Gene sets involved in prostate cancer based on differential expression

Hamid Alavi Majd[1], Soheila Khodakarim[1,*], Mostafa Rezaei Tavirani[2], Farid Zayeri[2], Nasrin Dehghan Nayeri[3],Seyyed Mohammad Tabatabaee[4]

[1]Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences,Tehran, Iran.
[2]Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[3]Department of Proteomics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[4]Department of Medical Informatics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

* Corresponding Author: email address: lkhodakarim@gmail.com (S. Khodakarim)

ABSTRACT

Prostate cancer is the second most common cancer in men. In spite of on-going researches in this filed, the specific causes of prostate cancer are so far unknown. In this study, we used two methods of Gene Set Analysis to improve the biological interpretation of the observed expression patterns in prostate cancer. The Gene Set Analysis is a computational method to discover gene sets whose expression is associated with a phenotype of interest. In addition, we used these methods to search gene sets defined by KEGG and BioCarta. Although, our results showed that most of the gene sets were associated with prostate cancer in the Category and Hotelling's T^2 methods, the power of the Hotelling's T^2 was more than Category method in either KEGG or BioCarta gene sets. The concordance between the results of Pubmed articles and KEGG gene sets was more than the results of Pubmed articles and BioCarta gene sets.

Keywords: Prostate Cancer; Gene Set Analysis; Category method; Hotelling'sT^2 method

INTRODUCTION

Prostate cancer is a form of cancer that develops in the prostate, a gland in the male reproductive system. Although most prostate cancers are slow growing, there are cases of aggressive prostate cancers [1]. Prostate cancer is the second most common cancer in men and the fifth in both sexes combined. However, 14% of all new male cancer cases have been related to this cancer, in the world [2]. In Iran, the incidence rate of prostate cancer was 5.1 per 100,000 person-years [3]. In spite of on-going researches in this filed, the specific causes of prostate cancer are so far unknown [4].

The integration of biology and statistics sciences in Gene Set Analysis (GSA) has been transformed into a strong arm which enables researchers to assay differential gene expression for finding related biomarkers. A gene set is a group of genes that is defined based on prior biological knowledge on gene functions available from public databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG)[5], BioCarta [6] and Gene Ontology (GO)[7]. The discovery of biomarker based on differentially expressed gene set rather than individual gene increases statistical power and enhances interpretability and more direct biological meaning.

Many statistical approaches have been proposed to accomplish GSA methods. Some of them calculate the gene set statistic based on gene level statistic [8-12]. Another group of GSA methods used of multivariate techniques to calculate gene set statistic [13-14]. In this study, two GSA methods, the Category [8] and Hotelling's T^2 [9], were utilized to identify the gene sets defined by the KEGG and BioCarta, which were strongly associated with prostate cancer.

MATERIAL AND METHODS

Category method

The Category method is the rich extension of GSA methods. In this method, *t*-statistic for all genes in the dataset is calculated as gene level statistic. The mean of the absolute *t*-statistics belonging to the same set is calculated as set level statistic:

$$\sum_t = \frac{1}{m}\sum_{i=1}^{m}|t_i| \quad (1)$$

Where t_i is the *t*-statistic for the *i*th gene and *m* is the number of genes in a gene set. This idea that the changes of gene expressions in each

gene set is either up or down regulated seems not to be true, thus we preferred to use the absolute *t*-statistics instead of *t*-statistics. The subject sampling has been used to determine permutation *p*-values. The subject sampling takes the subject (sample) as the sampling unit [8, 15]. The Category package in Bioconductor implements this method. The correlation structures within each set were not considered in this method because the set level statistic is computed based on the gene level statistic.

Hotelling's T^2

The Hotelling's T^2 statistic tested the hypothesis $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$, if F_1 and F_2 are multivariate normal distributions with common covariance matrix. Let m denote the number of genes in a gene set, n_i denote sample size for ith phenotype $(i=1,2)$.This statistic is:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(X_1 - X_2)V^{-1}(X_1 - X_2)' \quad (2)$$

where V is the covariance matrix of the gene expression and $\bar{X}_i$ is the m-vector of means for the ith phenotype.

One of the important problems in a gene expression study is that the number of sample is always much less than the number of gene. For this reason, to calculate the inverse of covariance matrix, one needs to use additional steps. Tsai and Chen used the shrinkage estimator to calculate the inverse of covariance matrix [13]. The shrinkage covariance matrix estimator (V^*_{ij}) proposed by Schafer and Strimmeris [16] can be written:

$$V^*_{ij} = \begin{cases} v_{ii} & \text{if} \quad i = j \\ r^*_{ij}\sqrt{v_{ii}v_{jj}} & \text{if} \quad i \neq j \end{cases} \quad (3)$$

and

$$r^*_{ij} = r_{ij} \times \min\{1, \min(0, 1-\hat{\lambda})\} \quad (4)$$

Where v_{ii} and r_{ij}, respectively, denote the empirical sample variance and sample correlation, and the optimal shrinkage intensity $\hat{\lambda}^*$ is estimated by:

$$\hat{\lambda}^* = \frac{\sum_{i \neq j} \hat{\mathrm{var}}(r_{ij})}{\sum_{i \neq j} r^2_{ij}} \quad (5).$$

Moreover, this method took account either the correlation structure among genes or both up-regulated and down-regulated gene expressions. The permutation *p*-values were calculated based on subject sampling in this package.

Datasets

The prostate cancer dataset was downloaded from Gene Expression Omnibus (GEO). This dataset which hybridized to affymetrix human genome HG-U133A platform consists of a total RNA from 148 prostate samples with various percentages of tumors determined by pathologist [15]. 12 samples whose percentage of tumor was not registered were excluded and the other samples were all included in the processing. We divided 136 samples based on percentage of tumor into two groups, tumor *(n=65)* and non-tumor *(n=71)*. The percentage of tumor range 0 to 0.1 in tumor group and 0.1 to 0.8 in another group. The null hypothesis tested here is various percentages of tumors with respect to their overall gene expression pattern. There are 22,283 probe sets in this platform.

The normalization of microarray data by the robust multi-array average (RMA) [18] algorithm was implemented using the Babelomics suite (an integrated web tool for microarray data analysis and functional profiling of genome-scale experiments) [19]. The Babelomics was also used to categorize 22,283 probe sets to 106 KEGG gene sets and 312 BioCarta gene sets. The KEGG and BioCarta gene sets contained 2,665 and 2,046 probe sets, respectively. Hence a high number of probes on the prostate cancer lacked gene set. We revealed differentially expressed KEGG and BioCarta gene sets between tumor and non-tumor prostate cancer samples by the Category and Hotelling's T^2 methods.

RESULTS

In this research, we found that 1,303 probe sets out of 2,665 related KEGG were statistically different, while 928 probe sets out of 2,046 probe related BioCarta were statistically different between tumor and non-tumor samples.

The Hotelling's T^2 method revealed that 105 KEGG gene sets were significant (*p*-values less than 0.05), while 65 KEGG significant gene sets were observed in the Category method. The top ten KEGG significant gene sets for two methods were shown in table 1.

Table **1.** The KEGG gene sets with p < 0.05 by the two GSA methods in the prostate cancer dataset.

KEGG Gene Set		Size	Statistics	Permutation P-value
Category				
1	DNA replication	10	-6.99515	0.000
2	Mismatch repair	12	-8.83720	0.000
3	Primary immunodeficiency	47	-8.26923	0.000
4	One carbon pool by folate	57	-12.9652	0.000
5	Tryptophan metabolism	13	-7.46058	0.000
6	Renin-angiotensin system	35	-7.60349	0.000
7	Base excision repair	20	-10.11030	0.000
8	Limonene and pinene degradation	19	-5.99054	0.000
9	Primary bile acid biosynthesis	39	-9.54988	0.002
10	Riboflavin metabolism	4	-3.71500	0.002
Hotelling's T2				
1	Nitrogen metabolism	75	511.077790	0.000
2	Arginine and proline metabolism	69	506.142051	0.000
3	Drug metabolism - other enzymes	86	480.856398	0.000
4	Glyoxylate and dicarboxylate metabolism	68	477.615468	0.000
5	Starch and sucrose metabolism	78	470.832097	0.000
6	Glycosphingolipid biosynthesis	70	447.099907	0.000
7	mTOR signaling pathway	78	444.186458	0.000
8	Pentose phosphate pathway	80	427.867381	0.000
9	Caffeine metabolism	84	423.784298	0.000
10	Linoleic acid metabolism	71	419.313996	0.000

In BioCarta, the Hotelling's T^2 method showed that 312 gene sets were significant (*p*-values less than 0.05), while 54 significant gene sets were observed in the Category method. The top ten BioCarta significant gene sets by the two methods were displayed in table 2.

Table **2.** BioCarta gene sets with p < 0.05 by the two GSA methods in the prostate cancer dataset.

BioCarta Gene Set		Size	Statistic	Permutation p-value
Category				
1	Integrin Signaling Pathway	11	13.82763	0.000
2	Role of PI3K subunit p85 in regulation of Actin Organization and Cell Migration	37	16.50176	0.000
3	Role of Erk5 in Neuronal Survival	29	11.09129	0.000
4	IL12 and Stat4 Dependent Signaling Pathway in Th1 Development	24	13.49084	0.000
5	VEGF, Hypoxia, and Angiogenesis	26	12.03878	0.000
6	The information-processing pathway at the IFN-beta enhancer	27	12.19200	0.000
7	Inhibition of Huntington's disease neurodegeneration by histone deacetylase inhibitors	24	9.94616	0.001
8	Oxidative Stress Induced Gene Expression Via Nrf2	26	13.56122	0.001
9	CXCR4 Signaling Pathway	16	10.24244	0.001
10	Opposing roles of AIF in Apoptosis and Cell Survival	12	6.139434	0.004
Hotelling's T2				
1	ALK in cardiac myocytes	46	336.0954	0.000
2	Low-density lipoprotein (LDL) pathway during atherogenesis	39	326.3753	0.000
3	Agrin in Postsynaptic Differentiation	44	325.3372	0.000
4	Nuclear Receptors in Lipid Metabolism and Toxicity	55	312.7743	0.000
5	Electron Transport Reaction in Mitochondria	37	308.0685	0.000
6	Synaptic Proteins at the Synaptic Junction	53	299.9156	0.000
7	Antigen Processing and Presentation	32	299.1207	0.000
8	Ceramide Signaling Pathway	26	296.0046	0.000
9	Role of BRCA1, BRCA2 and ATR in Cancer Susceptibility	37	292.5901	0.000
10	Role of PI3K subunit p85 in regulation of Actin Organization and Cell Migration	37	291.208	0.000

We excluded genes present in more than 30 gene sets in BioCarta. There were no relationships between them and prostate cancer in Pubmed literatures. The name of these genes has been listed in table 3.

Table3. The shared genes between pathways in BioCarta

Row	Name	Gene ID	# involved gene sets
1	MAP2K1	Hs.132311	54
2	MAPK3	Hs.861	59
3	HRAS	Hs.37003	55
4	PRKCA	Hs.349611	45
5	RAF1	Hs.257266	43
6	JUN	Hs.78465	40
7	CAM1	Hs.512804	40
8	NFKB1	Hs.160557	35
9	MAPK1	Hs.324473	36
10	GRM1	Hs.32945	34
11	RB1	Hs.408528	36
12	GRB2	Hs.411366	37
13	SHC1	Hs.433795	34
14	AKT1	Hs.368861	30
15	SOS1	Hs.326392	30

DISCUSSION

Prostate Cancer is one of the most challenging cancers in the medical field and its mechanism still remains completely unclear. There are more than 20 types of prostate cancer. No single theory can provide a perfect definition for different cases of this cancer. The GSA of microarray data not only shows a consistent alteration in a cancer state, but also is a valid method to reduce a major deviation. In addition, the GSA methods enable us to obtain common gene sets by integration differently expressed genes.

Hence, we used the GSA methods for exploration of genes in prostate cancer which are difficult to detect by individual gene analysis because of their subtle change. In the present study, the Hotelling's T^2 and Category were applied to prostate cancer dataset to extract biological insights involved in this cancer by defined gene sets in KEGG and BioCarta.

Our finding showed that the power of multivariate analysis (Hotelling's T^2) is more than univariate analysis (Category) in KEGG and BioCarta. These results were in agreement with published findings elsewhere [13, 27-28]. This conclusion is anticipated because the Hotelling's T^2 takes account the complicated correlation structure and interaction among genes, unlike the Category that is based on univariate analysis (*t*-statistics).

Our results showed that most of the gene sets were associated with prostate cancer. According to GSA results, we discussed several differentially expressed gene sets and genes shared among gene sets which suggested the role of these gene sets and genes in prostate cancer. The concordance between the results of Pubmed articles and KEGG gene sets was more than BioCarta gene sets.

Furthermore, prostate cancer pathway (ID 5215) and Transcriptional misregulation in cancers (ID 5202) are connected pathways with prostate cancer in KEGG. The Thyroid cancer which related pathways with Transcriptional misregulation in cancer was shown in table 1.

ACKNOWLEDGEMENTS

We would like to thank the Centro de Investigacion Principe Felipe, Valencia Spain and Clinical Proteomics Research Centre Tehran Iran for their supports.

REFERENCES

1. American Cancer Society. http://www.cancer.org/Cancer/ProstateCancer/OverviewGuide/prostate-cancer-overview-what-is-prostate-cancer. 2011.

2. Ferlay J, Shin H, Bray F, Forman D, Mathers C, Parkin D. GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer, 2010. http://globocan.iarc.fr.

3.Sadjadi A, Nooraie M, Ghorbani A, Alimohammadian M, Zahedi MJ, Darvish-Moghadam S, et al. The Incidence of Prostate Cancer in Iran: Results of a Population-Based Cancer Registry. Arch Iranian Med. 2007;10(4):481-5.
4.Hsing A, Chokkalingam A. Prostate cancer epidemiology. Frontiers in Bioscience. 2006;11:1388–413.
5.Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.
6.BioCarta. www.biocarta.com.
7.Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25:25-9.
8.Gentleman R. Using Categories to Model Genomic Data 2010. Available from: http://www.bioconductor.org/ packages/2.3/ bioc/vignettes/Category/inst/doc/ Category.pdf.
9.Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA. 2005;102:13544 - 9.
10.Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003;34:267-73.
11.Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25; 102(43):15545-50.
12.Irizarry RA, Wangy C, Zhouz Y, Speed TP. Gene Set Enrichment Analysis Made Simple. Stat meth Med Res. 2009; 18(6):565-75.
13.Tsai CA, Chen JJ. Multivariate analysis of variance test for gene set analysis. Bioinformatics. 2009;25(7):897-903.
14.Nettleton D, Recknor J, Reecy JM. Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. Bioinformatics. 2008 Jan 15; 24(2):192-201.
15.Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5:R80.
16.Chiou SS, Huang JL, Tsai YS, Chen TF, Lee KW, Juo SH, et al. Elevated mRNA transcripts of non-homologous end-joining genes in pediatric acute lymphoblastic leukemia. Leukemia. 2007;21(9):2061-4.
17.Wang Y, Xia XQ, Jia Z, Sawyers A, Yao H, Wang-Rodriquez J, et al. In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res. 2010;70(16):6448-55.
18.Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KtJ, Scherf U, et al. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostat. 2003;4(2):249-64.
19.Al-Shahrour F, Minguez P, Tarraga J, Montaner D, Alloza E, Vaquerizas JM, et al. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res. 2006;34(Web Server issue):W472-6.
20.van den Hoogen C, van der Horst G, Cheung H, Buijs JT, Pelger RC, van der Pluijm G. The aldehyde dehydrogenase enzyme 7A1 is functionally involved in prostate cancer bone metastasis. Clin Exp Metastasis. 2011;28(7):615-25.
21.Lin J, Xu J, Tian H, Gao X, Chen Q, Gu Q, et al. Identification of candidate prostate cancer biomarkers in prostate needle biopsy specimens using proteomic analysis. Int J Cancer. 2007;121(12):2596-605.
22.Persano L, Moserle L, Esposito G, Bronte V, Barbier iV, Iafrate M, et al. Interferon-alpha counteracts the angiogenic switch and reduces tumor cell proliferation in a spontaneous model of prostatic cancer. Carcinogenesis. 2009;30(5):851-60.
23.Leon C, Locke J, Adomat H, Etinger S, Twiddy A, Neumann R, et al. Alterations in cholesterol regulation contribute to the production of intratumoral androgens during rogression to castration-resistant prostate cancer in a mouse xenograft model.
. Prostate. 2010;70(4):390-400.
24.Locke J, Wasan K, Nelson C, Guns E, Leon C. Androgen-mediated cholesterol metabolism in LNCaP and PC-3 cell lines is regulated through two different isoforms of acyl-coenzyme A:Cholesterol Acyltransferase (ACAT). Prostate. 2008;68(1):20-33.
25.Benavides F, Blando J, Perez C, Garg R, Conti C, DiGiovanni J, et al. Transgenic overexpression of PKCε in the mouse prostate

induces preneoplastic lesions. Cell Cycle. 2011;10(2):268-77.
26.Martinez J, Sali T, Okazaki R, C A, Hollingshead M, Hose C, et al. Drug-induced expression of nonsteroidal anti-inflammatory drug-activated gene/macrophage inhibitory cytokine-1/prostate-derived factor, a putative tumor suppressor, inhibits tumor growth.
. J Pharmacol Exp Ther. 2006;18(2):899-906.
27.Kong SW, Pu WT, Park PJ. A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics. 2006;22(19):2373-80.
28.Khodakarim S, Alavimajd H, Rezaei-tavirani Mostafa, Zayeri Farid, Dehghan Nasrin. A Comparison of Univariate and Multivariate Gene Set Analysis in Acute Lymphoblastic Leukemia. APJCP. 2012 ;3(2) (in press).