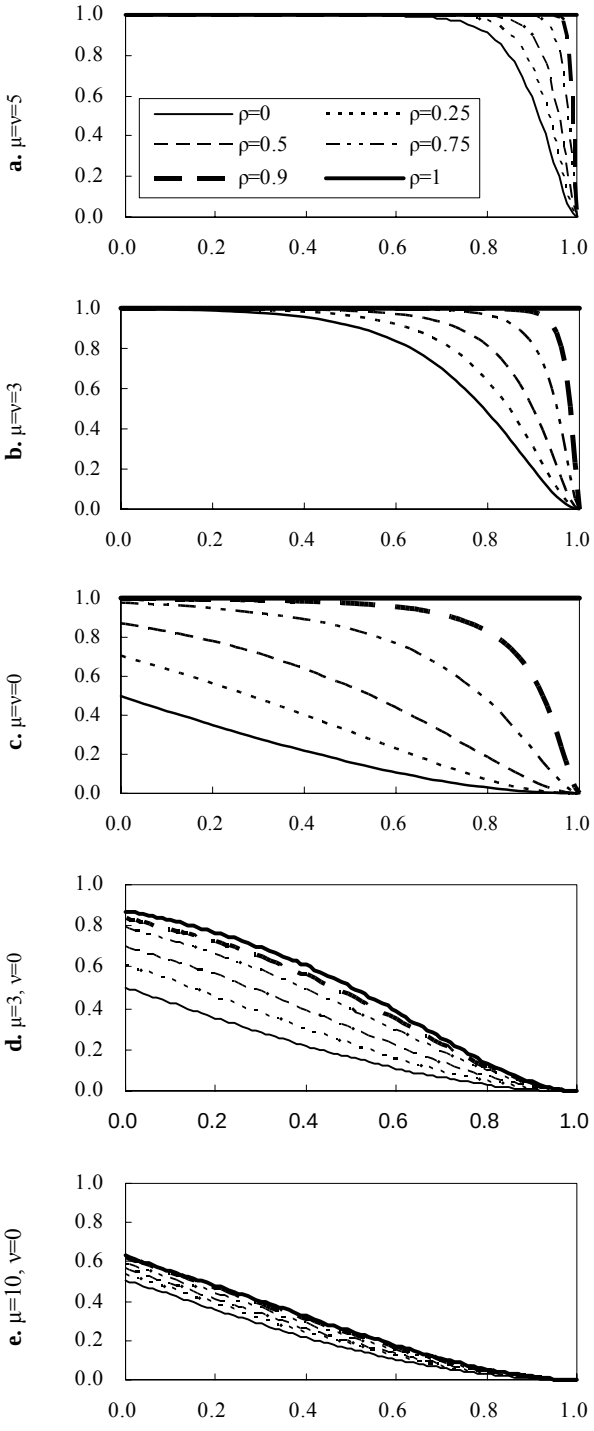

Fig. 3. Each curve is based on a gene pair expressed in 3 controls and 3 treatments. For a control, the expression of the pair is sampled from the bivariate normal distribution with mean 0, variance 1, and correlation ρ, and for a treatment, the expression of the pair is sampled from the bivariate normal distribution with mean $(\mu,\nu)'$, variance 1, and correlation ρ. Horizontal axis: ρ_0 . Vertical axis: $\Pr(\hat{\rho} > \rho_0)$, where $\hat{\rho}$ is gene-to-gene correlation

pression $(\mu,\nu)'$ the proba-bility $\Pr(\hat{\rho} > \rho_0)$ increased with population correlation ρ (Fig. 3 and Fig. S2). This is consistent with classical sample correlation. These figures also show three particular characteristics of gene-to-gene correlations, which clearly differ from classical sample correlations.

First, the distribution of the correlation between two similarly up- or down-regulated genes (i.e., $\mu = \nu \neq 0$) can be dramatically affected by the magnitude of differential μ. The larger the magnitude, the larger the probability $\Pr(\hat{\rho} > \rho_0)$, and the more likely the two genes will be clustered together. As shown in Fig. 3a-b and Fig. S2a-b, the correlation between two differentially expressed genes is likely large if the magnitude of the differential is large. Surprisingly, this is also true for two independent genes ($\rho = 0$). As $\mu = 5$ and $\rho = 0$, $\Pr(\hat{\rho} > 0.8)$=0.9091 in Fig. 3a and = 0.9585 in Fig. S2a. The probability decreases as μ decreases. As μ =3 and $\rho = 0$, $\Pr(\hat{\rho} > 0.8)$= 0.4798 in Fig. 3b and =0.0072 in Fig. S2b. Given $\rho = 0$, $\Pr(\hat{\rho} > 0.8)$ decreases to 0.032 in Fig. 3c and 0 in Fig. S2c when μ decreases to 0. Since high gene-to-gene correlations likely occur between similarly up- or down-regulated genes with large differentials, the proposed method tends to assign those genes with the largest differentials to a common cluster in early iteration steps.

Second, the correlation between a stably expressed gene and a differentially expressed gene is unlikely to be large. This is especially the case when the magnitude of differential is large, even when the population correlation ρ is close to 1. As μ =3, $\Pr(\hat{\rho} > 0.8) \le 0.1354$ in Fig. 3d and $\Pr(\hat{\rho} > 0.8) \le 0.0001$ in Fig. S2d for $\rho \in [0,1]$. As $\mu = 10$, $\Pr(\hat{\rho} > 0.8) \le 0.0518$ in Fig. 3e and $\Pr(\hat{\rho} > 0.8)$=0 in Fig. S2e for $\rho \in [0,1]$. These upper bounds are achieved at $\rho = 1$, and for fixed μ the probability $\Pr(\hat{\rho} > \rho_0)$ declines as ρ decreases. Thus, the proposed forward-search tight clustering method can reduce the chance of assigning a stable gene to a differentially expressed cluster, and vice versa.

Third, Fig. 3c-e and Fig. S2c-e show two properties of the corre lation $\hat{\rho}$ between a stably expressed gene and a differentially expressed gene: i) The distribution of $\hat{\rho}$ is invariant to μ if the two genes are independent ($\rho = 0$). Precisely, independence means $\tau = \sqrt{n-2}\hat{\rho}/\sqrt{1-\hat{\rho}^2} \sim t_{n-2}$ (the student t with $n-2$ degrees of freedom) for arbitrary μ. ii) For arbitrary $\rho \in [0,1]$, τ converges in distribution to t_{n-2} as $\mu \to +\infty$. As μ increases, the curves with respect to positive population correlations decline toward the benchmark curve of an independent stably expressed pair. By these properties, one may control the possibility of clustering a gene of fixed differential expression with a stably expressed gene by choosing a suitable tightness correlation threshold.

Classical correlation theory applies for the correlation $\hat{\rho}$ between two stably expressed genes ($\mu = \nu = 0$). In such a case, $\hat{\rho}$ is well-known to converge in probability to the population correlation ρ as sample size increases. The distribution of $\hat{\rho}$ is mainly affected by ρ for a finite sample size. The larger the population correlation, the more likely $\hat{\rho}$ is to be larger than a given threshold, see Fig. 3c and Fig. S2c. In microarray experiments, there are more stably expressed genes than differentially expressed genes. Hence, the forward-search tight clustering is especially efficient as there are large population correlations among stably expressed genes.

4 APPLICATIONS TO REAL MICROARRAY DATASETS

4.1 Nitrogen deficiency in Populus

We applied FCPC to analyze the transcriptomic response of *Populus* fremontii × angustifolia to nitrogen deficiency using the GeneChip® Poplar Genome Array (Affymetrix). Raw hybridization signals were processed by the Affymetrix MAS 5.0 software, and only probe-sets identified as "present" in all 3 control and 3 nitrogen stress replicates were analyzed further. The resultant 13507 probe-sets were separated into up- (7228) and down-regulated (6279) strata, with their substrata and associated parameters summarized in Table 2. Of the 7228 probe-sets in the up-regulated stratum, 7218 were represented by 374 tight clusters, and likewise, 6266 of the 6279 probe-sets in the down-regulated stratum were covered in 343 tight clusters. To control the overall FDR at $\alpha = 0.05$, we allocated half α (i.e., 0.025) each to the up- and down-strata to identify differentially expressed genes by the BH95 correction. The FCPC method detected 435 significantly up- and 12 significantly down-regulated genes. However, using BH95 at the same α level, we only identified 204 up- and 14 down-regulated genes, altogether 229 less than the FCPC.

Table 2. Distribution of substrata, tight clusters and probe-set numbers within the up- and down-regulated strata in the nitrogen stress experiment of *Populus*

Up-regulated stratum				Down-regulated stratum			
RR	SS	NC	NGC	RR	SS	NC	NGC
17	1			56	1		
11	1			23	1		
10	4	1	4	21	1		
9	6	2	6	19	1		
8	6	1	5	18	2		
7	15	2	14	17	2		
6	16	2	15	15	1		
5	58	4	57	14	5	1	5
4	211	14	211	13	3	1	3
3	1961	71	1959	12	2	1	2
2	4949	277	4947	11	3	1	3
				10	4	1	3
				9	3	1	3
				8	2	1	2
				7	8	1	7
				6	18	2	17
				5	37	4	37
				4	57	5	57
				3	204	14	204
				2	1563	61	1562
				1	4361	249	4361
Total	7228	374	7218	Total	6279	343	6266

RR: Relative expression ratio; SS: Substratum size; NC: Number of clusters; NGC: Number of probe-sets in clusters.

Further examination of the significant results from each procedure reveals discrepancies in terms of the genes identified. Table 3 lists the 10 discoveries based on the highest expression differentials and the 10 discoveries based on the smallest expression differentials found by FCPC and BH95. Half of the discoveries by FCPC from the up-regulated stratum were not captured by BH95 (see the genes and relative expression ratios in bold face). FCPC

outperformed BH95 in capturing strongly up-regulated candidate genes (Table 3a), as well as those were weakly but statistically significantly up-regulated (Table 3b). For instance, 9 of 10 weakly up-regulated genes discovered by FCPC were missed by BH95. All 10 FCPC discoveries shown in Table 3b were less than 1.3-fold up-regulated, versus only 2 by BH95. Although biological significance of these weakly up-regulated candidate genes requires follow-up analysis, FCPC nevertheless provides a more sensitive means than BH95 in capturing more candidate genes for subsequent investigations.

Table 3. Partial lists of up-regulated candidate genes in nitrogen-stressed *Populus*

a. The top 10 significant discoveries with the largest relative expression ratios

BH95		FCPC	
Gene name	RR	Gene names	RR
Ptp.459.1.S1_s_at	15.95	Ptp.459.1.S1_s_at	15.95
PtpAffx.24885.1.A1_a_at	9.20	**PtpAffx.113871.1.A1_at**	**9.27**
PtpAffx.27718.1.S1_s_at	8.77	PtpAffx.24885.1.A1_a_at	9.20
PtpAffx.6111.2.S1_a_at	7.92	**Ptp.3642.1.A1_at**	**8.87**
Ptp.3539.1.S1_x_at	7.92	PtpAffx.27718.1.S1_s_at	8.77
PtpAffx.74725.1.S1_at	7.51	PtpAffx.6111.2.S1_a_at	7.92
PtpAffx.101017.1.A1_s_at	7.43	Ptp.3539.1.S1_x_at	7.92
PtpAffx.618.1.S1_x_at	7.12	**PtpAffx.40333.1.S1_at**	**7.78**
Ptp.3539.1.S1_at	6.99	PtpAffx.74725.1.S1_at	7.51
Ptp.1516.3.S1_s_at	6.33	PtpAffx.101017.1.A1_s_at	7.43

b. The last 10 significant discoveries with the smallest relative expression ratios

BH95		FCPC	
Gene name	RR	Gene name	RR
Ptp.7283.1.S1_s_at	1.39	**PtpAffx.32381.1.S1_s_at**	1.29
PtpAffx.19580.1.S1_at	1.39	**PtpAffx.152585.1.S1_at**	1.29
Ptp.2604.1.S1_x_at	1.38	**PtpAffx.12414.1.S1_at**	1.28
PtpAffx.207565.1.S1_at	1.37	**PtpAffx.87139.1.A1_at**	1.28
Ptp.7637.1.A1_at	1.37	**Ptp.513.1.S1_at**	1.26
PtpAffx.216289.1.S1_at	1.33	**PtpAffx.85691.1.S1_s_at**	1.25
PtpAffx.10425.1.S1_at	1.32	**Ptp.4891.2.A1_at**	1.25
PtpAffx.6384.1.A1_s_at	1.31	**PtpAffx.2360.3.S1_at**	1.25
Ptp.4961.1.S1_at	1.23	**PtpAffx.5275.1.A1_at**	1.20
PtpAffx.206462.1.S1_at	1.19	PtpAffx.206462.1.S1_at	1.19

RR: Relative expression ratio.

4.2 Human diseases

Having applied the FCPC method to the low-replicate plant microarray experiment, we then turned to investigate its performance in microarray datasets of human cancers (breast cancer, colon cancer, and leukemia), in which there were more biological replicates. The original breast cancer dataset is from van't Veer *et al*. (2002), based on the Agilent Hu25K oligo array platform. We use the files ArrayData_less_than_5yr.zip and Array_Data_greater_than_5yr.zip, which correspond to 34 patients that developed metastases within 5 years and 44 individuals that remained disease-free for over 5 years, respectively. As the authors did, we selected only the genes that were "significantly regulated" (see their definition in the paper and supplemental material), which resulted in a total of 4869 clones. We excluded the 10^{th} individual from the "diseased" dataset (sample 54, IRI000045837, in the original data files), because it had over 44% missing values out of the entire 24481 clones. The colon cancer dataset is from Alon *et al*. (1999). In that dataset, expression indices of 40 tumor and 22 normal colon tissues for 6600 human genes were measured using the Affymetrix GeneChip. A subset of 2000 genes with the highest minimal signal intensity across the samples was chosen by the authors for further analysis. The leukemia dataset is from Golub *et al*. (1999). We used the data of 11 AML and 27 ALL from the original paper. This dataset contains expression indices of 6817 genes, and 3051 genes remained after filtering and preprocessing as done by the authors. The basic features of the three data sets are summarized in Table 4a.

Table 4. Basic features and significant gene discoveries of the three human cancer datasets

a.

	Numbers of		
Cancers	filtered genes	controls	Treatments
Breast	4869	44	33
Colon	2000	22	40
Leukemia	3051	11	27

b.

	BH95			FCPC		
Cancers	Up	Down	Total	Up	Down	Total
Breast	66	2	68	127	2	129
Colon	201	129	340	234	166	400
Leukemia	362	324	686	410	372	782

FCPC detected significantly higher numbers of differentially expressed genes than BH95 in all three datasets, as shown in Table 4b. Using the leukemia dataset as an example, BH95 discovered 362 significantly up-regulated genes and 324 significantly down-regulated genes, while FCPC discovered 410 and 372, respectively. The numbers of significant discoveries from these three real datasets can be validated in a relevant study by Meinshausen and Bühlmann (2005), who estimated the lower bounds of the numbers of differentially expressed genes in these datasets. The number of significant genes discovered by FCPC was very close to or greater than the lower bounds given by these authors. For the breast cancer, colon cancer and leukemia, the estimated lower bounds at $\alpha = 0.01$ were 126, 245 and 811 respectively, while FCPC identified 129, 400 and 782 differentially expressed candidate genes respectively. In contrast, the number of significant genes detected by BH95 was smaller than the estimated lower bound in 2 of the 3 datasets.

5 DISCUSSION

In this article, we present a powerful and computationally simple method, FCPC, to detect differentially expressed genes from microarray data. The method integrates the strengths of stratification, tight clustering, data compression, and standard Benjamini-Hochberg FDR correction. We evaluated FCPC by simulation studies as well as by application to real datasets. Simulation results showed that FCPC controls FDR and is much more powerful than the popular FDR correction when the number of genes is large. The basis for the FCPC approach is two-fold. First, expression indices that vary between different experimental conditions can reveal certain regulatory strata. Genes within one common stratum may be more closely related functionally, at the organismal level, than genes from different strata. This serves as the basis for post stratification. Second, many expression indices within one com-

mon stratum are strongly correlated. This serves as the basis for correlation-based clustering.

The rationality of the iterative clustering method we employed deserves special mention. Clustering was done progressively, and with a correlation threshold in order to maximize the tightness of early-formed clusters. We designed the method according to our observations on the sampling distributions of gene-to-gene correlation. A correlation threshold can be chosen such that the sample correlation between similarly up- or down-regulated genes will most likely exceed that threshold, while the sample correlation between a stably expressed gene and a differentially expressed gene is unlikely to meet the threshold. Therefore, the proposed clustering method can distinguish differentially expressed genes from stably expressed genes during the early iterations, and organize them into tight clusters. In addition, the correlation between two stably expressed genes is likely larger than a threshold if the population correlation is strong. This integration proved to efficiently prevent loss of statistical power and the flood of FDR.

The observations in Section 3.3 are helpful for defining the tightness levels for tight clustering to find stable clusters. It is very difficult to find the optimal set of tightness levels without information about the population correlation ρ. According to our simulations and observations of the properties of gene-to-gene correlation, {0.8, 0.7, 0.6, 0.5} is a reasonable choice. Analysis of real data as in Section 4 appeared to validate this estimate, as use of these tightness levels resulted in the clustering of nearly all genes.

FCPC appears conservative in analysis where the number of genes is small, even more conservative than the BH95. One may improve testing power by creating larger and/or more clusters using a set of more flexible tightness levels.

ACKNOWLEDGEMENTS

This work was supported by NSF Plant Genome Project DBI-0421756, NIH grants R01 GM069940, R03 HG003613, and R01 HG003054. We thank Guohua Zou, Lindsey Tuominen, Michael Senkow, Christy Oslund and Jill Olson for their helpful comments on the manuscript.

REFERENCES

Alon, U. *et al*. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96**, 6745–6750.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29** (4), 1165–1188.

Brem, R.B. *et al*. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

Chesler, E.J. *et al*. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, **37** (3), 233–242.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd ed. New York, John Wiley.

Dudoit, S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Science*, **18** (1), 71–103.

Feng S. and Shi X. (1996) *Survey sampling—Theory, Methods and Practice.* Shanghai Technology Press.

Gadbury, G.L. *et al*. (2003) Randomization tests for small samples: an application for genetic expression data. *Appl. Statist.*, **52**, Part 3, 365–376.

Golub, T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.

Holt D. and Smith, T.M.F. (1979) Post stratification. *J. R. Stat. Ser. A*, **142** (1), 33–46.

Hubner, N. *et al*. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, **37**, 243–253.

Li, S.S. *et al*. (2005) FDR-controlling testing procedures and sample size determination for microarrays. *Stat. Med.*, **24**, 2267-2280.

Manly, K.F. *et al*. (2004) Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.*, **14**, 997–1001.

Mehrabian, M. *et al*. (2005) Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nature Genetics*, **37**, 1224–1233.

Meinshausen, N. and Bühlmann, P. (2005) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, **92** (4), 893–907.

Morley M. *et al*. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.

Pawitan, Y. *et al*. (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21** (13), 3017–3024.

Pounds, S.B. (2006) Estimation and control of multiple testing error rates for microarray studies. *Brief. Bioinformatics*, **7** (1), 25–36.

Pounds, S. and Cheng, C. (2005) Sample size determination for the false discovery rate. *Bioinformatics*, **21** (23), 4263–4271.

Reiner A. *et al*. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19** (3), 368–375.

Scheetz, T.E. *et al*.(2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci.*, **103** (39), 14429–14434.

Sima, C. and Dougherty, E.R. (2006) What should be expected from feature selection in small-sample settings. *Bioinformatics*, **22**, 2430–2436.

Tsai C.-J. *et al*. (2006) Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in Populus. *New Phytologist*, **172**, 47–62.

Tseng G.C. and Wong W.H. (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

van't Veer L. J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Verhoeven, K.J.F. *et al*. (2005) Implementing false discovery rate control: increasing your power. *OIKOS*, **108**, 643–647.

Yang, H. and Churchill, G. (2007) Estimating *p*-values in small microarray experiments. *Bioinformatics*, **23** (1), 38–43.

Yvert, G. *et al*. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, **35**, 57 – 64.