Identifying Gene Signatures from Cancer Progression Data Using Ordinal Analysis

Yoon Soo Pyon and Jing Li*
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, OH 44106
{yoon.pyon, jingli}@case.edu

***Abstract*—A comprehensive understanding of cancer progression may shed light on genetic and molecular mechanisms of oncogenesis, and it may provide much needed information for effective diagnosis, prognosis, and optimal therapy. However, despite considerable effort in studying cancer progressions, their molecular and genetic basis remains largely unknown. Microarray experiments can systematically assay gene expressions across genome, therefore they have been widely used to gain insights on cancer progressions. In general, expression data may be obtained from different stages of the same samples. More often, data were obtained from individuals at different stages. Existing methods such as the Student's t-test and clustering approaches focus on identification of differentially expressed genes in different stages, but they are not suitable for capturing real progression signatures across all progression stages. We propose an alternative approach, namely a multicategory logit model, to identify novel genes that show significant correlations across multiple stages. We have applied the approach on a real data set concerning prostate cancer progression and obtained a set of genes that show consistency trends across multiple stages. Further analysis based on Gene Ontology (GO) annotations, protein-protein interaction networks and KEGG pathways databases, as well as literature search demonstrates that our candidate list not only includes some well-known prostate cancer related genes such as MYC and AMACR, but also consists of novel genes (e.g. CKS2)that have been confirmed by very recent independent studies. Our results illustrate that ordinal analysis of cancer progression data has the potential to obtain a set of promising candidate genes. Such a list can be further prioritized by combining other existing biomedical knowledge to identify therapeutic targets and/or biomarkers of cancer progressions.**

Keywords-ordinal analysis; gene expression; cancer progression

I. INTRODUCTION

Understanding the biology of disease progression at the genetic and molecular level may provide much needed information for effective diagnosis and optimal therapy. However, the biological and genetic basis of cancer progression is usually poorly understood. The past few years have witnessed tremendous interests in investigating genetic signatures of cancer progression using high-throughput gene expression experiments [1,6,8,18]. Some commonly used approaches include clustering analysis to group genes with similar expression profiles, ANOVA (analysis of variance) analysis to compare gene expressions across stages, and t-test to identify differentially expressed gene in the different stages. For example, a two-sample t-test has been used in analyzing prostate [*6*] and breast [1] cancer progressions. Clustering methods (such as principal component analysis and ensemble consensus clustering) have been used in investigating portraits of breast cancer progression [*8*]. Most of the previous studies focused on discovering the most up- or down-regulated genes within each stage, or discovering genes with different expressions across two different stages. Therefore, such analyses can only obtain different sets of genes from different stages that show significant over or under expressions, but neglect an important feature from progression data that cancer stages develop in an ordinal fashion. In contrast, our hypothesis is that genes with expression levels showing concordance or discordance with cancer development stages are more likely to reveal cancer progressions, and they may sever as genetic biomarkers for diagnosis, prognosis and selection of treatments. Therefore, our goal is to identify one set of such genes that will show consistently increasing/ decreasing expressions with the cancer progression and development.

Notice that in most cases, expression data on cancer progression are different from time-series expression data because studies usually collect different samples from different stages. Though it is expected that in general there should be an ordinal relationship in expression levels of genes responsible for cancer progression, real observations can be very noisy and one cannot directly apply techniques used in time-series data analysis. In this study, we propose an alternative approach to address this issue and compare its performance with a few commonly used approaches. We use a multi-category logit model that takes ordinal response with continuous explanatory variables. Logit models have been commonly used for binary response variables such as disease status. Here we use its extension to ordinal response variables with multiple levels because it can capture the natural ordinal relationship among development stages. Different from a commonly used statistical technique Analysis of Variance (ANOVA), which takes stages as explanatory variables and compares the expression levels in different stages, logit models take progression stages as response variables and gene expression as explanatory variables. *T*omlins et al [18] is one of few studies that also considered the ordinal relationship among stages though the authors mainly emphasized the "concept analysis" in their paper. The authors stated that they used Pearson's correlation for their multiclass ordinal analyses but no details were given in their paper. We suspect that their approach is equivalent to a modified $\chi 2$ test of independence that also incorporates the order of each class. (But our results turn out that results from these two approaches are quite different.) A major difference between the modified $\chi 2$ test and the ordinal logit model is that the $\chi 2$ test only takes categorical inputs.

Therefore valuable information can be lost during the discretization step of microarray data.

We use false discovery rate (Q-value) as a control of significance level. From a total of 20,000 probes, we identify a set of robust progression signatures comprising genes whose expression increase or decrease during the progression. We further investigate the gene lists based GO enrichment analysis, and results show that the set of genes with increased expressions returned by the logit model is extremely enriched with genes related to "cell cycles", while none significant GO terms have been found from the set returned by the χ^2 test. We also examine the distribution of the genes in existing biological networks (protein-protein interactions and pathways). Results show that these genes are more tightly connected than random in these networks.

The remainder of the paper is organized as follows. We discuss our methods in Section 2. We then apply these methods to a data concerning prostate cancer progression obtained from [18] and present our results in Section 3. We conclude our work in Section 4.

II. Materials and Methods

A. Data Source and Preprocessing

We first obtained normalized gene expression data of prostate cancer progression [18] from Gene Expression Omnibus at NCBI (GSE6099). The data set contains 84 cell populations from four different stages (22 are benign epithelium, 13 are prostatic intraepithelial neoplasia (PIN), 32 are prostate cancer (PCA) and 17 are metastatic prostate cancer). The expression profile of each cell population consists of 20,000-cDNA microarrays. We first filtered the data using two criteria. Probes with low values (less than the 10th percentile of all values) and probes with small variances (less than the 10th percentile of all the variances) were removed before further analysis. The quality of data with low expression values is usually bad due to large quantization errors or poor spot hybridization. Genes with small variances could be housekeeping genes which are out of our interest. After this filtering step, the number of cDNA probes was reduced to 17,904.

B. Multicategory Logit Model for Ordinal Response

The goal of the study is to identify robust progression signatures showing consistent increasing or decreasing patterns across cancer stages. We take cancer stages as the response variable and gene expressions as the explanatory variable. An ordered categorical response such as cancer progression stages, can be analyzed using the multicategory logit model for ordinal response [2], also known as the proportional odds model. Briefly, suppose that a certain cancer has 1 to J progression stages. The probability of an individual is in stage j is denoted as π_j. The cumulative probability of a response Y less than j ($1 \le j \le J$-1) is:

$$P(Y \le j) = \pi_1 + \pi_2 + ... + \pi_j = \sum_{h=1}^{j} \pi_h .$$

Then the cumulative logit is defined as

$$\log\left(\frac{P(Y \le j)}{P(Y > j)}\right) = \log\left(\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_J}\right).$$

It mimics a binary logistic regression model while stages 1 to j form one new category and stages j+1 to J form a new category. The cumulative log odd is then explained by a linear combination of the explanatory variable (*i.e.*, expression level of a gene) x for each j, where $1 \le j \le J$-1.

$$L_j = \log\left(\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_J}\right) = \alpha_j + \beta x$$

Notice that the model shares a common coefficient for the predictor variable across j with different intercepts. Essentially it assumes that the effect of x is identical for all J-1 cumulative logits. More detailed treatment about the model can be found in [2]. Many statistical tools such as Matlab provide functions for the above model under generalized linear models. Parameters including α's and β can be estimated through numerical methods. Statistical significance for the null hypothesis H_0: β=0 is reported using a one-sample t-test (which is equivalent to the Wald test).

It is well known that for gene expression analysis, one has to correct the significance level due to the multiple testing problem. A commonly used solution is the false discovery rate (FDR) Q instead of the p-value based on single tests. In this study, we use the Benjamini-Hochberg procedure [5] to obtain Q-values based on p-values. Basically, probes are ordered based on their p-values in the increasing order, and the Q-value of each probe is just its p-value weighted by the ratio of the total number of probes and its rank. Theoretical justifications about FDR can be found in [17]. A threshold of 0.05 is used to identify gene signatures with increasing (up-expressed) or decreasing (down-expressed) expressions during cancer progression.

C. Modified χ^2 Test for Ordinal-Categorical Analysis

Tomlins et al [18] used Pearson's correlation for their multiclass ordinal analyses but no details were given in their paper. Here we use a modified χ^2 test, which should be equivalent to Pearson's correlation. But results show these two approaches actually return different sets of genes. We compare the results of the logit model with those obtained from the modified χ^2 test and those from [18]. Different from the traditional χ^2 test of independence, which treats both variables as nominal, the modified χ^2 test [2] can treat one or both variables as ordinal with a score being assigned to each category. Therefore, it can naturally capture the ordinal trend of cancer progression stages. To construct a contingency table for each probe, we have to *discretize* the expression levels first. A probe is regarded as over-expressed (or under-expressed), if its expression level is greater than the third (first) quartile of all expression values. Then a 2×N contingency table can be built for each probe, where N is the number of progression stages. The statistics based on the contingency table is defined as $M^2 = (n-1)r^2$, where n is the sample size and correlation r is defined as

$$r = \frac{\sum_{i,j} u_i v_j n_{ij} - (\sum_i u_i n_{i+})(\sum_j v_j n_{+j})/n}{\sqrt{\left[\sum_i {u_i}^2 n_{i+} - \frac{(\sum_i u_i n_{i+})^2}{n}\right] \cdot \left[\sum_j {v_j}^2 n_{+j} - \frac{(\sum_j v_j n_{+j})^2}{n}\right]}}.$$

In the above formula for correlation r, u_i denotes the score value for the i^{th} row in the contingency table and v_j denotes the score value for the j^{th} column. Because we only use

two levels for gene expression, u_i can take any two values (e.g., 0 or 1) without affecting the final result. However, choosing an appropriate column-wise score scale is important and it is sometimes the main obstacle of this approach. In the case of prostate cancer, we believe that Gleason score would be a good candidate for this score, which is adopted in our analysis. The number n_{ij} is the number of observations in each cell, and n_{i+}/n_{+j} are the column/row summations. The null hypothesis is the independence of the two variables. Large statistic indicates that expression levels of the probe tend to be increasing or decreasing as cancer progresses. For large sample sizes, the test statistic follows approximately a χ^2 distribution with 1 degree of freedom.

D. Gene Ontology Enrichment Analysis

Statistically analyses of microarray data usually return a large list of candidate genes. To gain insights of the functional characteristics of these genes, a Gene Ontology (GO) enrichment analysis is usually followed. Basically, in order to identify significantly enriched GO terms, for each GO term, one compares the number of genes annotated with the GO term (say k) in the list of n genes with the total number of K genes with the annotation in a reference list of N genes. A statistic based on the hypergeometric distribution can be used. Then significance of enrichment for a given GO term is determined as $P = \sum_{i=k}^{n} \frac{\binom{N-K}{n-i}\binom{K}{i}}{\binom{N}{n}}$.

E. Network Analysis

In addition to gene expression data, other existing data sources may provide additional information to further prioritize the gene list. Towards that end, we first construct gene co-expression networks based on the returned lists. Two additional networks (protein-protein interactions and molecular pathways) are then compiled from existing databases. In this study, we just perform some preliminary studies by examining the overlaps between the co-expression networks and the other two networks. In addition, for directly linked gene pairs in the co-expression networks, we examine the distributions of their shortest distances in other two networks.

III. Results

To verify our hypothesis, we first tried some commonly used approaches for microarray data analysis including the hierarchical clustering algorithm, a two-sample *t*-test for each adjacent pair of stages, and the standard χ^2 test of independence on the prostate cancer progression data from [18]. For the clustering analysis, Pearson's correlation coefficient between each pair of genes was used as the similarity measure. An inherited difficulty in such an analysis is that one cannot automatically determine which groups are associated with the progression (results now shown). The *t*-tests for benign epithelium samples vs. PIN samples, PIN samples vs. PCA samples, and PCA samples vs. metastatic prostate cancer samples, which were also performed in [18], each returned a set of "significant" genes. However, those genes (at Q=0.01) actually do not have any overlaps, indicating that they may not serve as the signature across stages. The standard χ^2 test can identify some genes that seem not independent (Fig.S1) from cancer stages. However their expressions may not necessarily be concordant with cancer progression stages (Fig.S1). Therefore, these commonly used algorithms may not work well for our problem.

A. Overlaps between Different Approaches

We further applied the logit model and the modified χ^2 test on the dataset, and compared the significant gene signatures with those from the original paper [18] using multiclass Pearson's correlation. To declare significance, all three methods used Q-value of 0.05. Among the three approaches, the logit model was more sensitive than the other two. It returned 733/853 probes with increasing/decreasing expression patters across cancer stages. The modified χ^2 test was the most insensitive one and the numbers are 358/438 respectively for increasing/decreasing expression patters, while the numbers of genes from [18] were in-between (490/680). The number of overlaps among these three sets can be found in Fig 1. Though they share a significant portion, they do show some differences, especially for the modified χ^2 test and the Pearson's correlation approach in [18]. Due to lack of details in [18], we do not totally understand what caused the difference between these two.

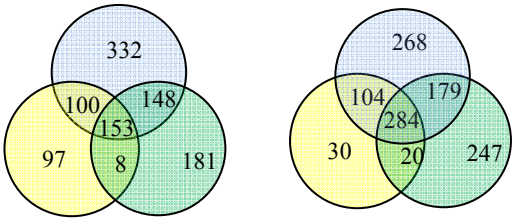

Figure **1**. Overlaps of genes identified by three approaches: logit (top circle), modified χ^2 (left circle), and Tomlins et al. (right circle). Left: up-expressed genes. Right: down-expressed genes.

B. Top Ranked Genes

We examined the top-ranked genes from each list. They were in general consistent with each other. For example, for the top 20 ranked genes in the up/down-expressed list from [18], 15/13 were found by the logit model, and all of them are ranked above 31/33. In addition, further examinations indicated that many known prostate cancer related genes are ranked highly in both lists. These genes include AMACR ($3^{rd}/5^{th}$ in the logit model/Tomlins et al. [18]) and MYC($11^{th}/17^{th}$) in the up-expressed lists and MME ($17^{th}/9^{th}$) in the down-expressed lists [9,15]. Gene CKS2($2^{nd}/ 7^{th}$), has subsequently identified by other studies [14]. The logit model also found some genes related with other types of cancer which were not identified by Tomlins et al. For example, FLJ10849 (5^{th}) is a septin family gene involved in the pathogenesis of 11q23 associated leukemia [13] and EVA1 (2^{nd}) is a papillary thyroid cancer (PTC) related gene [11]. Another interesting fact is that after randomly permuting the labels of samples, the logit model did not select any significant genes. This provides compelling evidences that gene signatures we identified indeed have some biological meaning.

C. Gene Ontology Enrichment Analysis

To systematically characterize the gene signatures, we performed Gene Ontology enrichment analysis as outlined in subsection 2.4 using an online tool named Ontologizer [4]. Gene Ontology (GO) is widely used system with annotations for gene products in many organisms. Ontologizer statistically computes a set of significant GO terms that are found to be enriched in a given gene list comparing to a set of reference genes. In our analysis, we first mapped probes to genes according to the information from GSE6099. Probes that could not be mapped to any genes were removed for this analysis. If multiple probes were mapped to a single gene, only the one with highest significance was retained. The reference gene list consisted of all the genes from our input data. We applied GO enrichment analysis to the up/down regulated genes from the logit model as well as the modified χ^2 test. There were 17 enriched terms from the list of up-regulated genes by the logit model, and about half of the germs were related to "cell cycles" including "cell cycle process", "cell cycle phase" and so on (Fig.S2). This is consistent to the results from [18]. However, no significant terms were found from the list of up-regulated genes from the modified χ^2 test. This may be due to the loss of information during discretization. For the down-regulated signatures from the logit model, 9 out of 13 were actually child nodes of the "cellular component" node (Fig.S2). This was also observed from the modified χ^2 list (7/10) (Fig.S2). Based on these results, it makes more sense to focus on genes with increasing expression levels with cancer progression. But this may also be due to bias/incompleteness of GO. Due to page limitation, significant GO terms and their diagrams were provided as supplementary materials.

D. Network analysis

In order to further analyze the gene signatures by taking advantage of other existing sources, we have compiled two additional networks: one for human protein interaction pairs based on public sources including DIP [16], Reactome [19], BIND [3] and MINT [7], and the other for human gene pairs that co-occur in any disease pathway in the KEGG database [12]. Our goal is to identify those highly co-expressed genes from the lists of gene signatures that also physically interact or co-occur in any disease pathways. Therefore, we first constructed two gene co-expression networks from the up/down expressed gene lists returned by the logit model. To determine a proper threshold, we constructed a null distribution for each gene list by randomly permuting the stage labels among samples and took the 0.005 and 0.995 percentiles as our thresholds. We then examined the overlaps between our co-expression networks and the PPI/ pathway networks. For the up-regulated genes, we identified around 25 connected components (Fig S5), with the largest one shown in Fig 2. We suspect that this set of genes are particularly relevant and need more attentions. In addition to the fact that their expressions are in concordance with cancer progression, further evidences suggest that they might be directly interact. We examined a few nodes in the largest component with high degrees of interactions. The gene in the middle with ID 4609 is MYC (v-myc myelocytomatosis viral oncogene homolog (avian)). The protein encoded by this gene is a multifunctional, nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis and cellular transformation. It functions as a transcription factor that regulates transcription of specific target genes. Mutations, overexpression, rearrangement and translocation of this gene have been associated with a variety of hematopoietic tumors, leukemias and lymphomas [provided by RefSeq]. A recent study shows that Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis [10]. For the largest component with down-expressions, genes 7157, 2033 and 1655 have high connectivities in the PPI network (Fig 2). Gene with ID 7157 is TP53 (tumor protein p53). This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. It is postulated to bind to a p53-binding site and activate expression of downstream genes that inhibit growth and/or invasion, and thus function as a tumor suppressor. Gene with ID 2033 is EP300 (E1A binding protein p300), which is important in the processes of cell proliferation and differentiation. Defects in this gene play a role in epithelial cancer. Gene with ID 1655 is DDX5 DEAD (Asp-Glu-Ala-Asp box polypeptide 5). Based on their distribution patterns, some members of this family are believed to be involved in cellular growth and division [provided by RefSeq]. The overlaps with the pathway network actually consisted of more connected subgraphs. This is probably because we had compiled different disease pathways into one network. Further analysis using individual pathways may provide a clearer picture about their involvements in disease pathways.

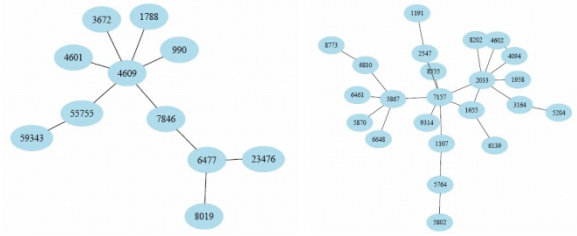

Figure 2 Largest connected components from up-expressed (left) and down-expressed (right) genes in the PPI network. See Fig.S3 for all connected components.

For those nodes that do not direct interact with other nodes in the PPI/pathway networks, we further examined their shortest distances in the PPI/pathway networks and compared the distance distributions with those of a set of randomly selected nodes of equal size from the two networks. Our results (Fig.S4) show that the nodes are significantly condensed, comparing to randomly selected nodes of equal sizes. The *p*-values based on homogeneity χ^2 tests are less than 2.2e-16 for all cases. This result further supports that the gene signatures identified in our study should have biological means associated with them.

IV. Conclusions and Discussions

Understanding of genetic basis of cancer progression is of practical importance for cancer diagnosis and prognosis. There have been tremendous efforts to identify gene signatures or biomarkers using high throughput microarray technology. Traditional statistical methods such as Student's t-test or ANOVA analysis can only identify differentially expressed genes in different progression stages. In this study, we adopted the multi-class logit

model for ordinal analysis to identify gene signatures that show consistent increasing or decreasing expressions in accordant with cancer progressions. We then applied the method to a real data set concerning about prostate cancer progressions. Our results show that the logit model is more sensitive than the modified χ^2 test. Among top ranked genes, we have found several known prostate related genes, as well as other cancer related genes. We further performed GO enrichment analysis and network analysis to gain more information about our selected genes. Results from both analysis show that many genes do have distinct functions that are associated with cell cycles, and they are significantly condensed in other networks. Although biological validation is beyond the scope of the current study, our results provides compelling evidence that gene signatures identified by the approach represent biologically meaningful results. Though we chose prostate cancer data as an example, the method can be applied to any cancer progression data.

ACKNOWLEDGMENT

This work is supported in part by NIH/NLM grant LM008991.

REFERENCES

[1] Abba, M. C., J. A. Drake, et al. (2004). "Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression." Breast Cancer Res 6(5): R499-513.

[2] Agresti, A. (1996). An introduction to categorical data analysis. New York, Wiley.

[3] Alfarano, C., C. E. Andrade, et al. (2005). "The Biomolecular Interaction Network Database and related tools 2005 update." Nucleic Acids Res 33(Database issue): D418-24.

[4] Bauer, S., S. Grossmann, et al. (2008). "Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration." Bioinformatics 24(14): 1650-1.

[5] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc., B 57, 289-300.

[6] Calvo, A., N. Xiao, et al. (2002). "Alterations in gene expression profiles during prostate cancer progression: functional correlations to tumorigenicity and down-regulation of selenoprotein-P in mouse and human tumors." Cancer Res 62(18): 5325-35.

[7] Chatr-aryamontri, A., A. Ceol, et al. (2007). "MINT: the Molecular INTeraction database." Nucleic Acids Res 35(Database issue): D572-4.

[8] Dalgin, G. S., G. Alexe, et al. (2007). "Portraits of breast cancer progression." BMC Bioinformatics 8: 291.

[9] De Marzo, A. M., T. L. DeWeese, et al. (2004). "Pathological and molecular mechanisms of prostate carcinogenesis: implications for diagnosis, detection, prevention, and treatment." J Cell Biochem 91(3): 459-77.

[10] Gurel, B., T. Iwata, et al. (2008). "Nuclear MYC protein overexpression is an early alteration in human prostate carcinogenesis." Mod Pathol 21(9): 1156-67.

[11] Jarzab, B., M. Wiench, et al. (2005). "Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications." Cancer Res 65(4): 1587-97.

[12] Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res 28(1): 27-30.

[13] Kojima, K., I. Sakai, et al. (2004). "FLJ10849, a septin family gene, fuses MLL in a novel leukemia cell line CNLBC1 derived from chronic neutrophilic leukemia in transformation with t(4;11)(q21;q23)." Leukemia 18(5): 998-1005.

[14] Lan, Y., Y. Zhang, et al. (2008). "Aberrant expression of Cks1 and Cks2 contributes to prostate tumorigenesis by promoting proliferation and inhibiting programmed cell death." Int J Cancer 123(3): 543-51.

[15] Rubin, M. A., T. A. Bismar, et al. (2005). "Decreased alpha-methylacyl CoA racemase expression in localized prostate cancer is associated with an increased rate of biochemical recurrence and cancer-specific death." Cancer Epidemiol Biomarkers Prev 14(6): 1424-32.

[16] Salwinski, L., C. S. Miller, et al. (2004). "The Database of Interacting Proteins: 2004 update." Nucleic Acids Res 32(Database issue): D449-51.

[17] Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A 100(16): 9440-5.

[18] Tomlins, S. A., R. Mehra, et al. (2007). "Integrative molecular concept modeling of prostate cancer progression." Nat Genet 39(1): 41-51.

[19] Vastrik, I., P. D'Eustachio, et al. (2007). "Reactome: a knowledge base of biologic pathways and processes." Genome Biol 8(3): R39.

V. APPENDIX

Supplementary materials including the lists of significant genes and large figures can be found online at http://www.eecs.case.edu/~ysp2/bibm.html.

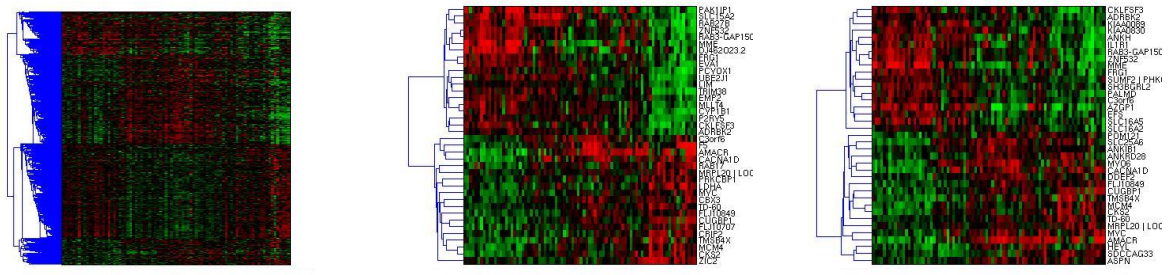

Fig S1. Hierarchical Clustering of genes selected by regular χ^2 test of independence (left), multicategory logit model (middle), and modified χ^2 test of independence (right). Samples are arranged from benign stage to metastatic stage. Genes may not show monotonic increasing/decreasing trends over stages in the regular χ^2 test of independence.

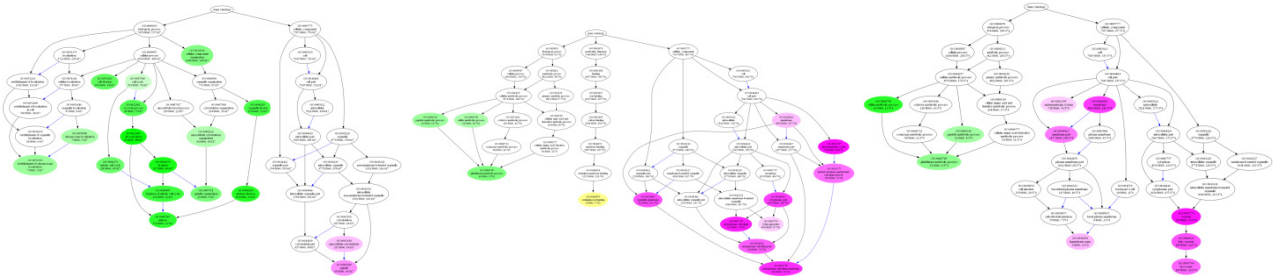

Fig S2. Significant GO terms associated with the gene signatures. Left: GO terms of genes showing increasing expressions identified by the logit model. Middle: GO terms of genes showing decreasing expressions identified by the logit model. Right: GO terms of genes showing decreasing expressions identified by the modified χ^2 test.

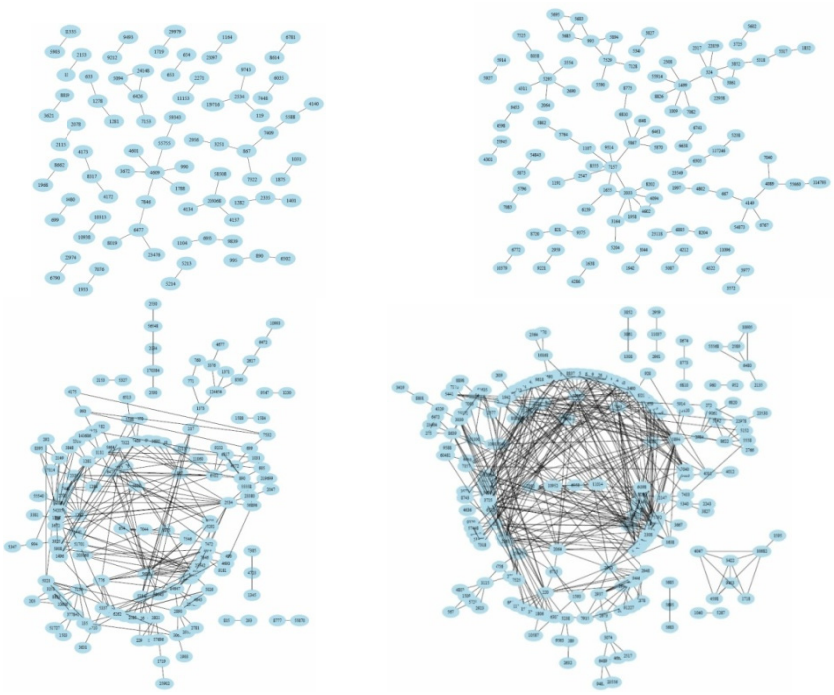

Fig S3. Common connected components (CCC) among different gene networks. Up left: CCC of the up-expressed gene network and the PPI network. Up right: CCC of the down-expressed gene network and the PPI network. Bottom left: CCC of the up-expressed gene network and the disease pathway network. Bottom right: CCC of the down-expressed gene network and the disease pathway network. Each gene is labeled using their NCBI gene id.

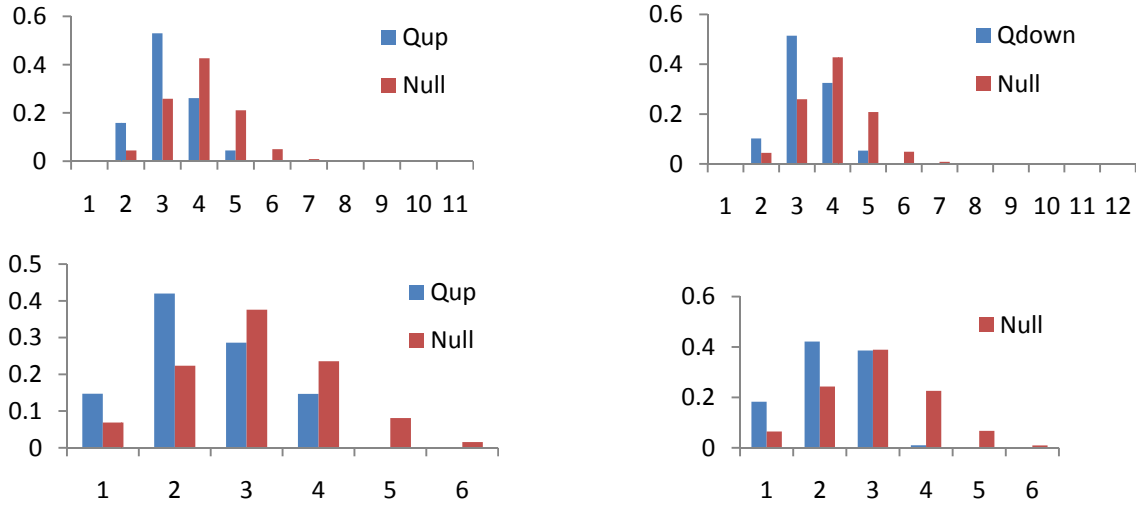

Fig S4. Shortest distance path distribution. The *x*-axis is distance between network nodes and the *y*-axis is the frequency of the distance. Pairs of nodes from the co-expression network are more closely linked in the PPI/pathway networks. Up left: up-expressed genes in the PPI network. Up right: down-expressed genes in the PPI network. Bottom left: up-expressed genes in the pathway network. Bottom Right: down-expressed genes in the pathway network.