Technology in Cancer Research and Treatment
ISSN 1533-0346
Volume 9, Number 3, June 2010
©Adenine Press (2010)

Discovering Disease-specific Biomarker Genes for Cancer Diagnosis and Prognosis

www.tcrt.org

Hung-Chung Huang, Ph.D.[1,2]
Siyuan Zheng, Ph.D.[1,2,3]
Vincent VanBuren, Ph.D.[5]
Zhongming Zhao, Ph.D.[1,2,3,4,*]

[1]Bioinformatics Resource Center,
[2]Functional Genomics Shared Resource,
[3]Department of Biomedical Informatics,
[4]Department of Cancer Biology,
Vanderbilt University Medical Center
Nashville, TN 37232, USA
[5]Department of Systems Biology and
Translational Medicine, College of
Medicine, Texas A&M Health Science
Center, Temple, TX 76504, USA

The large amounts of microarray data provide us a great opportunity to identify gene expression profiles (GEPs) in different tissues or disease states. Disease-specific biomarker genes likely share GEPs that are distinct in disease samples as compared with normal samples. The similarity of the GEPs may be evaluated by Pearson Correlation Coefficient (PCC) and the distinctness of GEPs may be assessed by Kolmogorov-Smirnov distance (KSD). In this study, we used the PCC and KSD metrics for GEPs to identify disease-specific (cancer-specific) biomarkers. We first analyzed and compared GEPs using microarray datasets for smoking and lung cancer. We found that the number of genes with highly different GEPs between comparing groups in smoking dataset was much larger than that in lung cancer dataset; this observation was further verified when we compared GEPs in smoking dataset with prostate cancer datasets. Moreover, our Gene Ontology analysis revealed that the top ranked biomarker candidate genes for prostate cancer were highly enriched in molecular function categories such as 'cytoskeletal protein binding' and biological process categories such as 'muscle contraction'. Finally, we used two genes, *ACTC1* (encoding an actin subunit) and *HPN* (encoding hepsin), to demonstrate the feasibility of diagnosing and monitoring prostate cancer using the expression intensity histograms of marker genes. In summary, our results suggested that this approach might prove promising and powerful for diagnosing and monitoring the patients who come to the clinic for screening or evaluation of a disease state including cancer.

Key words: Gene expression profile; Cancer biomarker; Pearson correlation coefficient; Kolmogorov-Smirnov distance; Cancer diagnosis and prognosis.

Introduction

DNA microarray experiments allow us to simultaneously examine the expression levels of many thousands of genes so that the effects of certain treatments, diseases, or developmental stages on gene expression can be detected (1-5). DNA microarrays have been widely applied in cancer research for better diagnosis and prediction of the disease states (6-8). Traditionally, most microarray studies aim to identify differentially expressed genes (DEGs) by comparing the average gene expression levels between two groups (*e.g.,* the treated vs. control or cancer vs. non-cancer) based on statistical analysis such as Student t-test (9, 10) and SAM (11). To account for gene-specific fluctuations, SAM defines a statistic based on the ratio of the change of expression means (*e.g.,* between two states) of a gene to the standard deviation in the data for that gene. Because SAM is based on the mean of the replicates, it can accurately predict the differentially expressed genes when the expressed intensities are very similar among the replicates in the same state. However, gene expression level of the samples in each state (*e.g.,* normal or cancer group) may be very different. For example, for some genes, while the averages of the $\log_2$ expression intensities in two comparing groups are

*Corresponding Author:
Zhongming Zhao, Ph.D.,
E-mail: zhongming.zhao@vanderbilt.edu

close, the Pearson Correlation Coefficient (PCC) of the two comparing profile vectors could be close to 0 (*i.e.,* no correlation between the profiles) and the Kolmogorov-Smirnov distance (KSD) of comparing profile vectors could be close to 1 (the KSD maximum), indicating a large difference (distance) between the two comparing profiles. Thus, profile-based metrics may be more appropriate in such analysis on identifying the biomarker genes with different expression profiles on two comparing groups (*e.g.,* cancer vs. normal). Gene expression profiles from the samples in normal and diseased groups respectively can also be utilized to predict the disease state as described in this work. So far, the shape of the expression intensity distribution, or gene expression profiles (GEPs), from the samples of each comparison group are often ignored in this type of analysis.

In this study, we defined the GEP of a gene as the distribution of the $\log_2$ values of its normalized expression signal intensities across the samples in the similarly studied microarrays. We hypothesized that the biomarker genes that distinguish cancer cells from normal cells might form distinct GEPs between comparison groups. GEPs are potentially useful for a better prediction of clinical outcome. We applied Pearson Correlation Coefficient (PCC) and Kolmogorov-Smirnov Distance (KSD) to evaluate the similarity and distance of two comparing GEPs respectively. The possible range is from 0 to 1 for KSD while the range is from -1 to 1 for PCC. As illustrated in Figure 1, although the means of the $\log_2$ (expression intensity) are similar in the two groups in comparison, differences in GEPs could be detected by high KSD and low PCC values. Therefore, GEP shapes could be compared by using KSD and PCC metrics. To demonstrate the utility of GEP analysis, we used PCC and KSD methods to evaluate 14,902 human genes in the GDS534 (12) and GDS2771 (13) datasets retrieved from Gene Expression Omnibus (GEO) database (14-16). These two datasets are based on the studies of gene changes caused by smoking (smoker vs. non-smoker, GDS534) and by cancer (lung cancer vs. non-cancer among smokers, GDS2771), respectively. We further analyzed more than 37 thousands of genes and expression sequence tags in other three datasets (GDS2545, GDS2546, GDS2547) (17) that were used in prostate cancer studies, to examine the robustness of discovering the corresponding biomarker candidate genes.

One question that a physician or biomedical researcher often asks is “What is the likelihood that this patient may have this disease (cancer)?”. In response to this question, we described a method to predict the likelihood of a tissue sample being cancerous by measuring the sample’s gene expression of a set of proposed cancer biomarker genes. In this distribution-based method, the prediction power is based on the probability density (histogram) of the biomarker’s expression intensity in different groups of populations (*e.g.,* normal vs. cancer groups). When population information (in microarray data) for a cancer type of interest becomes sufficiently accumulated and the pathological confirmation of the cancer class becomes relatively accurate, the power of our proposed method for predicting unknown sample’s disease state will be likely high. This application holds promise for the patients coming to the clinics for diagnosis and prognosis purposes.

Materials and Methods

Datasets

Datasets GDS534 (12), GDS2771 (13), and (GDS2545, GDS2546, GDS2547) (17) were obtained from the NCBI

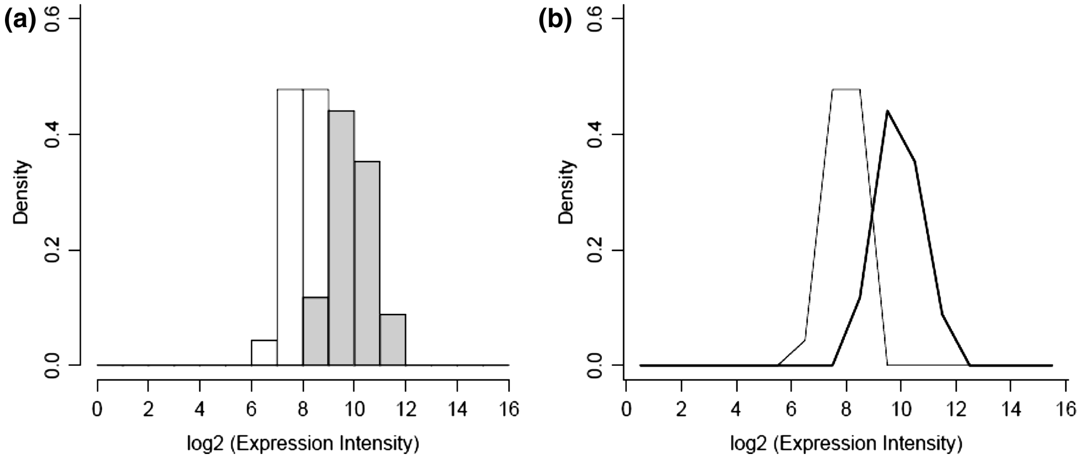

Figure 1: (**a**) Probability density histogram and (**b**) probability density line plot for a representative example comparing the GEPs of gene *GPX2* in smoker and never-smoker samples from GDS534 dataset. Grey bar in (a) and thicker line in (b) represent smoker (mean $\log_2$(Expression Intensity) = 9.9). White bar in (a) and thinner line in (b) represent never-smoker (mean $\log_2$(Expression Intensity) = 7.9). In this example, the KSD was 0.96 and PCC was –0.02 between the two comparing GEPs in each plot.

GEO database (14-16). GDS534 is from a smoking-related study; we compared current smoker vs. never smoker in this dataset. GDS2771 is from a lung cancer related study with two major groups of samples, *i.e.,* lung cancer smokers vs. non-cancer smokers. The other three datasets (GDS2545, GDS2546, and GDS2547) are from the same study on prostate cancer; they were produced by Affymetrix HG-U95A, HG-U95B, and HG-U95C array platforms in the same project (17). All data in these three datasets were obtained by the same normalization procedure so genes are comparable between the sample groups. We compared the GEPs in two major groups of samples in these datasets, *i.e.,* “normal prostate cells” vs. “prostate tumor cells”.

GEP Construction and Biomarker Identification

The GEP of a gene is defined as the distribution of the $\log_2$ values of normalized expression signal intensities across the samples in a set of studied arrays (*e.g.,* in a group of interest). This GEP may be represented by the probability density histogram plot based on the $\log_2$ intensity data as shown in Figure 1. Note that Figure 1 is to represent the probability density histogram, thus, the value on the Y-axis is probability density, not the frequency of the $\log_2$ (expression intensity) for each examined bin interval on the X-axis, The total area under the histogram curve should be 1. By this presentation, the distributions of the gene expression intensities in two different groups of populations could be compared even with different sample size in each group. Histogram bin size could be estimated to be 2 according to the formula “$R/(1 + \log_2 N)$” where R is the range of data (*e.g.,* 0-16 in Figure 1a) and N is the sample size (18). We tested our prostate cancer prediction (described below) with bin size 1, 2, or 3 and found the prediction results with bin size 1 or 2 were very similar and overall better than those with bin size 3 (data not shown). For better resolution on the plot, we decided to use a histogram bin size of 1 for the disease state predictions in our analysis of the prostate cancer datasets.

We examined GEPs based on the different groups of interest in the disease-specific datasets. For GDS534, the GEPs for current-smoking and never-smoking groups were constructed respectively and then compared. For GDS2771, the GEPs for lung cancer and non-cancer groups among smokers were constructed respectively and then compared. For three prostate cancer datasets (GDS2545, GDS2546, and GDS2547), the GEPs for prostate cancer and non-cancer groups were derived and compared. Strong candidates of biomarker genes correlating with phenotypic distinction are expected to have distinctive GEPs between the two comparing groups.

As what we recently proposed (19), group profile distinction for each gene can be systematically assessed using the Kolmogorov-Smirnov Distance (KSD) (20, 21) and Pearson Correlation Coefficient (PCC) (22, 23) computed between the two comparing GEPs based on the probability density histogram data. These measures were used to compare a gene’s GEP in one group to the same gene’s GEP obtained from the other group. The most promising group-specific biomarker candidate genes are expected to have large KSD and low PCC values between the two comparing GEPs, as seen in the upper left square of the KSD vs. PCC metric plot in Figure 2. The most promising group-specific biomarker candidate genes were thus selected with combined KSD and PCC cutoff values. The cutoff values may be chosen by different ways, or arbitrarily, depending on datasets and investigator’s interest. In this study, we used cutoff values that allowed us to select approximately one to two hundred top ranked genes as candidate biomarkers (~0.5-1% of the total genes investigated in microarray). A better approach to identifying optimal cutoffs for the biomarker candidate genes may be developed in future. For example, we may perform permutation for each gene regarding PCC and KSD values. For PCC, the P-value may be calculated as the proportion of sampled permutations whose PCC value is less than the observed PCC value. For KSD, the P-value may be calculated as the proportion of sampled permutations whose KSD value is larger than the observed KSD value. If we can find most genes in the upper left corner in the plot have both KSD and PCC P-values very small (*e.g.,* <0.001), we may use this P-value to define the upper left corner. In this study, after the biomarker genes were selected, we may have a visual assessment of the GEP profile specificity between two comparing groups for a given gene, as the example of the plot shown in Figure 1b.

Prediction of Disease State

The GEP histograms of several biomarker genes (as a biomarker set) may be combined to predict the disease state of a patient. The likelihood of a gene being in a normal (or diseased) state may be predicted according to the probability density of the expressed gene intensity in the normal (or diseased) GEP histogram, as shown in Figure 1a. For this purpose, it prefers that the selected biomarker genes are graded, predominantly-on, or predominantly-off genes and the switch-like (bimodal) genes are not appropriate for prediction due to their non-normal distribution, as described in our recent work (19). Graded genes can be expressed in rheostatic levels and states (24-26); predominantly-on genes are in activated high-expression states most of the time (27, 28); predominantly-off genes are in repressed low-expression states most of the time (29, 30); switch-like bimodal genes can switch between repressed and activated states (31, 32).

In practice, the tissue sample of a patient can be analyzed with microarray experiment followed by the same normalization procedure as the one applied to the array data that

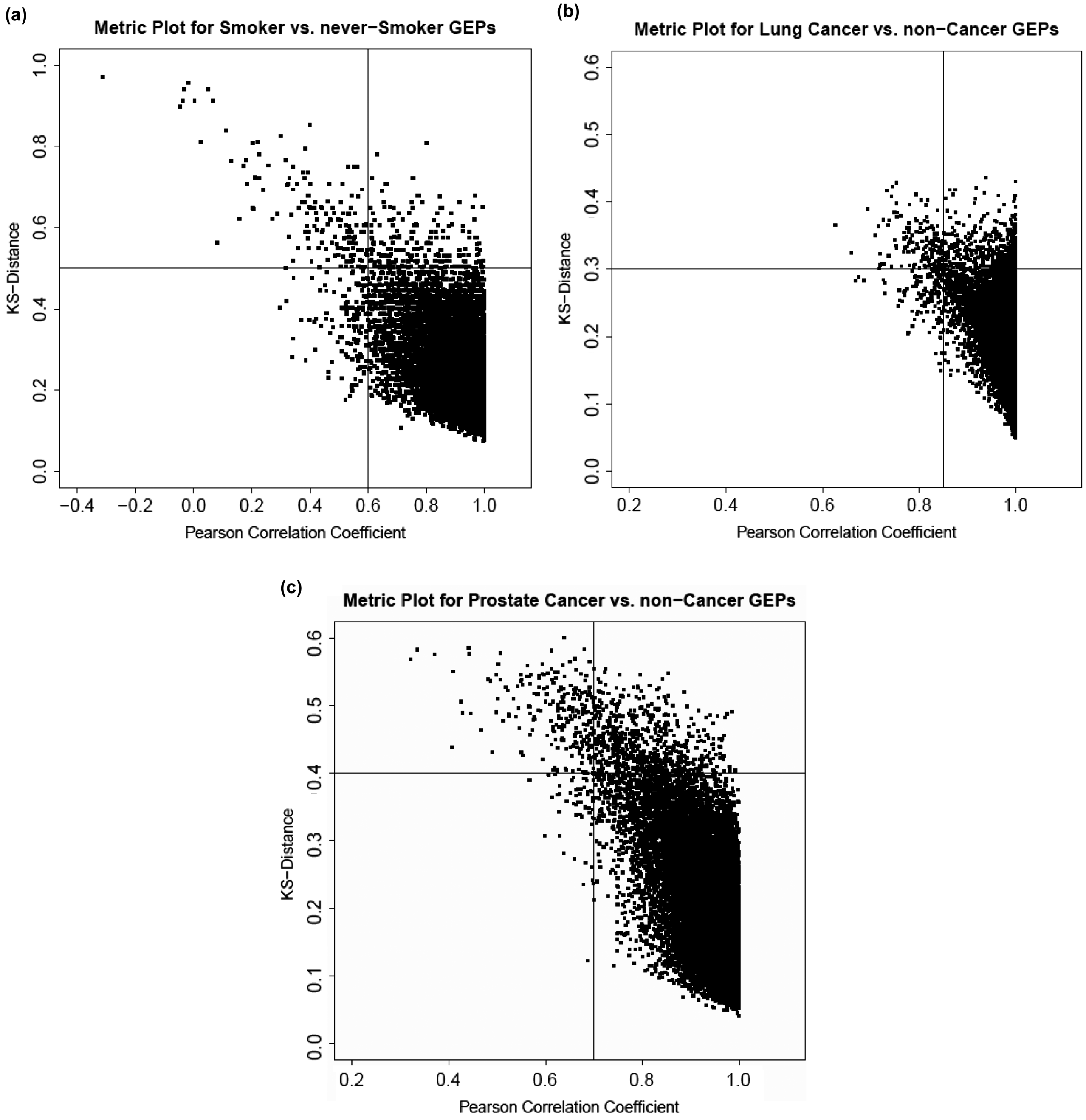

Figure 2: Metric plot of Kolmogorov-Smirnov Distance (KSD) vs. Pearson Correlation Coefficient (PCC) that was used to compare the GEPs of two opposing groups. Each dot on the plot represents a gene. (**a**) The plot for the genes studied in the GDS534 dataset. The GEPs between 34 current-smoker and 23 never-smoker sample groups were compared. KSD > 0.5 and PCC < 0.6 were used to obtain 159 biomarker candidate genes in the upper left square of the plot. (**b**) The plot for the genes studied in the GDS2771 dataset. The GEPs between 97 lung cancer and 90 non-cancer sample groups were compared. KSD > 0.3 and PCC < 0.85 were used to obtain 157 biomarker candidate genes in the upper left square of the plot. (**c**) The plot for the genes studied in the GDS2545, GDS2546, and GDS2547 datasets. The GEPs between prostate cancer and non-cancer sample groups were compared. KSD > 0.4 and PCC < 0.7 were used to obtain 230 biomarker candidate genes in the upper left square of the plot. Note the scale on Y-axis is different in (a) from (b-c).

were used for the construction of GEP probability density histograms for the normal and diseased samples respectively. After this procedure, for each gene in the biomarker set, its normalized expression value in patient can be used to obtain a likelihood value based on the interval it belongs to (on the X-axis) in the expression (probability density) histogram for normal tissues GEP (from normal individuals) and another likelihood value from diseased tissues GEP histogram (from patients with cancer). This step can be automated by probability density matrix representing the data for the histogram described above. The likelihood ratio between cancer and normal states can be normalized to have a sum of 1. Finally, all the normalized likelihood values obtained from normal samples' GEPs for the genes in the biomarker set were averaged to obtain a value indicating how likely the patient's tissue is in normal state. We can also take the average of all the normalized likelihood values obtained from cancer tissues' GEPs for the genes in the biomarker set to obtain another value that indicates how likely the patient's tissue is in the cancer state. The ratio between these two calculated values may help a physician to assess how likely the patient is in normal or cancer state. In real practice, the biomarker set may include several disease-specific biomarker genes to increase the accuracy and confidence in the prediction of disease state. If it succeeds, the method will be promising and powerful for diagnostic and prognostic monitoring on individuals who visit the clinic for screening or evaluation of a specific cancer disease.

Results and Discussion

Biomarker Candidate Genes for the Changes Due to Smoking

Cigarette smoking is the major cause of lung cancer, which is a leading cause of cancer death. Spira *et al.,* (12) had studied the effects of cigarette smoking on the airway transcriptome. We compared the GEPs derived from the current smokers and never-smokers in the dataset GDS534 retrieved from the GEO database using the KSD and PCC metrics described in "Materials and Methods". We used the cutoffs of KSD > 0.5 and PCC < 0.6. Figure 2a displays the 159 genes that satisfied these criteria. They were considered as biomarker candidate genes for smoking.

Among these biomarker candidates, the top ranked genes (*i.e.,* closest to the upper left corner of the square in Figure 2a) were also identified as the most significant genes by conventional method by Spira *et al.,* (12). For example, *GPX2*, *CYP1B1*, *ALDH3A1*, *CEACAM6*, *CX3CL1*, *CA12*, and *SLIT1* were found as top ranked genes in both our study and Spira *et al.,* Among these genes, *GPX2* and *ALDH3A1* function as antioxidants; *CYP1B1* has xenobiotic function; *CEACAM6* is a cell adhesion molecule and putative oncogene; *CA12* is a putative oncogene; expressions for *CX3CL1* and *SLIT1* were both decreased; and *SLIT1* is a putative tumor suppressor gene whose decreasing expression would cause tumor outgrowth.

According to Spira *et al.,* (12), in general, genes whose expression increased in smokers tended to be involved in regulation of oxidant stress and glutathione metabolism, xenobiotic metabolism, and secretion. Our results confirmed this observation. For example, our Gene Ontology (GO) term enrichment tests of the 159 biomarker candidates using the WebGestalt program (33) revealed that the most enriched term is "oxidoreductase activity" under GO principle "molecular function" ($P = 1.04 \times 10^{-13}$, Table I). A number of cytochrome P450 xenobiotic polypeptides (CYP1A1, CYP1B1, CYP27A1, CYP2A13, CYP2W1, and CYP4F11) and several antioxidants were categorized in this GO category (Table I). It is possible that mutations in these proteins might switch the bronchial epithelial cell to malignant state and subsequently cause lung cancer.

We used Ingenuity system (http://www.ingenuity.com/) to identify the perturbed pathways. For those biomarker candidate genes obtained from smoking versus non-smoking analysis, the most significantly enriched pathways were "Metabolism of Xenobiotics by Cytochrome *P450*" ($P = 6.36 \times 10^{-9}$, Fisher's exact test) and "*NRF2*-mediated Oxidative Stress Response" ($P = 6.97 \times 10^{-6}$, Fisher's exact test). These results were consistent with what we expected because *P450* is a well known protein for toxic chemical metabolism in human body while tobacco smoking contains miscellaneous chemical poisons such as acetone and nicotine. Oxidative stress was also found to be associated with smoking (34, 35). These results suggested that the candidate genes identified above might serve as useful biomarkers.

Biomarkers Candidate Genes for Lung Cancer

We studied the second dataset GDS2771 (13) that contained samples from lung cancer smokers and non-cancer smokers. This dataset may provide us important insights on what genes and how they trigger the lung cells to transform into cancerous state among smokers. As seen in Figure 2b, the GEPs between lung cancer and non-cancer samples from smokers were mostly similar. They distributed in a small area near the bottom right corner of the plot, that is, with high PCC and low KSD values. This distribution is remarkably different from that observed in smoking data set (Figure 2a, note the different scale on the Y-axis). This may suggest that smoking has major effect on gene expression changes (*e.g.,* GEPs), which overdominate the effect caused by cancer, as hundreds of poisons in cigarette burning might be toxic for cell development and growth. To further examine those genes having different GEPs in two comparing groups (cancer vs non-cancer smokers), we used relaxed cutoff values (KSD > 0.3 and PCC < 0.85) in order to select about one to two hundreds of candidate genes. This process resulted in 157 genes for further consideration as candidate biomarkers between these two groups (Figure 2b).

For these 157 genes, we performed GO term enrichment tests. Interestingly, the most significant GO term is "response to DNA damage stimulus" ($P = 1.39 \times 10^{-4}$, Table I). DNA damage is one of the mechanisms for cellular response to

Table I
Biomarker candidate genes highly enriched in specific Gene Ontology (GO) category.

Dataset*	GO term†	No. of genes (P value)	Gene symbols
GDS534	MF: oxidoreductase activity	32 (1.04×10^{-13})	ADH7,AKR1B10,AKR1C1,AKR1C2,AKR1C3,ALDH3A1,BDH1,CBR1,CBR3,CYP1A1,CYP1B1,CYP27A1,CYP2A13,CYP2W1,CYP4F11, DHRS3,DUOX2,FMO2,GCLM,GPX2,HGD,HPGD,MAOB,ME1,NQO1, PGD,PRDX1,PRODH,SEPX1,SOD1,TXN,TXNRD1
	BP: electron transport	17 (2.67×10^{-8})	ADH7,AKR1C1,AKR1C3,ALDH3A1,CYP1A1,CYP1B1,CYP27A1,CYP2A13,CYP2W1,CYP4F11,DUOX2,FMO2,MAOB,NQO1,PGD,TXN,TXNRD1
	CC: vesicular fraction	9 (5.63×10^{-6})	CYP1A1,CYP1B1,CYP2A13,CYP4F11,FMO2,UGT1A1,UGT1A3,UGT1A6,UGT1A9
GDS2771	BP: response to DNA damage stimulus	9 (1.39×10^{-4})	BTG2,CDK7,DCLRE1C,FANCF,GTF2H3,GTSE1,MLH1,MSH2,USP1
	CC: nuclear envelope-endoplasmic reticulum network	5 (1.48×10^{-3})	BSCL2,DHCR7,EXT2,SLC33A1,SSR4
GDS2545-2547	MF: cytoskeletal protein binding	18 (4.37×10^{-8})	ACTA1,CALD1,CAPG,CFL2,ENAH,FLNA,FLNC,KLHL5,MSN,PARVA,PRNP,SMTN,SORBS1,SYNPO2,TNS1,TPM1,TPM2,VCL
	BP: muscle contraction	10 (1.57×10^{-6})	CALD1,DES,FXYD1,GJA1,KCNMB1,MYL9,PPP1R12B,SLMAP,SMTN,TPM1
	CC: cytoskeleton	27 (7.53×10^{-8})	ACTA1,ACTC1,BICD1,CALD1,CAPG,CFL2,DES,DMN,ENAH,FLNA,FLNC,KIF20A,KLHL5,KRT15,KRT5,MSN,MYL9,PARVA,PDLIM7,PKP3,SGCB,SMTN,SORBS1,TNS1,TPM1,TPM2,VCL

*Datasets in which biomarker candidate genes were identified. GDS534 was used for "current-smokers vs. never-smokers"; GDS2771 was used for "lung cancer vs. non-cancer smokers"; GDS2545, GDS2546, and GDS2547 were used for prostate cancer analysis.
†Gene Ontology organization principles: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

in vitro carcinogen exposure and it has been found to be associated with lung cancer risk (36).

Molecular Association between Smoking and Lung Cancer

No overlap was found between the 159 biomarker candidate genes for smoking and 157 biomarker candidate genes for lung cancer. We further examined whether these genes share similar biological or subcellular function. We used the Fisher's exact test implemented in the Ingenuity Pathway Analysis (IPA) to find the molecular networks enriched by a set of genes. By using Fisher's exact $P < 0.001$, we found eight molecular networks enriched by the smoking biomarker candidate genes and seven networks enriched by the lung cancer biomarker candidate genes. Comparison of top functions of these molecular networks revealed some common features shared between smoking and lung cancer, such as "cancer", "genetic disorder", and "cell cycle". For smoking, "drug metabolism" is the top function among the eight enriched networks, reflecting the molecular mechanism of these biomarker genes in cellular system.

Figure 3 displays the most significant networks based on smoking and lung cancer biomarker genes, respectively. Interestingly, although no gene overlaps between these two sets, we found a common node*, NF-kB*, shared between these two networks (Figures 3a and b). Previous studies indicated *NF-kB* was frequently expressed in lung cancer (37), and smoking could activate *NF-kB* in human lymphocytes (38). Moreover, another gene, *NFKBIA* (*NF-kB* inhibitor, α), was identified as a biomarker candidate for smoking in our analysis of the GDS534 dataset. *NFKBIA* was down-regulated in the smoker group. Thus, it may potentially increase the risk to lung cancer by increasing the expression of *NF-kB* due to the down-regulation of *NFKBIA* caused by smoking.

Biomarkers Candidate Genes for Prostate Cancer

To demonstrate the power and feasibility of our approach, we explored biomarker genes in another type of cancer. We obtained three prostate cancer-related datasets (GDS2545, GDS2546, and GDS2547) from the same study (17) in the GEO database. For these three datasets, to identify the most distinctive biomarker genes that can be used to identify cancer tissue specifically, we combined the normal prostate tissue and normal tissue adjacent to tumor as one group (*i.e.,* normal sample) and the primary and metastatic prostate tissues as another group (*i.e.,* cancer sample). Using the same strategy and cutoff values KSD > 0.4 and PCC < 0.7, we obtained 230 biomarker candidate genes for further analysis (Figure 2c).

GO enrichment analysis revealed that the prostate biomarker candidate genes were highly enriched in structural molecular activities related to cytoskeleton and muscle

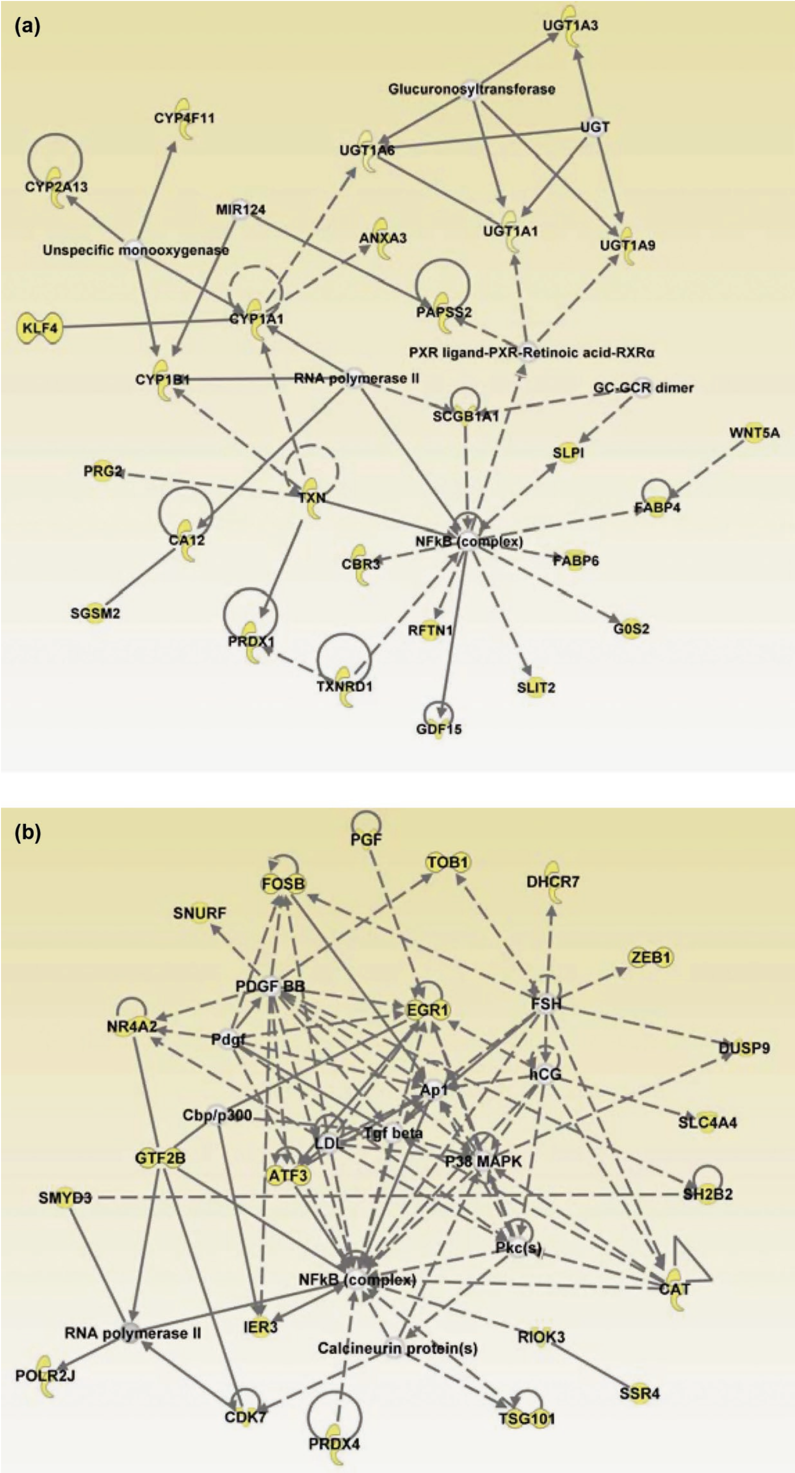

Figure 3: Molecular networks enriched by the biomarker candidate genes identified from (**a**) smoker vs. non-smoker analysis in GDS534 dataset (Figure 2a) and (**b**) lung cancer vs. non-cancer analysis in GDS2771 dataset (Figure 2b). Yellow nodes represent biomarker genes we identified. A solid line indicates a physical interaction, a dashed line with an arrow indicates a regulation relationship, and a solid line with an arrow indicates both a physical interaction and a regulation relationship.

contraction (Table I). Further, network/pathway analysis using the Ingenuity system revealed that the top pathway was "Actin Cytoskeleton Signaling" ($P = 5.46 \times 10^{-5}$, Fisher's exact test). Interestingly, the association between actin cytoskeleton and prostate cancer (PCA) has been documented in literature. For example, Zhang *et al.,* (33) found that *ZNF185* (a candidate PCA biomarker found in this analysis), which was down regulated in prostate cancer, encodes a novel actin-cytoskeleton protein. Papakonstanti *et al.,* (39) reported that the prostate cell line *LNCaP* had functional membrane testosterone receptors which could modify actin cytoskeleton and could increase the secretion of prostate specific antigen (PSA, a biomarker for prostate cancer). When we sorted the biomarker genes by PCC value, the *ACTC1* gene, which encodes an actin subunit, was ranked the first (*i.e.,* lowest PCC value) and *HPN* gene, which encodes hepsin and was previously found to be related to prostate cancer (40), was ranked the second. In the next section, we used these two genes as a biomarker set to explore how disease (cancer) states may be diagnosed and monitored.

Application of the Disease-specific Biomarker Genes for Cancer Diagnosis and Prognosis

As shown above, the disease-specific microarray study is likely to identify biomarker candidate genes that can distinguish the disease (cancer) group from normal group by using gene expression profiles. The top ranked biomarker genes may be applied to the disease-specific study to predict patient's disease (cancer) state according to the gene expression levels of those biomarker genes. That is, when several biomarker genes are discovered as the indicators for a disease (cancer state), the expression values of these genes in a new patient's sample can be obtained and assessed for the diagnosis of the disease, monitoring the patient's disease state, or even be evaluated on how seriously the patient has developed for this disease. We proposed three steps for this application. First, we use normalized array expression intensity to construct the probability density histograms (details in Materials and Methods) for the biomarker genes in normal samples and diseased (cancer) samples, respectively. Second, for a new patient with unknown disease state, his/her tissue sample is taken for microarray experiments followed by the same normalization procedure as in the microarray studies used for probability density histogram construction. The normalized gene expression intensity of the new patient will be used to estimate its probability of the disease or normal state by comparing the probability density histogram constructed in step one. Third, when additional studies with similar designs for the same disease become available, we can combine the previous microarray studies with the new datasets and construct a new version of the probability density histogram. This is an iterative approach, *i.e.,* by repeating step one. We expect the prediction performance will be improved when more microarray data for the same disease are available.

Here, for demonstration purpose, we tested our proposed prediction approach by using two top ranked genes (*ACTC1* and *HPN*) selected from prostate cancer studies using microarray data in GDS2545. First, we constructed probability density histogram for each of these two biomarker genes in normal and prostate cancer samples, respectively (Figure 4). For each gene, its distributions of histogram in normal and cancer samples were different (Figure 4). In the second step, to evaluate the prediction performance, each time we selected one sample in each group (*i.e.,* with known disease state: normal or cancer) to test the likelihood of our prediction, assuming the disease status of the selected sample is unknown. In this prediction, we applied leave-one-out cross validation (LOOCV) method, *i.e.,* the gene expression information of the sample under prediction was not included in the construction of the normal or cancer GEP. Then, each sample was predicted to be in a state (either cancerous or

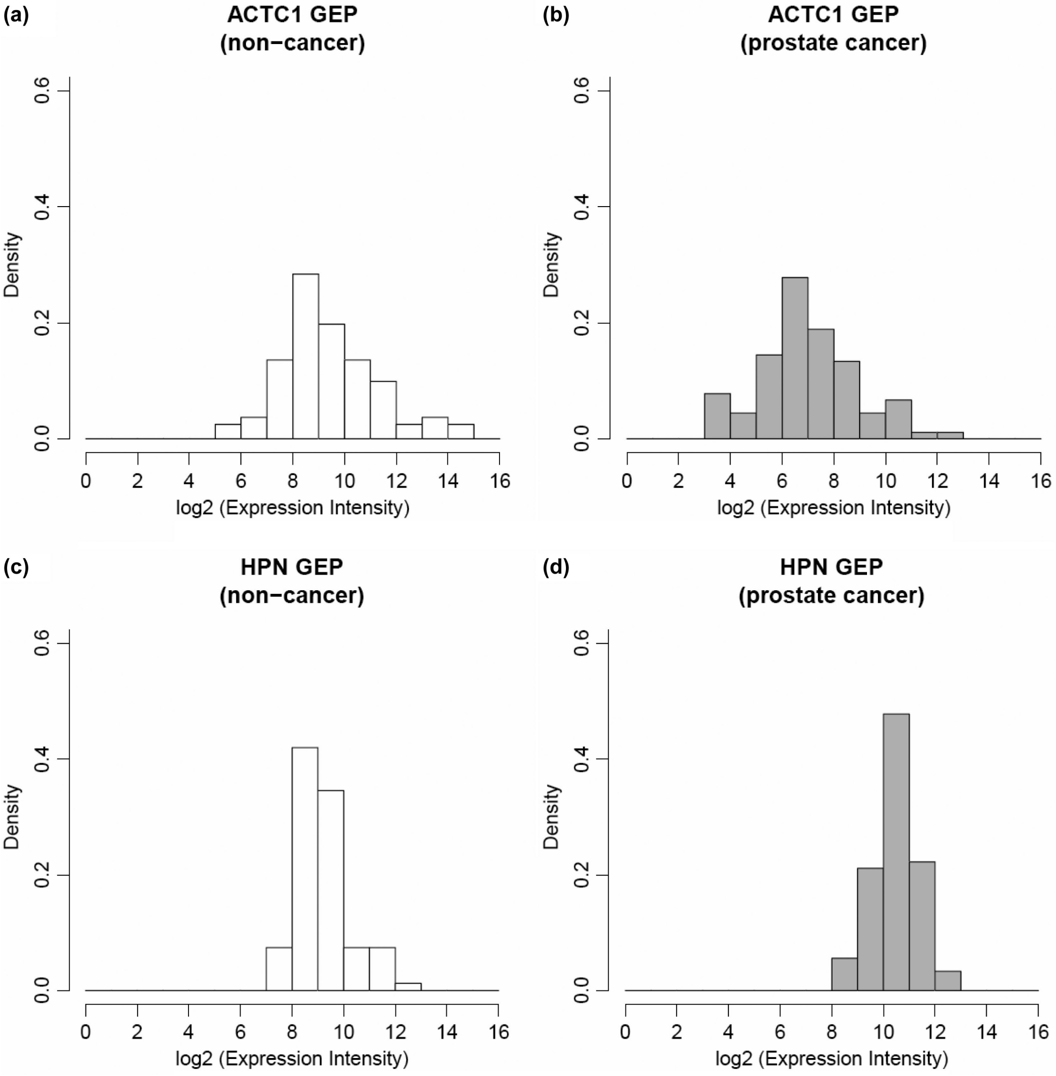

Figure 4: Probability density histograms for biomarker genes *ACTC1* and *HPN* in normal and cancerous samples. The interval (bin size) on X-axis for histogram construction is 1. (**a**) GEP for gene *ACTC1* in normal samples. (**b**): GEP for gene *ACTC1* in cancer samples. (**c**) GEP for gene *HPN* in normal samples. (**d**) GEP for gene *HPN* in cancer samples.

normal) based on the estimated likelihood. Specifically, for the selected sample, we calculated normalized gene expression intensity and then compared the normalize value to the corresponding probability density histogram constructed by using all sample (81 normal and 90 cancer samples, Figure 4) but excluding the sample under prediction.

Table II shows the prediction procedures of one sample in each group. GSM152804 was the first sample in the normal group. When we used marker gene *ACTC1,* the log$_2$(expression intensity) was 9.2. This corresponded to the probability density 0.19 in the histogram for normal samples and 0.04 in the histogram for cancer samples, respectively (Table II and Figure 4). The normalized probability density ratio between normal and cancer groups (N:C) was 0.808:0.192. This ratio indicated that there was strong likelihood of this sample to be normal. The similar likelihood was found by using marker gene *HPN*. Furthermore, when we combined both genes (*ACTC1+HPN*), we had N:C ratio 0.845:0.155. In summary, these two genes could serve as biomarkers to predict this sample to be normal. Similarly, we tested the first sample in the cancer group (GSM152931) in the original microarray sample dataset. Our results clearly indicated that these two genes could predict the cancer status of this sample (details in Table II).

Table III summarized the prediction of each sample in normal and cancer group by predicting one sample each time using the leave-one-out cross validation method. Among the 81 normal samples, 65, 68, and 68 were predicted to be normal according to the prediction by *ACTC1*, *HPN* and "*ACTC1+HPN*", respectively. The prediction by gene *HPN* alone is the same as that by using both genes, both of which had an 84% (cancer) true negative rate assuming the original disease diagnosis in GDS2545 was accurate. Among the 90 cancer samples, 66, 66, and 67 were predicted to be cancerous according to the prediction by *ACTC1*, *HPN*, and "*ACTC1+HPN*", respectively. The prediction by both genes had slightly higher (cancer) true positive rate (67/90) than that by one gene's prediction alone (66/90). When we had a more detailed check of the prediction results, we found a few samples were predicted to have different states by either *ACTC1*or *HPN* gene alone; however, when these two biomarker genes could give out the same disease state prediction, the same prediction was always made by using both genes (*ACTC1+HPN*).

Here, we described in more details two samples whose prediction was opposite by using one gene alone (*ACTC1* or *HPN*); but it gave out the correct prediction when both genes were used altogether. For example, GSM152946, a cancer sample in microarray GDS2545, was not predicted to be cancerous (*i.e.,* cancer negative, denoted by "–" sign in Table III) by *ACTC1*, while correctly predicted by *HPN*. When both genes were recruited, it had the correct prediction (*i.e.,* "+" in Table III). Similarly, GSM152973, another cancer sample, was not predicted cancerous by *HPN*, but was predicted cancerous by *ACTC1* alone and also by both genes. These results suggested that a combined set of biomarker genes may increase the prediction power. Each gene may utilize some positive information in the gene expression signals to help prediction. When more than one gene is used for the disease state prediction, more informative signals might be included so that an overall prediction could be made. However, this is only a putative explanation, as more noise might also be included at the same time.

Bueno *et al.,* (41) described a "3 ratio diagnostic test" on PCA diagnosis. They found a 90% accuracy of PCA diagnosis on the 20 test samples based on the "3 ratio diagnostic test" using the expression level ratios of *C7*-to-*HPN*, *MEIS2*-to-*HPN*, and *FN1*-to-*HPN* which were obtained by real-time quantitative RT-PCR experiments. The datasets and algorithms used for prostate cancer diagnosis are different between Bueno *et al.,* and ours. To have a comparable evaluation of these two methods, we first applied the "3 ratio diagnostic test" to predict the PCA cancer state of our 171 samples in GDS2545 dataset which contained those four genes' expression information. The accuracy turned out to be "70.2%" (120/171) by the "3 ratio diagnostic test", which is lower than our method (*e.g.,* "78.9%" (135/171) using gene pair *ACTC1*/*HPN*). Next, we applied the expression level ratio test in Bueno *et al.*, to our gene pair "*ACTC1*/*HPN*" to predict the PCA cancer state of the 171 samples in GDS2545 dataset.

Table II
Prediction of normal or diseased state on samples GSM152804 and GSM152931 in GDS2545 dataset.

		Probability density		
Gene symbol	*log$_2$(Expr)*[*]	Normal (N)	Cancer (C)	N:C Ratio[†]
Prediction of normal state for GSM152804				
ACTC1	9.2	0.19	0.04	4.21 (0.808:0.192)
HPN	8.7	0.41	0.06	7.40 (0.881:0.119)
ACTC1+HPN [**]				5.45 (0.845:0.155)
Prediction of cancer state for GSM152931				
ACTC1	6.7	0.04	0.27	0.14 (0.121:0.879)
HPN	10.9	0.07	0.47	0.16 (0.136:0.864)
ACTC1+HPN **				0.15 (0.128:0.872)

[*] log$_2$(Expr): normalized log$_2$(expression intensity).
[†] Normalized likelihood ratio between normal and cancer states (to have a total probability of 1).
[**] Prediction was based on both genes by taking the average of the normalized likelihood ratio values.
Note: Prediction was based on the histograms in Figure 4.

Table III
Evaluation of prostate cancer status of the two groups of samples in GDS2545 dataset.

	Prediction of disease state by		
	ACTC1	*HPN*	*ACTC1+HPN*[*]
Group	Number ratio of "normal : cancer" predicted states in each group[†]		
Normal group (81)	65 : 16	68 : 13	68 : 13
Cancer group (90)	24 : 66	24 : 66	23 : 67
Sample	Ratio for "normal : cancer" likelihood (determined state)**		
GSM152946	2.3 (0.697:0.303) [–]	0.16 (0.136:0.864) [+]	0.72 (0.417:0.583) [+]
GSM152973	0.14 (0.121:0.879) [+]	1.71 (0.631:0.369) [–]	0.60 (0.376:0.624) [+]

[*] Prediction was based on both genes by taking the average of the normalized likelihood ratio values.
[†] The ratio is between the "number of predicted normal states" and the "number of predicted cancer states" among each group. Each sample was predicted to be in a state (either normal or cancerous) based on the state with higher predicted likelihood value.
[**] Both samples GSM152946 and GSM152973 belonged to cancer group according to pathological classification. The ratio is for normalized likelihood ratio between normal and cancer possibilities. In parenthesis, "–" denotes predicted normal state and "+" denotes predicted cancer state.

The accuracy was "71.9%" (123/171), lower than that (78.9%) in our method. In principle, an evaluation of a method based on more samples (a total of 171 in our dataset) should be more reliable. Based on the comparison above, our method seems to outperform the expression ratio based method.

In the real world application, we can improve prediction power by the following three ways. First, we may identify more effective biomarker genes and consider them as a biomarker set for disease status prediction. Our results in Tables II and III suggest more than one biomarker genes might increase the accuracy on the prediction of patient's disease (cancer) state. The likelihood ratio predicted by multiple genes between normal and cancer state (*e.g.,* N:C ratio in Table II) can be useful to evaluate the status of a cancer state or monitor the progress during recovery process. Our further analysis of adding another biomarker gene *DMN* to *ACTC1* and *HPN* suggested a better performance in some cases. For example, when these three genes were used together, we found a higher true negative rate in the normal group in the prostate cancer prediction (data not shown). Second, we expect more samples will be helpful for the construction of GEPs, discovery of biomarker genes, and better plotting of the probability density histogram, thus, improving the prediction of the disease state. Since microarray technology becomes matured and available in most molecular biology core facility, we expect more and more studies for specific disease become publicly available. In fact, there have been more than 2000 curated microarray datasets and more than 369,000 samples in the GEO database alone. Third, the diagnosis quality of the samples used in construction of GEPs and probability density histogram is critical. The quality can be improved by better and careful pathological (or other) classifications for the disease states and precise extraction of the biopsy tissues.

Conclusion

In this study, we demonstrated that the biomarker candidate genes for disease-causal changes could be identified by the GEP comparisons between two groups of opposing samples (*e.g.,* cancer vs. normal). Both KSD and PCC are useful metrics for GEP comparison to evaluate the distinctness (KSD) and similarity (PCC) of the expression profiles. KSD vs. PCC metric plot may also provide an overview of the gene expression changes at the genome level in samples and be used to evaluate the magnitude of the distinctions between two comparing groups. We analyzed several real disease-related microarray gene expression datasets. We found that the number of genes with highly different GEPs between comparing groups in smoking dataset was much larger than that in lung cancer dataset; this observation was further verified when we compared GEPs in smoking dataset vs. prostate cancer datasets. We found that both genes in some gene pair, when utilized together, could effectively predict prostate cancer state. Although more work is needed, our results suggested that this approach might prove promising and powerful for diagnosing and monitoring the patients who come to the clinic for screening or evaluation of a disease state including cancer.

Acknowledgments

We thank the reviewers for their valuable suggestions. HCH thanks the Functional Genomics Shared Resource at the Vanderbilt-Ingram Cancer Center for the financial support. This work was partially supported by NIH grant (AA017437), Vanderbilt-Ingram Cancer Center and the SPORE in GI Cancer Grant (50CA95103), and an NARSAD Young Investigator Award (to Z.Z.).

References

1. Miller, M. B., Tang, Y. W. Basic concepts of microarrays and potential applications in clinical microbiology. *Clin Microbiol Rev 22*, 611-633 (2009).
2. Kuderer, N. M., Lyman, G. H. Gene expression profile assays as predictors of distant recurrence-free survival in early-stage breast cancer. *Cancer Invest 27*, 885-890 (2009).
3. Suh, I., Guerrero, M. A., Kebebew, E. Gene-expression profiling of adrenocortical carcinoma. *Expert Rev Mol Diagn 9*, 343-351 (2009).
4. Chambers, D., Lumsden, A. Profiling gene transcription in the developing embryo: microarray analysis on gene chips. *Methods Mol Biol 461*, 631-655 (2008).
5. Pritchard, C., Underhill, P., Greenfield, A. Using DNA microarrays. *Methods Mol Biol 461*, 605-629 (2008).

6. de Snoo, F., Bender, R., Glas, A., Rutgers, E. Gene expression profiling: decoding breast cancer. *Surg Oncol 18*, 366-378 (2009).
7. Reed, C. E., Graham, A., Hoda, R. S., Khoor, A., Garrett-Mayer, E., Wallace, M. B., Mitas, M. A Simple Two-Gene Prognostic Model for Adenocarcinoma of the Lung. *J Thorac Cardiovasc Surg 135*, 627-634 (2008).
8. Nakagawa, T., Kollmeyer, T. M., Morlan, B. W., Anderson, S. K., Bergstralh, E. J., Davis, B. J., Asmann, Y. W., Klee, G. G., Ballman, K. V., Jenkins, R. B. A Tissue Biomarker Panel Predicting Systemic Progression after PSA Recurrence Post-Definitive Prostate Cancer Therapy. *PLoS ONE 3*, e2318 (2008).
9. Hu, P., Greenwood, C. M., Beyene, J. Using the ratio of means as the effect size measure in combining results of microarray experiments. BMC Syst Biol 3, 106 (2009).
10. Jafari, P., Azuaje, F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med Inform Decis Mak 6*, 27 (2006).
11. Tusher, V. G., Tibshirani, R., Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A 98*, 5116-5121 (2001).
12. Spira, A., Beane, J., Shah, V., Liu, G., Schembri, F., Yang, X., Palma, J., Brody, J. S. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci USA 101*, 10143-10148 (2004).
13. Spira, A., Beane, J. E., Shah, V., Steiling, K., Liu, G., Schembri, F., Gilman, S., Dumas, Y. M., Calner, P., Sebastiani, P., Sridhar, S., Beamis, J., Lamb, C., Anderson, T., Gerry, N., Keane, J., Lenburg, M. E., Brody, J. S. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med 13*, 361-366 (2007).
14. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Edgar, R. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res 35*, D760-765 (2007).
15. Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Edgar, R. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res 37*, D885-890 (2009).
16. Edgar, R., Domrachev, M., Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res 30*, 207-210 (2002).
17. Chandran, U. R., Ma, C., Dhir, R., Bisceglia, M., Lyons-Weiler, M., Liang, W., Michalopoulos, G., Becich, M., Monzon, F. A. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer 7*, 64 (2007).
18. Wand, M. P. Data-Based Choice of Histogram Bin Width. *The American Statistician 51*, 59-64 (1997).
19. Huang, H.-C., Jupiter, D., VanBuren, V. Classification of Genes and Putative Biomarker Identification Using Distribution Metrics on Expression Profiles. *PLoS ONE 5*, e9056 (2010).
20. Chakravarti, Laha, Roy. *Handbook of Methods of Applied Statistics*, Vol. I, John Wiley and Sons (1967).
21. Stephens, M. A. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association 69*, 730-737 (1974).
22. Cohen, J., Cohen, P., West, S. G., Aiken, L. S. *Applied multiple regression/correlation analysis for the behavioral sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates (2003).
23. Rodgers, J. L., Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 59-66 (1988).
24. Kutejova, E., Briscoe, J., Kicheva, A. Temporal dynamics of patterning by morphogen gradients. *Curr Opin Genet Dev 19*, 315-322 (2009).
25. Charron, F., Tessier-Lavigne, M. The Hedgehog, TGF-beta/BMP and Wnt families of morphogens in axon guidance. *Adv Exp Med Biol 621*, 116-133 (2007).
26. Nishi, Y., Ji, H., Wong, W. H., McMahon, A. P., Vokes, S. A. Modeling the spatio-temporal network that drives patterning in the vertebrate central nervous system. *Biochim Biophys Acta 1789*, 299-305 (2009).
27. Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M. C., van Asperen, R., Boon, K., Voute, P. A., Heisterkamp, S., van Kampen, A., Versteeg, R. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science 291*, 1289-1292 (2001).
28. Zhou, Y., Luoh, S. M., Zhang, Y., Watanabe, C., Wu, T. D., Ostland, M., Wood, W. I., Zhang, Z. Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res 63*, 5781-5784 (2003).
29. Moehren, U., Eckey, M., Baniahmad, A. Gene repression by nuclear hormone receptors. *Essays Biochem 40*, 89-104 (2004).
30. Eckey, M., Moehren, U., Baniahmad, A. Gene silencing by the thyroid hormone receptor. *Mol Cell Endocrinol 213*, 13-22 (2003).
31. Ertel, A., Tozeren, A. Human and mouse switch-like genes share common transcriptional regulatory mechanisms for bimodality. *BMC Genomics 9*, 628 (2008).
32. Ertel, A., Tozeren, A. Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC Genomics 9*, 3 (2008).
33. Zhang, B., Kirov, S., Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res 33*, W741-748 (2005).
34. Agarwal, R. Smoking, oxidative stress and inflammation: impact on resting energy expenditure in diabetic nephropathy. *BMC Nephrol 6*, 13 (2005).
35. Csordas, A., Wick, G., Laufer, G., Bernhard, D. An Evaluation of the Clinical Evidence on the Role of Inflammation and Oxidative Stress in Smoking-Mediated Cardiovascular Disease. *Biomark Insights 3*, 127-139 (2008).
36. Wu, X., Roth, J. A., Zhao, H., Luo, S., Zheng, Y. L., Chiang, S., Spitz, M. R. Cell cycle checkpoints, DNA damage/repair, and lung cancer risk. *Cancer Res 65*, 349-357 (2005).
37. Tang, X., Liu, D., Shishodia, S., Ozburn, N., Behrens, C., Lee, J. J., Hong, W. K., Aggarwal, B. B., Wistuba, II. Nuclear factor-kappaB (NF-kappaB) is frequently expressed in lung cancer and preneoplastic lesions. *Cancer 107*, 2637-2646 (2006).
38. Hasnis, E., Bar-Shai, M., Burbea, Z., Reznick, A. Z. Mechanisms underlying cigarette smoke-induced NF-kappaB activation in human lymphocytes: the role of reactive nitrogen species. *J Physiol Pharmacol 58 (Suppl 5)*, 275-287 (2007).
39. Papakonstanti, E. A., Kampa, M., Castanas, E., Stournaras, C. A rapid, nongenomic, signaling pathway regulates the actin reorganization induced by activation of membrane testosterone receptors. *Mol Endocrinol 17*, 870-881 (2003).
40. Pal, P., Xi, H., Kaushal, R., Sun, G., Jin, C., Jin, L., Suarez, B., Catalona, W., Deka, R. Variants in the HEPSIN gene are associated with prostate cancer in men of European origin. *Hum Genet 120*, 187-192 (2006).
41. Bueno, R., Loughlin, K. R., Powell, M. H., Gordon, G. J. A diagnostic test for prostate cancer from gene expression profiling data. *J Urol 171*, 903-906 (2004).

Received: November 25, 2009; Revised: January 29, 2010;
Accepted: March 1, 2010