BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3146–3155
doi:10.1093/bioinformatics/bth379

A statistical method for identifying differential gene–gene co-expression patterns

Yinglei Lai[1], Baolin Wu[1], Liang Chen[3] and Hongyu Zhao[1,2,*]

[1]Department of Epidemiology and Public Health, [2]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA and [3]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA

Received on April 21, 2004; revised on May 20, 2004; accepted on June 19, 2004
Advance Access publication July 1, 2004

ABSTRACT

Motivation: To understand cancer etiology, it is important to explore molecular changes in cellular processes from normal state to cancerous state. Because genes interact with each other during cellular processes, carcinogenesis related genes may form differential co-expression patterns with other genes in different cell states. In this study, we develop a statistical method for identifying differential gene–gene co-expression patterns in different cell states.

Results: For efficient pattern recognition, we extend the traditional *F*-statistic and obtain an Expected Conditional *F*-statistic (*ECF*-statistic), which incorporates statistical information of location and correlation. We also propose a statistical method for data transformation. Our approach is applied to a microarray gene expression dataset for prostate cancer study. For a gene of interest, our method can select other genes that have differential gene–gene co-expression patterns with this gene in different cell states. The 10 most frequently selected genes, include *hepsin*, *GSTP1* and *AMACR*, which have recently been proposed to be associated with prostate carcinogenesis. However, genes *GSTP1* and *AMACR* cannot be identified by studying differential gene expression alone. By using tumor suppressor genes *TP53*, *PTEN* and *RB1*, we identify seven genes that also include *hepsin*, *GSTP1* and *AMACR*. We show that genes associated with cancer may have differential gene–gene expression patterns with many other genes in different cell states. By discovering such patterns, we may be able to identify carcinogenesis related genes.

Availability: The R-codes for our study are available at http://bioinformatics.med.yale.edu/microarray/BioSupp1.html

Contact: hongyu.zhao@yale.edu

INTRODUCTION

Microarrays enable us to simultaneously screen expression of thousands of genes and generate enormous amount of data (Schena *et al*., 1995; Lashkari *et al*., 1997). With such techniques, it is possible to explore cellular processes at the molecular level on a genomic scale (DeRisi *et al*., 1997). Highly correlated genes are likely to be involved in the same biological process (Eisen *et al*., 1998; Marcotte *et al*., 1999). The correlation coefficient and its variants (Eisen *et al*., 1998; Cherepinsky *et al*., 2003) are widely used to measure the correlation of two genes.

To understand cancer etiology, it is important to explore molecular changes in cellular processes from normal state to cancerous state. Microarray techniques can be used for molecular classification of cancer (Golub *et al*., 1999; van't Veer *et al*., 2002). Differentially expressed genes are potential markers for clinical diagnoses and medical treatments. The *F*-statistic and its variants are commonly used to identify differentially expressed genes, e.g. *t*-test, signal-to-noise statistic (Golub *et al*., 1999) and SAM method (Tusher *et al*., 2001).

However, since there may be no significant correlation between protein and gene expression abundance (Gygi *et al*., 1999; Chen *et al*., 2002; Washburn *et al*., 2003), some carcinogenesis related genes may not be identified by finding differentially expressed genes. The genome-wide co-expression dynamics (Li, 2002) shows that the pattern of gene–gene co-expression may depend on another gene's expression level. Similarly, carcinogenesis related genes may also form differential co-expression patterns with other genes in different cell states, which can be utilized as an alternative approach to identifying carcinogenesis related genes.

There are many possible reasons for the existence of differential gene–gene co-expression patterns. For example, depending on different stimuli, the transcription factor nuclear factor kappa B (NF-κB) can be either an activator or a repressor of its target genes (Campbell *et al*., 2004). NF-κB complexes are comprised of homo- or hetero-dimers formed from the multigene family of RelA (p65), c-Rel, RelB, NF-κB1 (p50/p105) and NF-κB2 (p52/p100). Stimulated with tumor necrosis factor (TNF), NF-κB induces the expression of antiapoptotic genes Bcl-XL, X-IAP and A20. Therefore, NF-κB has an antiapoptotic effect. On the other hand, stimulated with ultraviolet light (UV-C) or the chemotherapeutic drugs daunorubicin, NF-κB is converted into a

*To whom correspondence should be addressed.

Bioinformatics vol. 20 issue 17 © Oxford University Press 2004; all rights reserved.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

repressor of antiapoptotic gene transcription through inducing its association with histone deacetylases, which reduces the expression of Bcl-xL, X-IAP and A20. Therefore, NF-κB can also be proapoptotic. Another example is p53, a tumor suppressor gene (Willis *et al*., 2004). p53 is mutated in many human cancers, which generally results in the loss of its function. However, some of the p53 mutations are dominantly-negatively inhibiting the function of wild-type p53. These p53 mutants reduce the binding of wild-type p53 to the p53 responsive element in the target genes of p21, MDM2 and PIG3. Therefore, p53 positively correlates with its target genes in the $p53^{+/+}$ wild-type, and negatively correlates with its target genes in the $p53^{DN/+}$ dominant-negative mutant.

Motivated by Li's study (2002), we study differential gene–gene co-expression patterns in different cell states to understand molecular changes in cellular processes from normal state to cancerous state. Such study may provide insight to cancer etiology at molecular level. In contrast to Li (2002), who analyzed microarray gene expression data for yeast cell cycle (Spellman *et al*., 1998), we focus on classification data, and apply our method to a microarray gene expression dataset for prostate cancer (Singh *et al*., 2002). We chose to analyze this dataset because of its relatively large sample size and the importance of prostate cancer. In the United States, prostate cancer is one of the most common malignancies in men. It was estimated that about one in six men would be diagnosed with this disease, and there were about 189 000 diagnoses and 30 200 deaths in 2002 (DeMarzo *et al*., 2003a). A main problem with the prostate cancer is that its molecular mechanisms still remain unclear. Although numerous genes were discovered to be associated with prostate cancer, there is no detection of major predisposition genes (Visakorpi, 2003). Exploring differential gene–gene co-expression patterns in different cell states may lead to uncover major predisposition genes.

For the rest of the paper, we first introduce the microarray gene expression dataset used for our study, and then describe our statistical method for pattern recognition by extending the traditional F-statistic. We also propose a method for data transformation to handle outliers. Finally, we present our analysis results and discuss significance of biological findings.

METHODOLOGY

Gene expression data

In this study, we use a published microarray gene expression dataset for prostate cancer (Singh *et al*., 2002). Gene expression levels were measured for samples of prostate tumors ('cancerous') and adjacent prostate tissues not containing tumor ('normal'). Data were generated using Affymetrix oligonucleotides microarrays and GeneChip Software. The dataset contains probes for 12 600 genes and expressed sequence tags (ESTs). There are 50 normal samples and 52 cancerous samples, and no missing values in the dataset.

Before analyzing the data, we perform the following thresholding and filtering according to Singh *et al*. (2002). Any measurements below 10 are set as 10, and any measurements above 16 000 are set as 16 000. After thresholding, we find the maximum (Max) and the minimum (Min) measurements for each gene. A gene is excluded from the analysis if Max/Min <5 or Max−Min <50. There are 6034 genes left after filtering.

Expected conditional F-statistic

Suppose that there are g different sample groups. For a gene X, let x_{ij} be the j-th observation in the i-th group, n_i be the number of observations, $\bar{x}_i$ be the sample mean for the i-th group and $\bar{x}$ be the sample mean for all observations, where $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n_i$. Also, let $n = \sum_i n_i$ be the total number of observations. The following F-statistic is widely used to test whether gene X is differentially expressed in different sample groups:

$$F = \frac{\sum_i n_i(\bar{x}_i - \bar{x})^2/(g-1)}{\sum_{i,j}(x_{ij} - \bar{x}_i)^2/(n-g)}.$$

Under normality, when $\lim_{n\to\infty} n_i/n = p_i > 0$, we have a weak convergence of the modified F-statistic

$$\frac{g-1}{n-g}F = \frac{\sum_i n_i(\bar{x}_i - \bar{x})^2}{\sum_{i,j}(x_{ij} - \bar{x}_i)^2} \to \frac{\sum_i p_i(\mu_i - \sum_j p_j\mu_j)^2}{\sum_i p_i\sigma_i^2}$$

in probability, as $n \to \infty$; where μ_i and σ_i^2 are the mean and variance of the i-th group, respectively. Let λ denote the above limit. Following simple algebra, we have

$$\lambda = \frac{\sum_i p_i(\mu_i - \sum_j p_j\mu_j)^2}{\sum_i p_i\sigma_i^2} = \frac{\sum_{i<j} p_i p_j(\mu_i - \mu_j)^2}{\sum_i p_i\sigma_i^2}.$$

To extend F-statistic for identifying differential gene–gene co-expression patterns, we assume that two genes X and Y are normally distributed in the i-th group,

$$(X_i, Y_i) \sim N\left[(\mu_{Xi}, \mu_{Yi}), \begin{pmatrix} \sigma_{Xi}^2 & \rho_i\sigma_{Xi}\sigma_{Yi} \\ \rho_i\sigma_{Xi}\sigma_{Yi} & \sigma_{Yi}^2 \end{pmatrix}\right].$$

The conditional distribution of X_i given $Y_i = y$ is

$$X_i|Y_i = y \sim N\left[\mu_{Xi} + \rho_i(y - \mu_{Yi})\sigma_{Xi}/\sigma_{Yi}, \sigma_{Xi}^2\left(1 - \rho_i^2\right)\right].$$

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

Replacing means and variances in the formula of λ with the corresponding conditional means and variances, we have

$$\lambda_{X|Y=y} = \left[\sum_i p_i\sigma_{Xi}^2\left(1-\rho_i^2\right)\right]^{-1}\sum_{i<j} p_i p_j[(\mu_{Xi}-\mu_{Xj}) - (\mu_{Yi}\rho_i\sigma_{Xi}/\sigma_{Yi} - \mu_{Yj}\rho_j\sigma_{Xj}/\sigma_{Yj}) + (\rho_i\sigma_{Xi}/\sigma_{Yi} - \rho_j\sigma_{Xj}/\sigma_{Yj})y]^2.$$

For gene Y, let $f_{Yk}(y)\sim N(\mu_{Yk},\sigma_{Yk}^2)$ be the probability density function (p.d.f.) of group k, and $f_Y(Y) = \sum_k p_k f_{Yk}(y)$. Taking expectation over variable Y, we have

$$\begin{aligned}
&\mathbf{E}_Y(\lambda_{X|Y=y})\\
&= \int_Y \lambda_{X|Y=y} f_Y(y)\mathrm{d}y\\
&= \sum_k p_k \int_Y \lambda_{X|Y=y} f_{Yk}(y)\mathrm{d}y\\
&= \left[\sum_i p_i\sigma_{Xi}^2\left(1-\rho_i^2\right)\right]^{-1}\sum_{i<j}\sum_k p_i p_j p_k\{[(\mu_{Xi}-\mu_{Xj})\\
&\quad - (\mu_{Yi}\rho_i\sigma_{Xi}/\sigma_{Yi} - \mu_{Yj}\rho_j\sigma_{Xj}/\sigma_{Yj})\\
&\quad + (\rho_i\sigma_{Xi}/\sigma_{Yi} - \rho_j\sigma_{Xj}/\sigma_{Yj})\mu_{Yk}]^2\\
&\quad + (\rho_i\sigma_{X_i}/\sigma_{Y_i} - \rho_j\sigma_{Xj}/\sigma_{Yj})^2\sigma_{Yk}^2\}.
\end{aligned}$$

If $\sigma_{Xi} = \sigma_X$ and $\sigma_{Yi} = \sigma_Y$ for all i, then $\mathbf{E}_Y(\lambda_{X|Y=y})$ can be simplified as

$$\begin{aligned}
&\mathbf{E}_Y(\lambda_{X|Y=y})\\
&= \left[\sum_i p_i\left(1-\rho_i^2\right)\right]^{-1}\sum_k\sum_{i<j} p_k p_i p_j\{[(\mu_{Xi}-\mu_{Xj})/\sigma_K\\
&\quad - \rho_i(\mu_{Yi}-\mu_{Yk})/\sigma_Y + \rho_j(\mu_{Yj}-\mu_{Yk})/\sigma_Y]^2\\
&\quad + (\rho_i-\rho_j)^2\}.
\end{aligned}$$

This simplified formula incorporates two types of statistical measurements: non-central parameters of t-distribution and correlation coefficients. In general, $\mathbf{E}_Y(\lambda_{X|Y=y})$ incorporates both location and correlation information.

We estimate $\mathbf{E}_Y(\lambda_{X|Y=y})$ by estimating the parameters in the formula, and we obtain a statistic. We name this statistic Expected Conditional F-statistic (*ECF*-statistic) $\mathbf{E}_Y(F_{X|Y=y})$. The parameters in the formula are estimated using standard methods, i.e.

$$\widehat{\mu_{Xi}} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij},$$

$$\widehat{\sigma_{Xi}^2} = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu_{Xi}})^2.$$

For observations of a pair of genes (X,Y) in group i, we estimate

$$\hat{\rho}_i = \frac{\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu_{Xi}})(y_{ij}-\widehat{\mu_{Yi}})}{\sqrt{\sum_{j=1}^{n_i}(x_{ij}-\widehat{\mu_{Xi}})^2}\sqrt{\sum_{j=1}^{n_i}(y_{ij}-\widehat{\mu_{Yi}})^2}}.$$

Data transformation

To estimate the *ECF*-statistic for a pair of genes, we need to estimate the means and variances of two genes for each sample group. For this particular dataset, the sample sizes (50 normal samples and 52 cancerous samples) are adequate to obtain robust estimations for these statistics using traditional estimation methods. However, the sample sizes may not be adequate for robust bivariate analysis (e.g. correlation coefficient), and the existence of outliers may seriously influence the estimation of correlation coefficients. Therefore, we propose data transformation to handle potential outliers.

Another reason for data transformation is the underlying distribution of the data. The *ECF*-statistic is derived based on the normal distribution assumption. To achieve statistical efficiency, we propose data transformation so that the underlying distribution of the data agrees with the normal distribution.

In Li's study on genome-wide co-expression dynamics (Li, 2002), rank and inverse standard normal transformations were used for data transformation. In our study, we assume that gene expression has different distributions for different sample groups. Therefore, we generalize the data transformation used by Li (2002) and propose the following procedure.

For each gene passing the filtering criterion (see previous data description), let $X = (x_1, x_2, \ldots, x_n)$ be the data before thresholding, and $X' = (x_1', x_2', \ldots, x_n')$ be the data after thresholding. Also, let $\Phi_{(\mu,\sigma^2)}$ be the cumulative distribution function (c.d.f.) for normal distribution with mean μ and variance σ^2. We perform the following transformation procedure.

- Estimate μ_i and σ_i^2 for each group i using X'.
- Construct a mixture c.d.f. of normal distributions, $\Psi(x) = \sum_i p_i\Phi_{(\mu_i,\sigma_i^2)}(x)$.
- Rank all the observations using X, $(x_1, x_2, \ldots, x_n) \rightarrow (r_1, r_2, \ldots, r_n)$.
- Invert the mixture function, $z = \Psi^{-1}[r/(n+1)]$, $(r_1, r_2, \ldots, r_n) \rightarrow (z_1, z_2, \ldots, z_n)$.

In general, it is difficult to calculate the inverse function $\Psi^{-1}(x)$, so we utilize a large number of simulations to approximate the inverse function. The purpose of using the data before thresholding for ranking is to preserve the original order.

Significance assessment

To assess the significance level for an *ECF*-statistic, we consider the following null hypothesis for a pair of

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

genes $(X,\ Y)$.

- There is no distribution difference among sample groups for any gene expression.
- Two genes are independently expressed in any sample group.

We can simply use permutation test on the observed data to assess the significant level for an *ECF*-statistic. However, as there are about 18 201 561 possible pairs, this approach requires extremely intensive computation. Therefore, we choose not to use this method in our study.

Instead, we use simulations to generate a distribution for the null hypothesis to assess significance. When genes (X, Y) are normally distributed, our null hypothesis is equivalent to $\mu_{\cdot i} = \mu_{\cdot j}, \sigma_{\cdot i} = \sigma_{\cdot j}$ and $\rho_i = 0$ for any i, j. Therefore, we run a large number of simulations from two independent standard normal distributions and obtain an approximate distribution for the *ECF*-statistic.

Differential gene–gene co-expression patterns identification

We propose the following procedure for identifying differential gene-gene co-expression patterns in different sample groups. For a gene X of interest, this procedure screens all the other genes Y and select genes that have differential gene–gene co-expression patterns with gene X.

- Calculate two *ECF*-statistics: $\mathbf{E}_Y(F_{X|Y=y})$ and $\mathbf{E}_X(F_{Y|X=x})$.
- Select Y for X when both *ECF*-statistics are significant.

Note that the *ECF*-statistic is not symmetric. For a pair of genes (X, Y), two *ECF*-statistics $\mathbf{E}_Y(F_{X|Y=y})$ and $\mathbf{E}_X(F_{Y|X=x})$ may not be the same. For consistency purpose, a pair of genes will be considered to be significant and selected if both *ECF*-statistics are greater than a threshold value. When multiple pairs of genes are selected, we rank them by the combined *ECF* score that is defined as $ECF(X, Y) = \mathbf{E}_Y(F_{X|Y=y}) + \mathbf{E}_X(F_{Y|X=x})$.

RESULTS

Data transformation

For comparison, we use both F-statistic and *ECF*-statistic to analyze the data. Considering multiple comparison adjustments, we set a threshold value 10^{-6} for the P-value. The corresponding F-statistic is $\sim$27.2 from the theoretical F-distribution, and the corresponding *ECF*-statistic is $\sim$0.32 from the simulations. First, we show that our data transformation method is efficient based on the following observations.

We use the Shapiro–Wilk normality test (Shapiro and Wilk, 1965; Royston, 1982) to evaluate whether the underlying distribution of the data agrees with normal distribution. From Figure 1, before data transformation, we observe that most of the P-values are quite significant both in the normal group and the cancerous group, indicating deviation from the normal distribution for most of the gene expressions. After data transformation, only a small portion of the P-values are significant. Therefore, the underlying distribution of the data agrees more with the normal distribution after the data transformation.

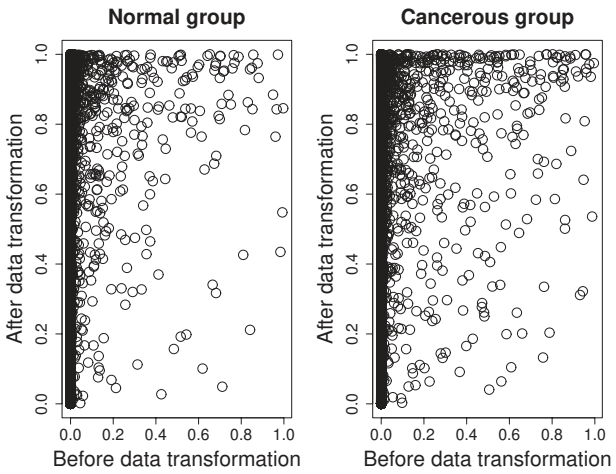

Fig. 1. Effect of data transformation on normality test. P-value of the Shapiro–Wilk normality test is calculated for each gene in the normal group and the cancerous groups. x-axis represents the P-value before data transformation and y-axis represents the P-value after data transformation.

Before data transformation, 101 differentially expressed genes are selected based on our criterion. The most significant gene is hepsin, which was recently proposed as a potential marker for prostate cancer (DeMarzo *et al.*, 2003b; Stephan *et al.*, 2004). In Figure 2, F-statistic after data transformation is compared to F-statistic before data transformation. Only two genes that are selected before data transformation are not selected after data transformation. But 65 genes that are not selected before data transformation are selected after data transformation. *GSTP1*, which is associated with prostate cancer (DeMarzo *et al.*, 2003a,b; Visakorpi, 2003), is among these 65 additionally selected genes.

In Figure 3, *ECF*-statistic for gene *TP53* [fix gene X as *TP53* in $\mathbf{E}_X(F_{Y|X=x})$] after data transformation is compared to *ECF*-statistic before data transformation. The *ECF*-statistics can be very different before and after data transformation. Figure 4 shows that outlier impact is reduced by data transformation. The pair of genes (*TP53* and cDNA DKFZp564G013) in the figure are selected by our method before data transformation. But this pair is selected simply because of the existence of an outlier, which makes the difference of correlation coefficients overestimated. After data transformation, this pair is no longer selected.

Most frequently selected genes

We perform our pattern recognition procedure for every gene passing filtering criterion (see previous method description).

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

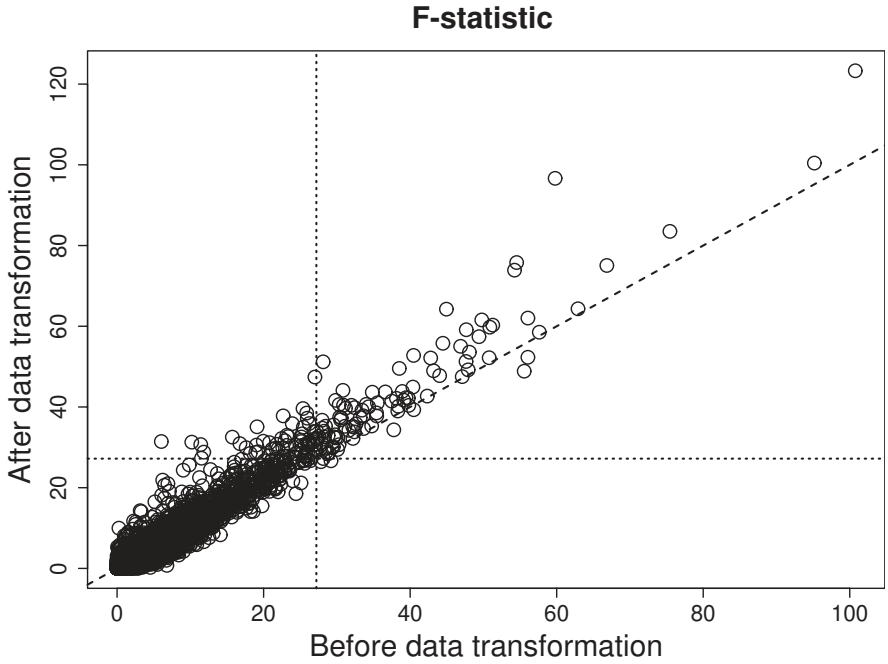

Fig. 2. Effect of data transformation on F-statistic. x-axis represents the F-statistic before data transformation and y-axis represents the F-statistic after data transformation. The dashed line represents where two values are equal. Two dotted lines represent the threshold values to select differentially expressed genes.

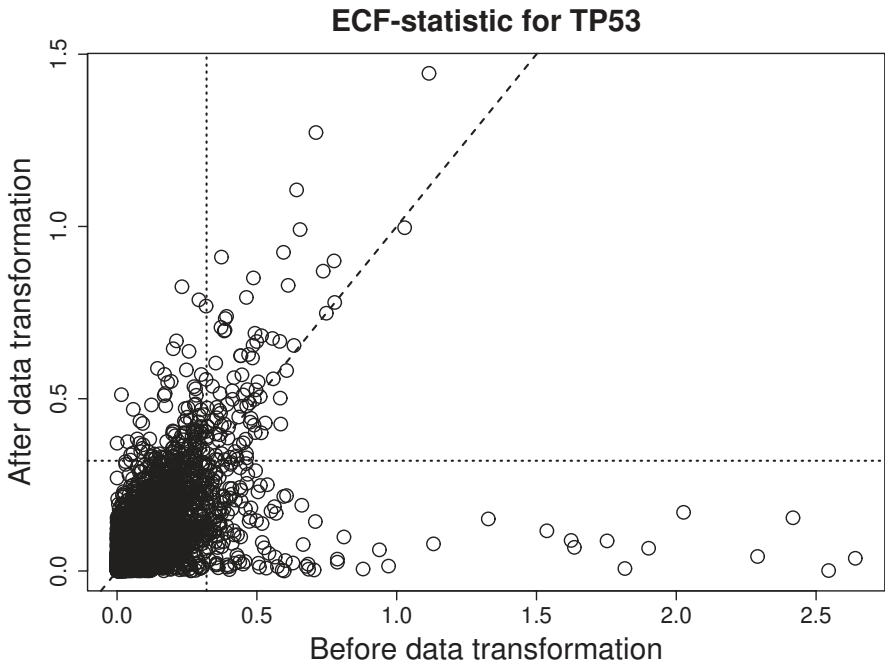

Fig. 3. Effect of data transformation on *ECF*-statistic. x-axis represents the *ECF*-statistic before data transformation and y-axis represents the *ECF*-statistic after data transformation for gene *TP53* and another gene. The dashed line represents where two values are equal. Two dotted lines represent the threshold values to select differential gene–gene co-expression patterns.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

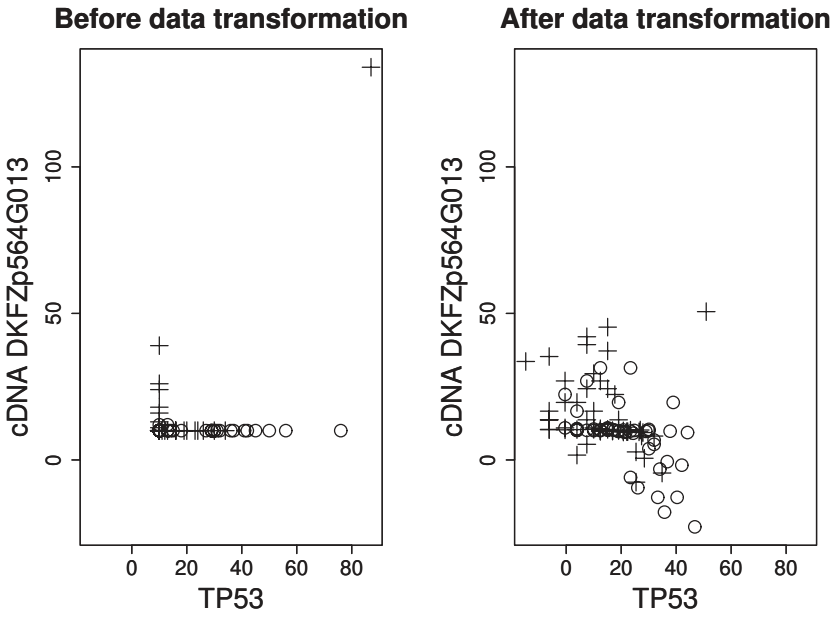

Fig. 4. Effect of data transformation on outliers. Before data transformation, our method selects a pair of genes *TP53* and cDNA DKFZp564G013. After data transformation, this pair is no longer selected. Circles represent observations from the normal group and crosses represent observations from the cancerous group.

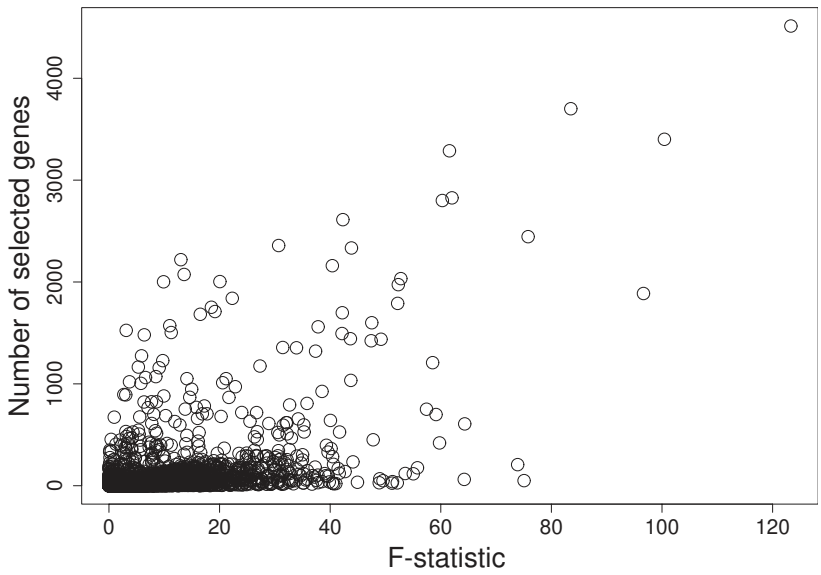

Fig. 5. Comparison between the F-statistic for a gene and the number of genes selected for this gene. x-axis represents the F-statistic after data transformation and y-axis represents the number of selected genes.

For each gene, the number of genes selected for it varies in a wide range. Among 6034 genes, there are 484 genes with no genes selected for them; 3574 genes with more than 10 genes selected; 583 genes with more than 100 genes selected; and 51 genes with more than 1000 genes selected. Figure 5 gives the comparison between the F-statistic for a gene and the number of genes selected for this gene. We observe that some genes with significant F-statistic have a small number of selected genes, and some genes with insignificant F-statistic have a large number of selected genes. Therefore, there is no trivial connection between the F-statistic and the ECF-statistic.

Table 1 lists 10 most frequently selected genes. *Hepsin* and *GSTP1* are among these genes, which can be identified by

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

Table 1. The most frequently selected genes

Gene	S	$F(R)$
X07732: hepatoma mRNA for serine protease hepsin	4513	123.3 (1)
M30894: T-cell receptor Ti rearranged gamma-chain mRNA V-J-C region	3701	83.5 (4)
M22382: mitochondrial matrix protein P1 (nuclear encoded) mRNA	3401	100.4 (2)
AF045229: regulator of G protein signaling 10 mRNA	3288	61.6 (11)
J03592: ADP/ATP translocase mRNA, 3′ end, clone pHAT8	2826	62.0 (10)
AL049969: mRNA for cDNA DKFZp564A072	2800	60.3 (12)
M84526: adipsin/complement factor D mRNA	2445	75.8 (5)
U21689: mRNA for glutathione S-transferase-P1c gene (GSTP1)	2358	30.7 (135)
X17620: mRNA for Nm23 protein, involved in developmental regulation	2334	43.8 (36)
AJ130733: mRNA 2-methylacyl-CoA racemase (AMACR)	2219	13.0 (842)

S represents the number of being selected, which is used for ranking; F represents the F-statistic for the gene and R in parentheses represents the number of corresponding ranks among 6034 genes.

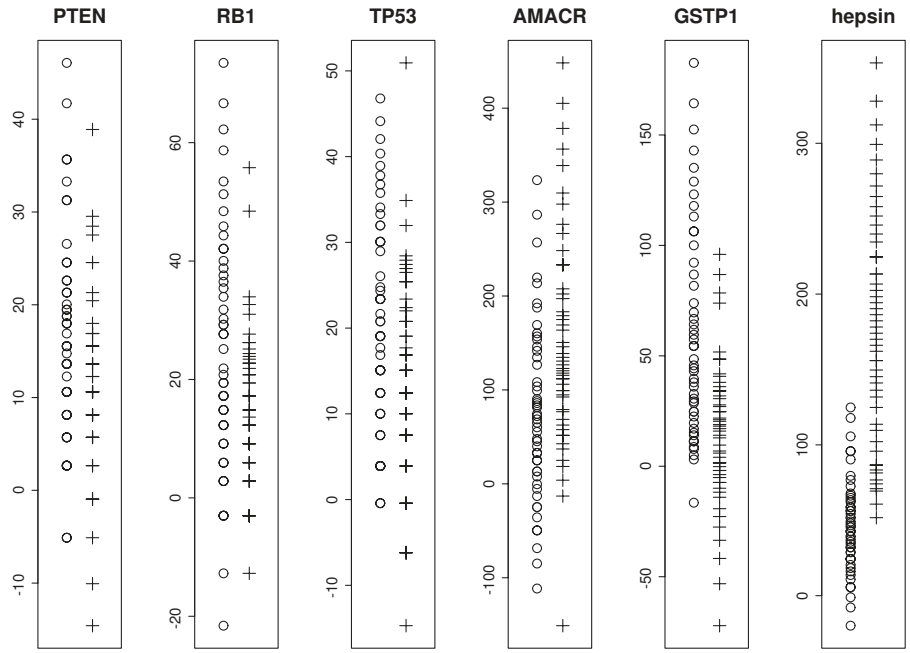

Fig. 6. Transformed data for six genes associated with prostate cancer. Circles represent observations from normal group and crosses represent observations from cancerous group.

the F-statistic. However, *GSTP1* ranks only 135 by the F-statistic. Furthermore, *AMACR*, which is another gene associated with prostate cancer (Rubin *et al*., 2002; DeMarzo *et al*., 2003a,b), is on this list. However, from Figure 6, we observe that gene *AMACR* cannot be identified by the F-statistic.

Selected pairs of genes

Figure 7 shows plots for six selected pairs of genes with the most significant combined *ECF* scores. *Hepsin* is always in these pairs. For the 5th pair, there is negative correlation in the normal group, and positive correlation in the cancerous group. For the other five pairs, there are strong positive correlations in the normal group, but no clear pattern in the cancerous group. These observations suggest differential gene–gene co-expression patterns in different cell states. Such molecular changes at 'pattern' level may provide important information to understand cancer mechanisms.

Tumor suppressor genes

Tumor suppressor genes play crucial roles in cancer pathways (Hahn and Weinberg, 2002). A recent review (DeMarzo *et al*., 2003b) summarizes several tumor suppressor genes associated with prostate cancer. Among them, *PTEN*, *RB1* and *TP53* are in the dataset and passed filtering criterion.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

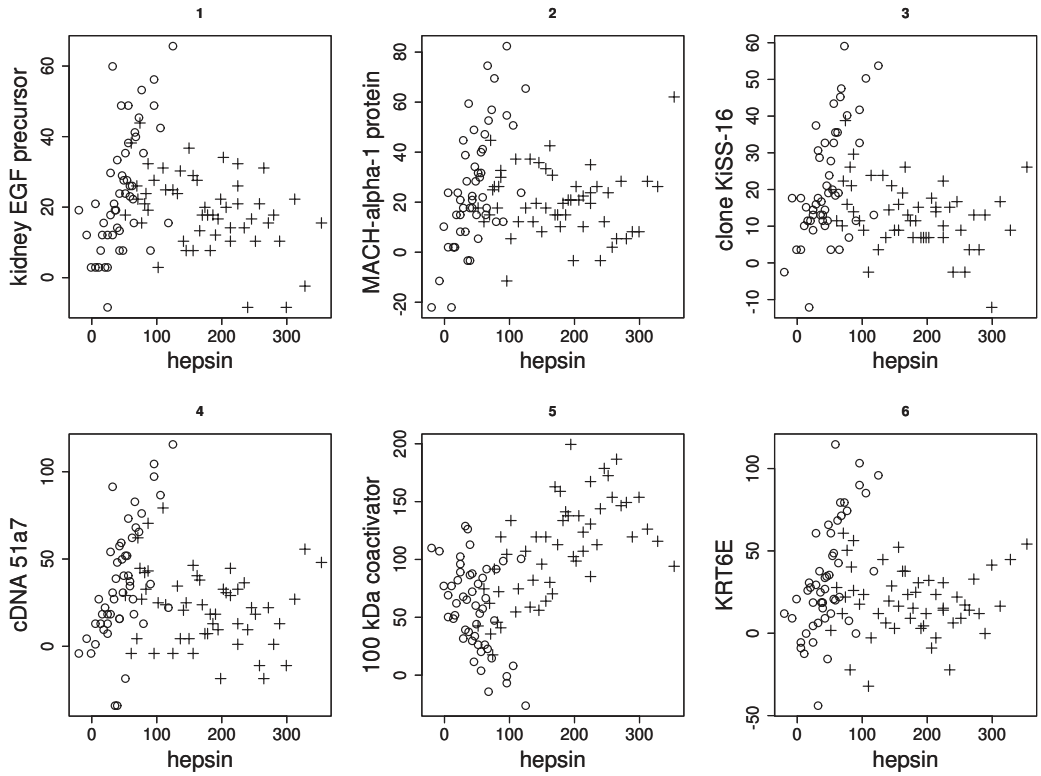

Fig. 7. Plots for six most significant pairs of genes selected by our method after data transformation. Circles represent observations from normal group and crosses represent observations from cancerous group.

Table 2. The number of selected genes for three tumor suppressor genes

Gene	*PTEN*	*RB1*	*TP53*
U92436: *PTEN*	15	11	8
L41870: *RB1*	—	66	35
X02469: *TP53*	—	—	67

The number in the diagonal cell represents the number of genes selected for the corresponding tumor suppressor gene. The number in the non-diagonal cell represents the number of genes that are selected in common for the two corresponding tumor suppressor genes.

PTEN, *RB1* and *TP53* all play a critical role in cell apoptosis (Hahn and Weinberg, 2002; Vousden and Lu, 2002; Chau and Wang, 2003), which is an important cellular mechanism to prevent cell proliferation and is targeted in cancer therapy to repress tumor cell proliferation. From Figure 6, we observe that there is almost no distribution differences in the normal group and the cancerous groups for these genes. However, they have different number of genes selected for them to form differential gene–gene co-expression patterns. Table 2 gives these numbers of genes selected for them. From these three sets of selected genes, there are seven genes in common including *hepsin*, *GSTP1* and *AMACR*, which implies the involvement of these genes in cancer pathways, and further supports the importance of these genes in prostate cancer studies.

Figure 8 shows plots for these differential gene–gene co-expression patterns. Tumor suppressor genes have similar patterns with *hepsin*, *GSTP1* and *AMACR*. *Hepsin* and *AMACR* are positively correlated with *PTEN*, *RB1* and *TP53* in the normal group; but there is no clear pattern in the cancerous group. *GSTP1* is negatively correlated with *PTEN*, *RB1* and *TP53* both in the normal group and the cancerous group; but the patterns varies in different groups at an elevated level.

DISCUSSION

There are complicated molecular changes during cellular processes from normal state to cancerous state, such as gain or loss some transcription factors, and change of chemical compounds. Such changes may result in differential gene–co-expression patterns in normal state and cancerous state. Our statistical method is capable of identifying differential gene–gene co-expression patterns. These patterns are certainly helpful for us to further understand cancer mechanism at molecular level.

We have proposed a statistical method for identifying pairs of genes with differential gene–gene co-expression patterns.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

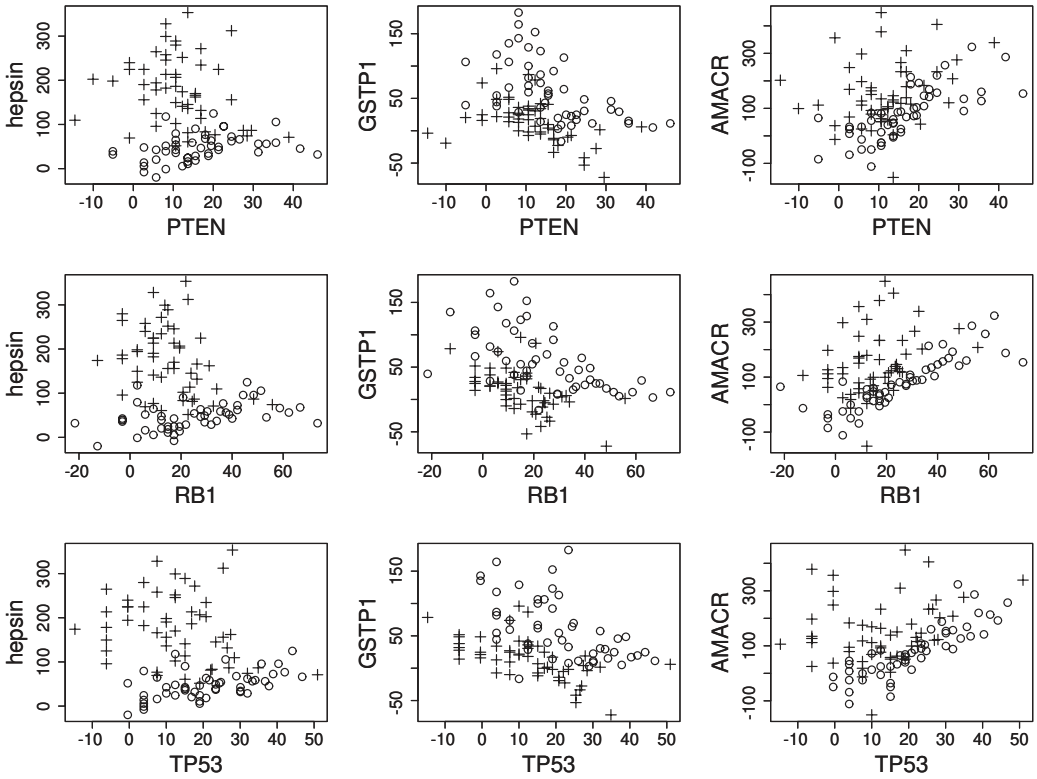

Fig. 8. Plots for genes *hepsin*, *GSTP1* and *AMACR* versus tumor suppressor genes *PTEN*, *RB1* and *TP53*. Circles represent observations from the normal group and crosses represent observations from the cancerous group.

We extend the traditional F-statistic to obtain the *ECF*-statistic, which incorporates statistical information of location and correlation. Our method is applied to a microarray gene expression dataset for prostate cancer, and numerous pairs of genes with differential gene–gene co-expression patterns are identified. Based on the number of being selected in pairs, genes are ranked by their importance. Several genes associated with prostate cancer are on the top list. Some of these genes cannot be identified by univariate analysis, such as the F-statistic and its variants. In addition, the results show differential gene–gene co-expression patterns in different cell states for genes associated with prostate cancer.

We observe that an efficient data transformation method can improve simple statistical methods. After data transformation, the underlying distribution of data agrees more with normal distribution. Also, more genes with biological significance are identified using F-statistic. Furthermore, outlier impact is reduced, leading to fewer false identified differential gene–gene co-expression patterns.

Since our *ECF*-statistic incorporates both location and correlation information, it can be used to identify various types of differential gene–gene co-expression patterns. In contrast, Li's method (2002) only focuses on the difference of correlations. There are some differential gene–gene co-expression patterns not identifiable using Li's method. For examples, our method identified pairs of genes *GSTP1* with tumor suppressor *PTEN*, *RB1* or *TP53*. Such pairs cannot be identified by Li's method because there are negative correlations in both the normal group and the cancerous group.

Our statistical method can also be applied to other types of data, such as mass spectrometry proteomics data. A biologically interesting topic for future research is to seek explanations for differential gene–gene co-expression patterns in the normal state and the cancerous state. Statistically, it is necessary to study the significance of the number of being selected for a gene. Also, the recent study on multiple hypothesis testing problem (Benjamini and Hochberg, 1995) will be useful for false control on numerous selected pairs of genes. Furthermore, it will be interesting to understand statistical properties of the *ECF*-statistic, such as its theoretical distribution and asymptotic behaviors.

ACKNOWLEDGEMENTS

We thank two anonymous reviewers for their valuable comments. This work was supported in part by NSF grant DMS 0241160 and NIH grant GM 59507.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010

REFERENCES

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.

Campbell,K.J., Rocha,S. and Perkins,N.D. (2004) Active repression of antiapoptotic gene expression by RelA (p65) NF-kappa B. *Mol. Cell*, **13**, 853–865.

Chau,B.N. and Wang,J.Y.J. (2003) Coordinated regulation of life and death by RB. *Nat. Rev. Cancer*, **3**, 130–138.

Chen,G.A., Gharib,T.G., Huang,C.C., Taylor,J.M.G., Misek,D.E., Kardia,S.L.R., Giordano,T.J., Iannettoni,M.D., Orringer,M.B., Hanash,S.M. and Beer,D.G. (2002) Discordant protein and mRNA expression in lung adenocarcinomas. *Mol. Cell. Proteomics*, **1**, 304–313.

Cherepinsky,V., Feng,J.W., Rejali,M. and Mishra,B. (2003) Shrinkage-based similarity metric for cluster analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 9668–9673.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.

DeMarzo,A.M., Meeker,A.K., Zha,S., Luo,J., Nakayama,M., Platz,E.A., Isaacs,W.B. and Nelson,W.G. (2003a) Human prostate cancer precursors and pathobiology. *Urology*, **62** (Suppl. 5A), 55–62.

DeMarzo,A.M., Nelson,W.G., Isaacs,W.B. and Epstein,J.I. (2003b) Pathological and molecular aspects of prostate cancer. *Lancet*, **361**, 955–964.

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

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

Gygi,S.P., Rochon,Y., Franza,B.R. and Aebersold,R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, **19**, 1720–1730.

Hahn,W.C. and Weinberg,R.A. (2002) Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer*, **2**, 331–341.

Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci., USA*, **94**, 13057–13062.

Li,K.C. (2002) Genome-wide co-expression dynamics: theory and application. *Proc. Natl Acad. Sci., USA*, **99**, 16875–16880.

Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Royston,J.P. (1982) An extension of Shapiro and Wilk's *W* test for normality to large samples. *Appl. Stat.*, **31**, 115–124.

Rubin,M.A., Zhou,M., Dhanasekaran,S.M., Varambally,S., Barrette,T.R., Sanda,M.G., Pienta,K.J., Ghosh,D. and Chinnaiyan,A.M. (2002) α-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *JAMA*, **287**, 1662–1670.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Shapiro,S.S. and Wilk,M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Singh,D., Febbo,P.G., Ross,K., Jackson,D.G., Manola,J., Ladd,C., Tamayo,P., Renshaw,A.A., D'Amico,A.V., Richie,J.P. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stephan,C., Yousef,G.M., Scorilas,A., Jung,K., Jung,M., Kristiansen,G., Hauptmann,S., Kishi,T., Nakamura,T., Loening,S.A. and Diamandis,E.P. (2004) Hepsin is highly over expressed in and a new candidate for a prognostic indicator in prostate cancer. *J. Urol.*, **171**, 187–191.

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

van't Veer,L.J., Dai,H.Y., van de Vijver,M.J., He,Y.D.D., Hart,A.A.M., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Visakorpi,T. (2003) The molecular genetics of prostate cancer. *Urology*, **62** (Suppl. 5A), 3–10.

Vousden,K.H. and Lu,X. (2002) Live or let die: the cell's response to p53. *Nat. Rev. Cancer*, **2**, 594–604.

Washburn,M.P., Koller,A., Oshiro,G., Ulaszek,R.R., Plouffe,D., Deciu,C., Winzeler,E. and Yates,J.R.,III, (2003) Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae. Proc. Natl Acad. Sci., USA*, **100**, 3107–3112.

Willis,A., Jung,E.J., Wakefield,T. and Chen,X.B. (2004) Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene*, **23**, 2330–2338.

Downloaded from http://bioinformatics.oxfordjournals.org at University of Wisconsin Madison - General Library System on June 7, 2010