Identification of Biomarkers for Prostate Cancer Prognosis Using a Novel Two-Step Cluster Analysis

Xin Chen[1,⋆], Shizhong Xu[2,⋆], Yipeng Wang[3], Michael McClelland[1,4], Zhenyu Jia[1,⋆⋆], and Dan Mercola[1,⋆⋆]

[1] Department of Pathology and Laboratory Medicine, University of California, Irvine
[2] Department of Botany and Plant Sciences, University of California, Riverside
[3] AltheaDx Inc. San Diego
[4] Vaccine Research Institute of San Diego
xinc6@uci.edu,
shxu@ucr.edu,
ywang@altheadx.com,
mmcclelland@sdibr.org,
zjia@uci.edu,
dmercola@uci.edu

Abstract. Prognosis of Prostate cancer is challenging due to incomplete assessment by clinical variables such as Gleason score, metastasis stage, surgical margin status, seminal vesicle invasion status and preoperative prostate-specific antigen level. The whole-genome gene expression assay provides us with opportunities to identify molecular indicators for predicting disease outcomes. However, cell composition heterogeneity of the tissue samples usually generates inconsistent results for cancer profile studies. We developed a two-step strategy to identify prognostic biomarkers for prostate cancer by taking into account the variation due to mixed tissue samples. In the first step, an unsupervised EM clustering analysis was applied to each gene to cluster patient samples into subgroups based on the expression values of the gene. In the second step, genes were selected based on χ^2 correlation analysis between the cluster indicators obtained in the first step and the observed clinical outcomes. Two simulation studies showed that the proposed method identified 30% more prognostic genes than the traditional differential expression analysis methods such as SAM and LIMMA. We also analyzed a real prostate cancer expression data set using the new method and the traditional methods. The pathway assay showed that the genes identified with the new method are significantly enriched by prostate cancer relevant pathways such as the wnt signaling pathway and TGF-β signaling pathway. Nevertheless, these genes were not detected by the traditional methods.

⋆ Xin Chen and Shizhong Xu are joint first authors.
⋆⋆ Corresponding author.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 63–74, 2011.
© Springer-Verlag Berlin Heidelberg 2011

1 Introduction

Prostate cancer is the most frequently diagnosed male cancer and the second leading cause of cancer death in men in the United States [1]. Majority of the diagnosed cases are "indolent" that may not threaten lives. Radical prostatectomy provides excellent outcomes for patients with localized disease. It is the subsequent disease recurrence and metastatic spread of the cancer that accounts for most of the mortality. In order to improve disease management and benefit for the patients, reliable indicators are needed to distinguish the indolent cancer from the cancer that will progress. This information would guide treatment choices, avoid inappropriate radical prostatectomy and provide guidance to those who may profit from adjuvant therapy in the post-prostatectomy setting, a period that is seldom utilized as a treatment window. Much effort has been made to identify gene expression changes between aggressive cases and indolent cases [2,3,4]. Standard analytical approaches, such as t-test, significance analysis of microarray (SAM) [5] and linear models for microarray data (LIMMA) [6], have been applied to these studies. However, few accepted and clinically employed biomarkers have been developed owing to the lack of consistency among these studies, *i.e.*, the significant gene set identified in one study has little overlap with the significant gene set identified in other studies. Similar phenomenon has been observed in breast cancer research as well [10]. We noted that a major reason accounting for such inconsistency across studies is the heterogeneity in terms of cell composition, *i.e.*, the tissue samples used for assays were usually mixture of various cell-types with varying percentages [11,12,13]. For example, prostate samples are usually composed of tumor, stroma and BPH (benign prostatic hyperplasia). Therefore, the observed gene expression changes among samples could be merely due to the difference in cell composition of these samples. Nevertheless, such composition heterogeneity is rarely taken into account in biomarker studies since there is no straightforward way to deal with such variation through regular gene expression analyses, leading to false discoveries and inaccurate conclusions in individual studies.

In this paper, we first use a linear combination model to integrate cell composition data (obtained by pathological evaluation or in silico method [11]) as shown in our previous studies [11,13]. We then propose a two-step strategy to identify genes that are associated with disease outcomes based on the assumption that the potential prognostic biomarker is able to partition patients into groups of various levels of risk. Step 1: For each gene, unsupervised cluster analysis involving an EM algorithm [15] was used to categorize the subjects (patients) into several groups (for example, 2, 3, or 4 groups) solely based on the expression values for the gene across subjects. Note that the number of groups (C) needs to be specified before implementing the EM algorithm. Models with different C will be fitted with the data. The optimal number C (or the optimal model) is can be determined through the Bayesian information criterion (BIC). Step 2: Chi-square test is utilized to select genes with strong associations between the groups obtained at Step 1 and the frequency of classification of a subject as relapse or non-relapse based on the recorded outcomes. The pool of genes with

significant Chi-square results defines a classifier with potential for predicting the risk of relapse for new subjects.

Our analyses of two simulated data sets using the traditional gene differential expression methods (SAM and LIMMA) consistently indicated that when gene expression data is substantially variable due to mixed cell-type composition, an improved method for identifying cases with high and low risk of relapse ought to be developed. By analyzing a real prostate cancer data set using the new method, we identified 648 genes which categorize prostate cancer cases into two groups of high and low risk for relapse. These genes are significantly enriched in a number of interested pathways disclosed by computer-assisted programs, such as DAVID [14]. Finally the method has potential for accounted for other sources of heterogeneity that bedevil the analysis of Prostate Cancer such as the polyclonal nature of this cancer.

We start the paper with describing the novel two-step approach biomarker identification. The new method was then compared to two commonly used methods by analyzing the simulated data sets as well as a real prostate cancer data set. This is followed by a detailed discussion of the desirable features of the new method and its application domains.

2 Theory and Method

2.1 Step 1: Unsupervised Cluster Analysis

Mixture Model of Gene Expression. Let N be the number of subjects (patients) for a microarray experiment. Let X be a $P \times N$ matrix for the tissue components percentage matrix determined by the pathologists, where P is the number of tissue types included in the model. Define $y = [y_1, \ldots, y_N]^T$ as an $N \times 1$ vector for the observed expression levels (the values transformed from the raw microarray data by logarithms) of a particular gene across N individuals. Note that the Gaussian error model is assumed as below. It is commonly accepted that the raw microarray data are log-normally distributed; therefore, logarithms can adequately transform microarray data to resemble a normal distribution [5,6,7,8,9]. We assume that each subject is sampled from one of C clusters of different levels of risk. The expression of the gene for individual i in the kth cluster ($k = 1, \ldots, C$) can be described by following model:

$$y_i|_{Z_i=k} = X_i^T \beta_k + \epsilon_i, \tag{1}$$

where Z_i is the cluster indicator for the ith individual, β_k is the cell-type coefficient vector ($P \times 1$) for the kth cluster, and $\epsilon_i \sim N(0, \sigma^2)$. Such MLR setting has been used in our previous studies [11,13]. Therefore, the likelihood for the Gaussian mixture can be constructed as follows,

$$L(\Theta; y, Z) = \prod_{i=1}^{N} \prod_{k=1}^{C} [f(y_i|Z_i = k; \beta_k, \sigma^2) P(Z_i = k)]^{\delta(Z_i,k)}, \tag{2}$$

where $\pi_k = P(Z_i = k)$ denotes the prior probability of $Z_i = k$, $\Theta = (\pi_1, \ldots, \pi_C, \beta_1, \ldots, \beta_C, \sigma^2)$ is a vector of parameters, $\delta(Z_i, k)$ is an indicator variable defined as

$$\delta(Z_i, k) = \begin{cases} 1, \text{ if } Z_i = k \\ 0, \text{ otherwise} \end{cases}, \tag{3}$$

and

$$f(y_i | Z_i = k; \beta_k, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - X_i^T \beta_k)^2}{2\sigma^2} \right] \tag{4}$$

is the normal density.

EM Algorithm for Cluster Analysis. The EM algorithm [15] is a numerical algorithm commonly used for clustering analyses. Generally, the EM algorithm consists of two major routines: an expectation (E) routine in which the expectation of the log-likelihood is computed, and a maximization (M) routine in which the MLE is calculated for each parameter. The E-routine and M-routine alternate until some criterion of convergence is reached. In the current study, the EM algorithm is implemented as follows,
Procedure 0: Initializing all parameters, $\Theta = \Theta^{(0)}$.
Procedure 1 (E): Update the posterior membership probability,

$$\pi_{ik} = E(\delta(Z_i, k)) = \frac{\pi_k f(y_i | Z_i = k; \beta_k, \sigma^2)}{\sum_{k'=1}^{C} \pi_{k'} f(y_i | Z_i = k'; \beta_{k'}, \sigma^2)} \tag{5}$$

Procedure 2 (M1): Update the mixing probability,

$$\pi_k = \sum_{i=1}^{N} \pi_{ik} \tag{6}$$

Procedure 3 (M2): Update the expression coefficient,

$$\beta_k = \left(\sum_{i=1}^{N} \pi_{ik} X_i X_i^T \right)^{-1} \left(\sum_{i=1}^{N} \pi_{ik} y_i X_i \right) \tag{7}$$

Procedure 4 (M3): Update the residual variance,

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{C} \pi_{ik} \left(y_i - X_i^T \beta_k \right)^2 \tag{8}$$

Procedure 5: Repeat procedure 1 through procedure 4 until a certain criterion of convergence is reached.

2.2 Step 2: Gene Identification Based on Correlation Analysis

The posterior probabilities $(\pi_{i1}, \ldots, \pi_{iC})$ are calculated for subject i where $i = 1, \ldots, N$. The $\tau_i = \max(\pi_{i1}, \ldots, \pi_{iC})$ determined the membership for subject i given $\tau_i \geq 0.6$. If $\tau_i < 0.6$, the subject cannot be assigned to any of the C clusters. In this study, we only consider genes that definitively determine the membership for $\geq 60\%$ of the subjects. Therefore, genes that do not meet this criterion will be considered irrelevant to the disease progression. A contingency table is constructed for each gene based on the membership indicator obtained by cluster analysis and the observed clinical outcomes, *i.e.*, relapse or non-relapse. A chi-square test is then performed for each contingency table to evaluate the association between these two variables. Genes that are strongly associated with the observed outcome are selected if their p-values are less than a pre-selected cutoff point, *e.g.*, 0.001, where p-values are calculated from the central χ^2 distribution.

3 Application

3.1 Simulation Study

To demonstrate the efficiency of the new method, we carried out two simulation studies each with 1000 genes and 200 subjects. Among the 200 subjects, 1 to 100 were designated as non-relapse patients and 101 to 200 were designated as relapse patients. We randomly generated three tissue types (tumor, stroma and BPH) with various component percentages for each subject. In the first simulation, the first 150 genes were set to be associated with the disease outcomes. The expression coefficients $\beta_{intercept}$, β_{tumor}, β_{stroma} and β_{BPH} for each gene represent the strength of gene-outcome association. They were generated as follows: (1) $\beta_{intercept}$ and β_{BPH} were not altered between relapse and non-relapse patients; (2) for genes 1 - 50, the relapse and non-relapse patients had different β_{tumor} but the same β_{stroma}, *i.e.*, these genes are differentially expressed in tumor between relapse and non-relapse groups (tumor related signatures); for genes 51 - 100, the relapse and non-relapse patients had different β_{stroma} but the same β_{tumor}, *i.e.*, these genes are differentially expressed in stroma between relapse and non-relapse groups (stroma related signatures); (3) for genes 101 - 150, both β_{tumor} and β_{stroma} were different between the relapse and non-relapse patients, *i.e.*, these genes are differentially expressed in both tumor and stroma between the relapse and non-relapse groups (tumor and stroma related signatures). Table 1 (column 1) shows the expression coefficients for the first 150 genes. We deliberately generated random dichotomous variables for the 200 patients. These random dichotomous variables are similar to but unrelated to the outcome variable. We let genes 151 - 300 to be associated with these simulated dichotomous variables, yielding a special group of control genes. The remaining 700 genes were set to have no association with any indicator variables. Residual errors were sampled from a normal distribution with mean 0 and variance 0.01. The two-cluster EM algorithm was performed for each gene as described below.

First, the 200 subjects were randomly (0.5 : 0.5) assigned into the relapse and non-relapse groups. A simple regression analysis was used to generate the two initial expression coefficients sets $\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}$ and $\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}$ for these two groups. The average of the estimated residual errors for the 200 subjects was used as the initial value of σ^2. The EM algorithm was then applied until the difference between successive log-likelihoods was smaller than 0.0001. We constructed a 2×2 contingency table and calculated the p-value based on the central χ^2 distribution. Using the cutoff point ($p < 0.001$) mentioned in Theory and Method, we detected 148 differentially expressed genes with no false detection. The two missing genes were gene 5 and 29, owing to the fact that less than 60% of the 200 subjects had been definitively clustered. However, the p-values of the χ^2 test for the two genes are still significant for the successfully clustered subjects.

Table 1. The number of the genes (out of the first 150 genes) identified by SAM, LIMMA, and the new method in the first simulation experiment. The first column represents the simulation scenarios: (+) and (-) represents positive and negative expression coefficient, respectively.

Gene		SAM	LIMMA	New method
1-30	Tumor(+)	23	21	28
31-50	Tumor(-)	14	16	20
51-80	Stroma(+)	30	30	30
81-100	Stroma(-)	17	18	20
101-110	Tumor(+) Stroma(+)	10	10	10
111-130	Tumor(+) Stroma(-)	3	3	20
131-140	Tumor(-) Stroma(-)	10	10	10
141-150	Tumor(-) Stroma(+)	4	6	10
1-150		111	114	148

We also analyzed the simulated data using two commonly used differential expression analysis software packages, SAM [5] and LIMMA [6]. False discovery rate (FDR) ≤ 0.05 was chosen as the criterion for gene detection. SAM identified 111 differentially expressed genes with no false detection. LIMMA identified 114 differential expressed genes with no false detection. Table 1 (columns 3, 4 and 5) shows the numbers of genes identified by SAM, LIMMA, and the proposed new method, suggesting the new method has improved the accuracy for gene identification over the two existing methods. Another advantage of the new method is that it can identify the cell origin of the observed expression changes, *i.e.*, the new method can determine if such expression changes occurred in tumor cells or stroma cells. The contingency test in Table 2 shows that our method can successfully identify 98% tumor-cell-related signatures, 100% stroma-related signatures, and 80% tumor-stroma related signatures. Ten tumor-stroma related genes were misclassified into stroma related genes. There are two possible explanations: (1) For these ten genes, the stroma expression coefficients (average 1.28) are much greater than the tumor expression coefficients (average 0.63), and (2)

The stroma percentage data are more variable than the tumor percentage data (about 50% of the simulated subjects had 0% tumor cells). Thus this simulation study reveals several advantages of the new method and highlights the improved efficiency.

Table 2. The performance of the new method in the first simulation experiment

# of estimate \ # of true	Tumor	Stroma	Tumor+stroma	Null
Tumor	48	0	0	0
Stroma	0	50	10	0
Tumor+stroma	0	0	40	0
Null	2	0	0	700

The second simulation experiment was based on the assumption that subtypes of tumors which are possibly polyclonal variants may exist for relapse patients, *i.e.*, different types of tumors may be subject to different levels of recurrence risk. In this simulation experiment, we divided relapse patients into low risk (subjects 101 - 150) and high risk relapse patients (subjects 151 - 200). We used the same X matrix as in the first simulation experiment. We let the first 150 genes (1-150) be associated with non-relapse/relapse status (2 clusters) but the next 150 genes (151-300) be associated with non-relapse/low risk/high risk status (3 clusters). The clusters were assigned different expression coefficients. Genes 301 - 350 and 351 - 450 were controlled by other randomly generated dichotomous and trichotomous variables respectively. The last 550 genes were not association with any indicator variables. The residual errors were sampled from normal distribution with mean 0 and variance 0.01. We used the two-cluster EM algorithm (EM2) and the three-cluster EM algorithm (EM3) to analyze the simulation data. A similar strategy was used for initializing parameters for each model. Comparison of these two models is shown in Figure 1. For genes 1 - 150, both EM2 and EM3 identified 145 of them, but each method identified two additional genes, making a shared proportion of 97.3% (see Figure 1(a)). Among the 149 identified genes, 132 genes were selected by the EM2 algorithm (2-cluster model) and the remaining 17 genes were selected by the EM3 algorithm (3-cluster model) based on the BIC scores (Model with lower BIC score was preferred.). For genes 151-300, 138 genes were identified by both the EM2 and EM3 algorithms (shared proportion of 92.6%, see Figure 1(b)). The two algorithms together identified 149 genes, among which 133 genes were selected by EM3 (3-cluster model) and 16 genes were selected by EM2 (2-cluster model) based on the BIC scores. These results are consistent with the setting of the simulation experiment, which shows that subtypes within a risk group may be identified by increasing the number of clusters in EM algorithm.

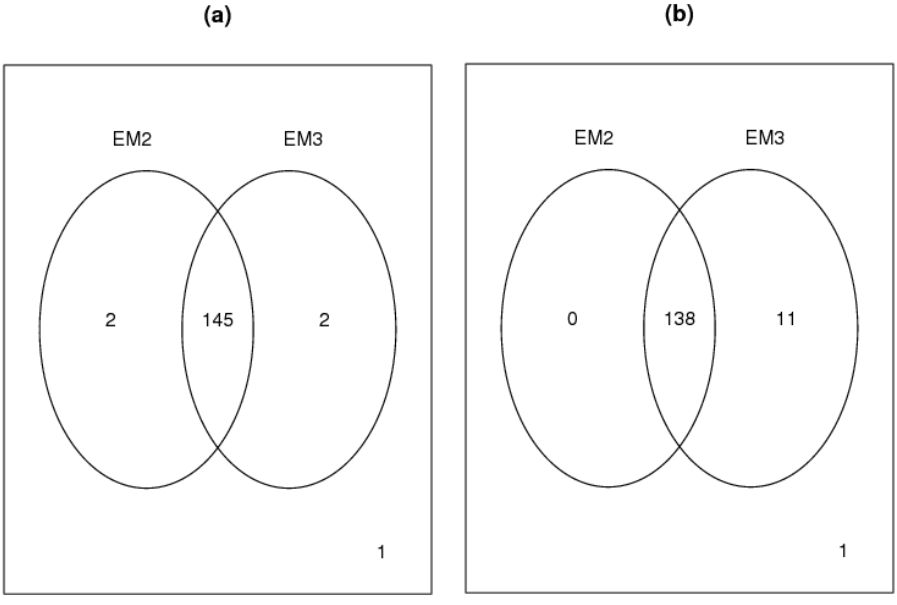

Fig. 1. The number of genes detected by EM2 and EM3 algorithms in the second simulation experiment. (a)Venn diagram of the detected genes for gene 1 - 150. For this analysis, EM2 identified 147 (2 + 145) genes and EM3 identified 147 (2 + 145) genes, with 145 genes identified by both models. Only 1 gene (bottom-right corner) has been missed by two models. (b)Venn diagram of the detected genes for gene 151 - 300.

3.2 Real Data Analysis

A publicly available prostate cancer data set [12] was analyzed here. A total 136 postprostatectomy frozen tissue samples were obtained from 82 subjects by informed consent using Institutional Review Board (IRB)-approved and HIPPA-compliant protocols. All tissues were collected at surgery and escorted to pathology for expedited review, dissection and snap freezing in liquid nitrogen. The tissue components (tumor epithelial cells, stroma cells, epithelial cells of BPH) were estimated by four pathologists. RNA samples prepared from the frozen tissue samples were hybridized to Affymetrix U133A GeneChip array. The data have been deposited in the Gene Expression Omnibus (GEO) database with accession number GSE08218. Out of the 136 samples, 80 samples were from relapsed patients, 50 samples from non-relapsed patients, and 6 samples from normal subjects.

The EM2 and EM3 algorithms were applied to the prostate cancer data following the same strategy as in the simulation studies. We only used 130 patient samples for the analysis. The cutoff point we used for gene selection was p-value ≤ 0.005 (χ^2 test), which resulted in 648 detected genes (215 genes by EM2, 324

genes by EM3, and 109 genes by both algorithms). A computer-assisted analysis of these 648 genes using DAVID Bioinformatics [14] indicated that these genes are significantly enriched in several prostate cancer pathways, such as ECM-receptor interaction, wnt signaling pathway, focal adhesion, TGF-β signaling pathway and gap junction (see Table 3). For comparison, we also analyzed the same data using SAM and LIMMA. The 650 highly ranked genes (similar in size to the detected genes by the new method) in each analysis were selected and then analyzed using DAVID Bioinformatics. The pathway analysis of the genes from SAM and LIMMA did not pull out relevant connections between these genes from literatures.

4 Discussion

We have developed a new gene differential expression method for identifying reliable prognostic signatures for prostate cancer. The new method consists of two steps - cluster analysis and correlation analysis. Two major contributions of the new method include (1) the use of cell-type distribution information, and (2) avoiding direct use of relapse/non-relapse states information, which is often not definitive due to data censoring. It is very common in the tumor marker literature that different labs produce inconsistent results [16,17]. One possible explanation among others is that tumor samples used for gene expression assays are highly heterogeneous in terms of cell composition. We have solved the problem by incorporating tissue component percentage into the analysis. Another advantage of incorporating cell-type distribution into the data analysis is that we were able to identify the cell origin of the observed gene expression changes. This provides a clue for identification of desirable therapeutic targets or better understanding the mechanism of cancer biology.

By using two sets of simulated data, we demonstrated that the new method is more desirable than the commonly used differential expression methods when the samples are highly heterogeneous in cell composition. Moreover, the new method can also identify the potential subtypes of tumors based on the gene expression profiles (represented by the genes detected in EM3 model). Given larger sample size, more complex models, for example EM4 or EM5, could be applied to identify finer tumor subtypes represented by special expression signatures. It has been hypothesized that these subtypes of tumor likely arise from the polycolonal nature of prostate cancer and they may have distinct potentials to progress [13,18,19,20]. Therefore, the gene signatures that are distinctive of these subtypes of prostate tumors may be useful for predicting patients' clinical outcomes. The pathway analysis based on the real prostate cancer data demonstrated that the new method identified genes significantly enriched or associated with prostate cancer related pathways. Genes selected by the traditional methods, however, did not show relevant connectivity in biological functions. The failure of the traditional methods might be explained by (1) the substantial variation caused by mixed tissue or (2) the incomplete recorded relapse/non-relapse data due to censoring.

Table 3. Pathway analysis of detected genes by three methods

EM clustering	SAM	LIMMA
ECM-receptor interaction	Huntington's disease	Huntington's disease
Wnt signaling pathway	Oxidative phosphorylation	Oxidative phosphorylation
Focal adhesion	Pyruvate metabolism	Pyruvate metabolism
TGF-β signaling pathway	Alzheimer's disease	Alzheimer's disease
Lysosome	Parkinson's disease	Parkinson's disease
RIG-I-like receptor signaling pathway	Valine,leucine, and isoleucine degradation	Valine, leucine, and isoleucine degradation
Biosynthesis of unsaturated fatty acids	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis
Ribosome	Fatty acid metabolism	Fatty acid metabolism
Regulation of autophagy	Cardiac muscle contraction	Cardiac muscle contraction
Gap junction	Propanoate metabolism	Tryptophan metabolism
Tryptophan metabolism		Terpenoid backbone biosynthesis
Oocyte meiosis		Endocytosis
Dilated cardiomyopathy		

In this study, we aim to identify expression changes that are associated with bad outcomes of prostate cancer and hope to translate these gene signatures to clinical use. Therefore, genes that do not contribute to accurate classification may not be of clinical use even though they may be biologically related to the disease. One reviewer suggested further Gene Ontology study or Enrichment analysis of the genes that have not been selected in gene identification process owing to the fact that they do not meet the membership criterion in step (2). We agree that such analysis would help gain valuable information conveyed by those "ignored" genes. We will definitely look into the biological functions of those genes in the subsequent biological validation. However, these efforts do not improve the performance of the classifier that we are trying to develop in the study.

To deal with uncertain issue related to data censoring, we first developed an unsupervised EM clustering method to cluster samples and then constructed a contingency table between our cluster variables and the recorded outcome variables (relapse/non-relapse) to select genes. Our new method avoided the direct use of recorded relapse/non-relapse information to identify reliable signatures for disease prognosis. To validate that the unsupervised clustering method is superior to the supervised clustering, we did another experiment. The classification obtained for the 113 significant genes (p-value ≤ 0.001 in the χ^2 test) were used for deriving the outcome variables for the 130 patient samples *via* majority voting, yielding 16 non-relapse samples and 114 relapse samples. We then simply reanalyzed the prostate cancer data using SAM and LIMMA with the outcome variable estimated from the 113 genes. The pathway analysis based on the most significant 650 genes from the SAM and LIMMA analyses identified relevant pathways, *e.g.*, the wnt signaling pathway from both methods. Other important cancer related pathways, such as ECM-receptor interaction, focal adhesion and

TGF-β signaling pathway were identified by SAM using the "corrected" outcome information. We conclude that the new method has potential for classification of prostate cancer into risk groups for relapse and non-relapse outcome and is worthy of further development. In particular it will be important to test the method by comparison to other methods like the forest of trees approach [21] which does not require cell-type express and by testing the method on additional real data sets with known or calculated cell type distributions. These studies are in progress.

Acknowledgments. This work was supported by NIH grants U01 CA114810-01 and UO1 CA152738-01.

References

1. A.C.S: American Cancer Society: Cancer Facts and Figures 2011 [online] (2011)
2. Barwick, B.G., Abramovitz, M., Kodani, M., Moreno, C.S., Nam, R., Tang, W., Bouzyk, M., Seth, A., Leyland-Jones, B.: Prostate cancer genes associated with TMPRSS2-ERG gene fusion and prognostic of biochemical recurrence in multiple cohorts. Br. J. Cancer 102, 570–576 (2010)
3. Bibikova, M., Chudin, E., Arsanjani, A., Zhou, L., Garcia, E.W., Modder, J., Kostelec, M., Barker, D., Downs, T., Fan, J.B.: Expression signatures that correlated with Gleason score and relapse in prostate cancer. Genomics 89, 666–672 (2007)
4. Bickers, B., Aukim-Hastie, C.: New molecular biomarkers for the prognosis and management of prostate cancer-the post PSA era. Anticancer Res. 29, 3289–3298 (2009)
5. 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)
6. Smyth, G.K.: Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. 3, Article3 (2004)
7. Ibrahim, J., Chen, M.-H., Gray, R.: Bayesian models for gene expression with dna microarray data. J. Am. Stat. Assoc. 97, 88–99 (2002)
8. Ishwaran, H., Rao, J.: Detecting differentially expressed gene in microarrays using bayesian model selection. J. Am. Stat. Assoc. 98, 438–455 (2003)
9. Lewin, A., Bochkina, N., Richardson, S.: Fully Bayesian mixture model for differential gene expression: Simulations and model checks. Stat. Appl. Genet. Mol. Biol. 6, 1–36 (2007)
10. Fan, C., Oh, D.S., Wessels, L., Weigelt, B., Nuyten, D.S., Nobel, A.B., Van't Veer, L.J., Perou, C.M.: Concordance among gene-expression-based predictors for breast cancer. N. Engl. J. Med. 355, 560–569 (2006)
11. Wang, Y., Xia, X.Q., Jia, Z., Sawyers, A., Yao, H., Wang-Rodriquez, J., Mercola, D., McClelland, M.: In silico Estimates of Tissue Components in Surgical Samples Based on Expression Profiling Data. Cancer Res. 70, 6448–6455 (2010)
12. Stuart, R.O., Wachsman, W., Berry, C.C., Wang-Rodriguez, J., Wasserman, L., Klacansky, I., Masys, D., Arden, K., Goodison, S., McClelland, M.: In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. Proc. Natl. Acad. Sci. U. S. A. 101, 615–620 (2004)

13. Jia, Z., Wang, Y., Sawyers, A., Yao, H., Rahmatpanah, F., Xia, X.Q., Xu, Q., Pio, R., Turan, T., Koziol, J.A.: Diagnosis of prostate cancer using differentially expressed genes in stroma. Cancer Res. 71, 2476–2487 (2011)
14. Dennis Jr, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., Lempicki, R.A.: DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 4, R60 (2003)
15. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. 39, 1–38 (1977)
16. Woodson, K., Tangrea, J.A., Pollak, M., Copeland, T.D., Taylor, P.R., Virtamo, J., Albanes, D.: Serum insulin-like growth factor I: tumor marker or etiologic factor? A prospective study of prostate cancer among Finnish men. Cancer Res. 63, 3991–3994 (2003)
17. Xu, L., Tan, A.C., Naiman, D.Q., Geman, D., Winslow, R.L.: Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics 21, 3905–3911 (2005)
18. Sutcliffe, P., Hummel, S., Simpson, E., et al.: Use of classical and novel biomarkers as prognostic risk factors for localised prostate cancer: a systematic review. Health Technol. Assess 13, 5 (2009)
19. Mucci, L.A., Pawitan, Y., Demichelis, F., et al.: Testing a multigene signature of prostate cancer death in the Swedish Watchful Waiting Cohort. Cancer Epidemiol. Biomarkers Prev. 17, 1682–1688 (2008)
20. Tomlins, S.A., Bjartell, A., Chinnaiyan, A.M., et al.: ETS gene fusions in prostate cancer: from discovery to daily clinical practice. Eur. Urol. 56, 275–286 (2009)
21. Díaz-Uriarte, R., de Andrés, A.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)