RESEARCH ARTICLE

Comparison of Univariate and Multivariate Gene Set Analysis in Acute Lymphoblastic Leukemia

Khodakarim Soheila[1], AlaviMajd Hamid[2]*, Zayeri Farid[2], Rezaei-Tavirani Mostafa[3], Dehghan-Nayeri Nasrin[4], Tabatabaee Syyed-Mohammad[5], Tajalli Vahide[6]

Abstract

Background: Gene set analysis (GSA) incorporates biological with statistical knowledge to identify gene sets which are differentially expressed that between two or more phenotypes. Materials and Methods: In this paper gene sets differentially expressed between acute lymphoblastic leukaemia (ALL) with BCR-ABL and those with no observed cytogenetic abnormalities were determined by GSA methods. The BCR-ABL is an abnormal gene found in some people with ALL. Results: The results of two GSAs showed that the Category test identified 30 gene sets differentially expressed between two phenotypes, while the Hotelling's T^2 could discover just 19 gene sets. On the other hand, assessment of common genes among significant gene sets showed that there were high agreement between the results of GSA and the findings of biologists. In addition, the performance of these methods was compared by simulated and ALL data. Conclusions: The results on simulated data indicated decrease in the type I error rate and increase the power in multivariate (Hotelling's T^2) test as increasing the correlation between gene pairs in contrast to the univariate (Category) test.

Keywords: Acute lymphoblastic leukemia - microarray - gene set analysis - category - Hotelling's T^2

Asian Pacific J Cancer Prev, **14 (3)**, 1629-1633

Introduction

Microarray technology is allowing researchers to measure the expression of thousands of genes simultaneously which this has translated a tool for identifying genes that have been expressed differentially among different phenotypes. Always a list of differentially expressed genes is the result of a microarray experiment. The main attention of the researchers is to translate such lists into a better understanding of the underlying biological phenomena related to interest phenotypes. This is the starting point for Gene Set Analyses (GSA) to incorporate biological into statistical knowledge (Emmert-Streib and Glazko, 2011).

From 2000, a number of enormous approaches with different statistical methods have been suggested to execute GSA, we divide them into two groups: 1- Based on univariate analyses, 2- Based on multivariate analyses. In univariate analyses, some researchers used tests based on contingency tables such as chi-square, Fisher-exact-test could not find the small differences between phenotypes. This subgroup was called Overrepresentation (Man et al., 2000, Al-Shahrour et al., 2004; Khatri and Draghici, 2005). Another subgroup of these methods is the Gene Set Enrichment Analysis (GSEA). Although it utilized the result from individual gene analyses, using of statistics such as Kolmogorov-Smirnov, Mean and Sum could increase the ability of these methods to identify differentially expressed gene sets (Mootha et al., 2003; Subramanian et al., 2005; Tian et al., 2005).

The GSEA presented by Mootha et al. (2003) could identify oxidative phosphorylation as a gene set with differential expression between normal and diabetes type II phenotypes while previous method could not do it. Although the GSEA methods obtain credible results, these approaches cannot take account of correlation structure between genes and cover the hypothesis of interest involving a group of genes.

In contrast to the GSEA approaches, multivariate analyses (Hotelling's T^2 and N-statistics) consider correlation structures between genes within each gene set (Kong et al., 2006; Nettleton et al., 2008; Tsai and Chen, 2009). A common drawback of microarray data is high dimension of microarray data -due to many genes with small samples that increases type I error rate- adjusted by statistical techniques such as principal component or shrinkage analysis (Liu et al., 2007; Tsai and Chen, 2009; Jacobson and Emerton, 2012). Goeman et al. (2004) and Hummel et al. (2008) proposed the Globaltest and ANCOVAGlobal test, respectively. These methods set in

*[1]Department of Epidemiology, Faculty of Public Health, [2]Department of Biostatistics, [4]Department of Proteomics, [5]Department of Medical Informatics, Faculty of Paramedical Sciences, [3]Proteomics Research Center, Shahid Beheshti University of Medical Sciences,[6]Department of Linguistics, Faculty of Literature and Human Sciences, Tehran University, Tehran, Iran *For correspondence: alavimajd@gmail.com*

the multivariate approaches, because they modeled gene expressions as random effects in a logistic regression model and calculated p-value use of the score test proposed by Le Cessie and Van Houwelingen (1995) and Houwing-Duistermaat et al. (1995).

In this study we evaluated two groups by simulated and acute lymphoblastic leukemia (ALL) microarray dataset with use of the Category and Hotelling's T^2 approaches.

Materials and Methods

Here we describe two gene set analysis methods, the Category based on univariate techniques and the Hotelling's T^2 based on multivariate techniques. Permutations were used for the statistical significance calculation in both methods.

Category

The Category analysis is the simple and wealthy extension of the GSEA. This method presents genes and gene sets that they are expressed differentially (Gentleman, 2010). This package found out p-values based on summing the t-statistics for the all members of each gene set and did permutation for calculating permutation-based p-value. We used this method by Category package in the Bioconductor (www.bioconductor.org).

Hotelling's T^2

Tsai et.al considered complicated correlation structures between genes and used of the Hotelling's T^2 statistic. However, one of the important statistical issues associated with differential expression detection for large scale microarray data lies in the extreme multiple testing and many false positives are likely to be identified just by chance. So, they modified the Hotelling's T^2 statistic by incorporating a shrinkage sample covariance matrix in the test statistics. This method is useful for identifying differentially expressed gene sets that contain both up- and down-regulated genes (Tsai and Chen, 2009).

Results

Simulation experiment

We carried out a set of simulations, to assess the performance of the two GSA methods (Category and Hotelling's T^2). The simulated data sets contained four gene sets, respectively with 3, 5, 10, and 20 genes. Expression of these 38 genes for the two groups was generated from a multivariate normal distribution with a mean vector μ and a diagonal variance-covariance matrix Σ. In this process, 38 elements of μ were generated as uniform and random variables in interval (0, 10) and the 38 diagonal elements of Σ were generated as uniform and random variables in interval (0.1,10). The first expression was uncorrelated among the genes within each set. In the next step we repeated the same process except for the off-diagonal elements of the variance-covariance matrix Σ. In this stage, the off-diagonal zero correlations between all pairs of the genes in each set were substituted with a correlation of 0.3, 0.5 and 0.9. For j = 1,..., 38, mean vectors $\mu i's$ for the two phenotypes (i = 1, 2) differ by. Here, we consider a range of γ from 0-3 with an increment of 0.3 (Liu et al., 2007; Tsai and Chen, 2009).

The simulation data were replicated μ_{1j}-μ_{2j}=γ 1000 times in each condition and for calculating permutation-based p-value has been done permutations 1000 times. Then, we checked the type I error and power of the two tests according to the simulation data. For comparing the type I error across the two tests, it was estimated by the observed proportion of replications with a p-value smaller than the size 0.05. For each permutation-based p-value, 1000 random permutations were carried out (γ=0). Also, to compare the power across the two tests, the observed proportion of the replications of an experiment in which the null hypothesis was correctly rejected estimated the power.

Table 1 shows type I error rate of two tests for correlations 0, 0.3, 0.5 and 0.9; gene set size of 3, 5, 10 and 20; and sample size of 10 and 25 in each group. As the correlations between gene pairs increase, we could not see any systematic pattern in changes of the type I error rate of the Category method, while the type I error rates of Hotelling's T^2 decreases. However, when the size of gene sets in each group increases, the range of the variation of the type I error decreases and in the Category method that it goes to zero.

The results of observed power using 10 samples in each scenario of each method are shown in Figure 1. As the correlation increases, the power of the Hotelling's T^2 method increases, however the power of the Category method decreases slightly. Both methods show decrease in power when the size of gene sets increase. We also estimated the power of the two tests using 25 samples, instead of 10 samples in each group, and observed similar patterns as shown in Figure 2.

Application

In this section, we used the described GSA methods

Table 1. The Type I Error of the Simulation Experiment

Gene		n=10				n=25			
Set Size	Test	r=0	r=0.3	r=0.5	r=0.9	r=0	r=0.3	r=0.5	r=0.9
3	Category	0.136	0.116	0.139	0.124	0.121	0.124	0.123	0.113
	Hotteling' T^2	0.038	0.020	0.022	0.007	0.037	0.032	0.038	0.016
5	Category	0.055	0.046	0.017	0.053	0.048	0.052	0.049	0.045
	Hotteling' T^2	0.002	0	0	0.001	0.003	0.005	0.004	0.002
10	Category	0.009	0.019	0.016	0.017	0.017	0.013	0.013	0.016
	Hotteling' T^2	0	0	0	0	0	0	0	0
20	Category	0	0	0	0.001	0.001	0.003	0.002	0.001
	Hotteling' T^2	0	0	0	0	0	0	0	0

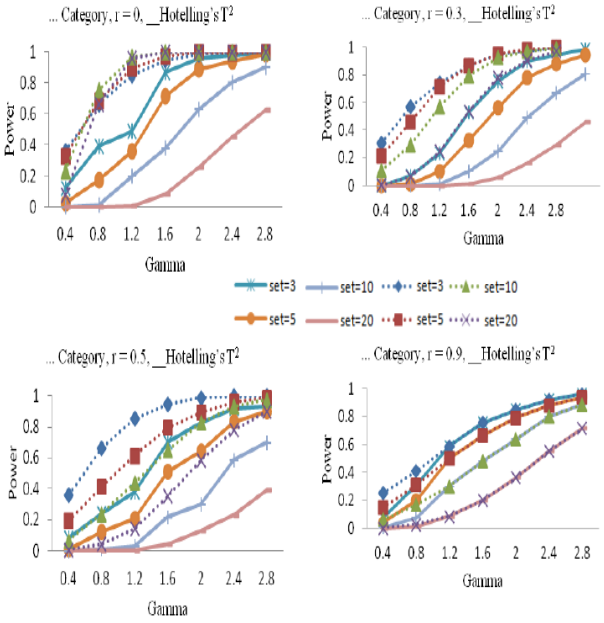

Figure 1. The Results of the Simulation Experiment, Power Analysis of Three Gene Set Analysis. 10vs.10 samples

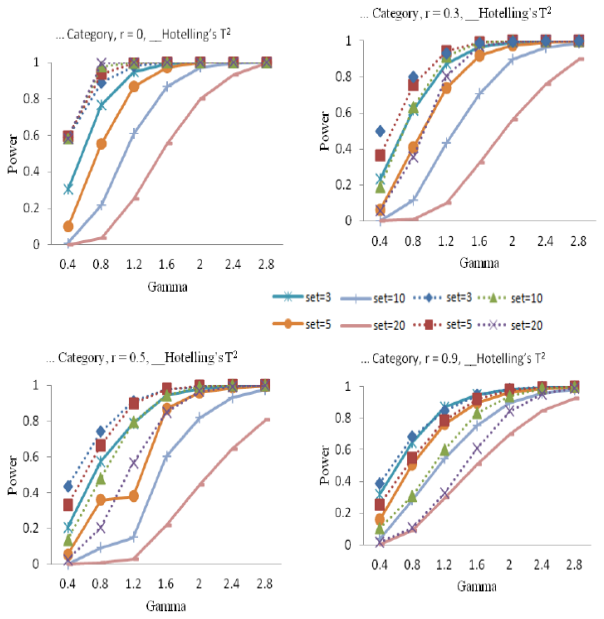

Figure 2. The Results of the Simulation Experiment, Power Analysis of Three Gene Set Analysis. 25vs.25 samples

(Category and Hotelling's T^2) to analyze the data from a microarray study of acute lymphoblastic leukemia. This data set is publicly available at the Bioconductor (Li, 2007). The ALL dataset contains 12625 genes and 128 samples. The filtered dataset contains 79 samples (37 ALL patients with BCR-ABL and 42 persons with no observed cytogenetic abnormalities (NEG) and 1857 genes.

These genes were categorized according to the Kyoto Encyclopedia Gene and Genome (KEGG) as 200 gene sets (Kanehisa and Goto, 2000).

The Category initiated 30 gene sets with p-values less than 0.05, while only 19 significant gene sets were observed in the Hotelling's T^2 methods. From these findings, it seems that the Category has much greater statistical power comparing to another method. In the Category method, our findings about the significant effect of DNA replication (p=1.24e-10), repair mismatch (p=9.72e-7), non-homologous end-joining and purine metabolism (p=0.000247) on ALL were in agreement with the results of the previously published surveys (van Laarhoven et al., 1983; Matheson and Hall, 1999; Shah and Rajshekhar, 2004; Chiou et al., 2007). Table 2 shows the significant gene sets in the Category method.

Table 3 shows the significant gene sets in the Hotelling's T^2 method. Our finding about the significant effect of D-glutamine and D-glutamate metabolism (p=0.03), Glycosphingolipid biosynthesis-globo series (p=0.034) and Renin-angiotensin system (p=0.032) on ALL were in agreement with the results of the previously published surveys (Merritt et al., 1988; Teresa Gomez Casares et al., 2002; Cory and Cory, 2006).

In this case, we found a number of shared genes among the significant gene sets. For example PCNA, RFC1, RFC2, RFC3, RFC4 and RFC5 were previously shown to be related to ALL elsewhere (Zolzer et al., 2010; Kobayashi et al., 1989; de Jonge et al., 2009; Koppen et al., 2010). These genes are a subset of common genes between DNA replication and repair mismatch. Moreover, there were eight common genes (POLA2, POLE3, POLA1, POLE, POLE2, PRIM1, PRIM2 and POLE4) between DNA replication and purine metabolism. According to our research, however, we could not find any relationship between ALL and these genes in other documents.

Table 2. Gene Sets in the ALL Dataset with P-value<0.05 by the Category Method

	Gene Set	P-value
1	Ribosome	3.59E-21
2	DNA replication	1.24E-10
3	Spliceosome	7.42E-07
4	Mismatch repair	9.72E-07
5	Homologous recombination	9.65E-05
6	Non-homologous end-joining	0.000247
7	Nucleotide excision repair	0.000308
8	Purine metabolism	0.000464
9	RNA polymerase	0.000859
10	Parkinson's disease	0.001277
11	Pyrimidine metabolism	0.002218
12	Terpenoid backbone biosynthesis	0.003047
13	O-Glycan biosynthesis	0.003543
14	Base excision repair	0.004281
15	Metabolic pathways	0.008337
16	Tyrosine metabolism	0.011487
17	Glycolysis / Gluconeogenesis	0.012894
18	Oxidative phosphorylation	0.014379
19	Citrate cycle (TCA cycle)	0.016137
20	Keratan sulfate biosynthesis	0.016518
21	Pyruvate metabolism	0.017790
22	Cysteine and methionine metabolism	0.023119
23	Porphyrin and chlorophyll metabolism	0.025649
24	Pentose phosphate pathway	0.028052
25	Proteasome	0.028645
26	Cardiac muscle contraction	0.029528
27	Heparan sulfate biosynthesis	0.031428
28	Glioma	0.039697
29	RNA degradation	0.041485
30	Basal transcription factors	0.045362

Table 3. Gene sets in the ALL dataset with P-value<0.05 by the Hotelling's T^2 method

	Gene Set	P-value
1	Primary bile acid biosynthesis	0.016
2	Sulfur metabolism	0.016
3	Phenylalanine, tyrosine and tryptophan biosynthesis	0.018
4	Lysine biosynthesis	0.023
5	Thiamine metabolism	0.028
6	D-Glutamine and D-glutamate metabolism	0.030
7	Renin-angiotensin system	0.032
8	Glycosphingolipid biosynthesis - globo series	0.034
9	Synthesis and degradation of ketone bodies	0.036
10	O-Glycan biosynthesis	0.039
11	Folate biosynthesis	0.039
12	Heparan sulfate biosynthesis	0.040
13	Linoleic acid metabolism	0.040
14	Maturity onset diabetes of the young	0.040
15	Methane metabolism	0.041
16	Pantothenate and CoA biosynthesis	0.042
17	Chondroitin sulfate biosynthesis	0.045
18	Histidine metabolism	0.050
19	Riboflavin metabolism	0.050

The relationship among gene sets that share some of their members and their proper interpretation are subject to further investigation.

Discussion

The methods based on multivariate techniques could regard as the all of the genes within each gene at the same time and account for the correlation structures between gene pairs. We expected the findings of these methods would be better than the methods based on univariate techniques. Because, the univariate methods such as the Fisher's exact test or GSEA test which calculate the p-values under the assumption of independence between genes will have incorrect type I error if genes are in fact correlated (Goeman and Buhlmann, 2007; Liu et al., 2007). In this paper, we evaluated the execution of the Category and Hotelling's T^2 methods (univariate and multivariate techniques) for analyzing of gene sets on simulated and real gene expression data. Both chosen methods are self-contained null hypotheses, because self-contained hypothesis tests have more power and more clear biological interpretation than competitive null hypothesis tests (Goeman and Buhlmann, 2007).

The results on simulated data were according to our expectation, they indicated decrease the type I error rate and increase the power in multivariate (Hotelling's T^2) test as increasing the correlation between gene pairs. In the scenarios with the correlation less than 0.5, the power of Hotelling's T^2 test was less than the Category test (one-sided tests).

In spite of the general belief that multivariate tests pay attention to a complex correlation structure between genes and, hence, may result in a better power compared to univariate tests, Emmert-Streib and Glazko (2011), Nettleton et al. (2008) and our results (when correlations is less than 0.5) did not confirm this belief.

Another reason to this non-ordinary result may be up-regulated expression in our simulated data (Kong et al., 2006). However, we know the Category method is a one-sided test, while the Hotelling's T^2 method is a two-sided test. 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 two-sided tests (the Hotelling's T^2 method) instead of one-sided tests (the Category method).

Perhaps a wrong assumption which data was simulated from a Multivariate Normal distribution has made dissimilarity between the results of the real and simulated data. Purdom and Holmes (2005) pointed to this common error in simulation while in many GSA studies have been used Multivariate Normal distribution to simulate gene expression data (Kong et al., 2006; Jiang and Gentleman, 2007; Liu et al., 2007; Dinu et al., 2008; Song and Black, 2008; Tsai and Chen, 2009).

References

Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578-80.

BioConductor: Open Source Software for Bioinformatics [cited; Available from: www.bioconductor.org].

Chiou SS, Huang JL, Tsai YS, et al (2007). Elevated mRNA transcripts of non-homologous end-joining genes in pediatric acute lymphoblastic leukemia. *Leukemia,* **21**, 2061-4.

Cory JG, Cory AH (2006). Critical roles of glutamine as nitrogen donors in purine and pyrimidine nucleotide synthesis: asparaginase treatment in childhood acute lymphoblastic leukemia. *In Vivo*, **20**, 587-9.

de Jonge R, Tissing WJ, Hooijberg JH, et al (2009). Polymorphisms in folate-related genes and risk of pediatric acute lymphoblastic leukemia. *Blood*, **113**, 2284-9.

Dinu I, Liu Q, Potter JD, et al (2008). A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. *Cancer Inform*, **6**, 357-68.

Emmert-Streib F, Glazko GV (2011). Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Computational Biology*, **7**, 1002053.

Gentleman R (2010). Using Categories to Model Genomic Data [cited; Available from: www.bioconductor.org].

Goeman JJ, Buhlmann P (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-7.

Goeman JJ, van de Geer SA, de Kort F, et al (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93-9.

Houwing-Duistermaat JJ, Derkx BHF, Rosendaal FR, et al (1995). Testing familial aggregation. *Biometrics*, **51**, 1292-301.

Hummel M, Meister R, Mansmann U (2008). GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **2**4, 78-85.

Jiang Z, Gentleman R (2007). Extensions to gene set enrichment. *Bioinformatics*, **23**, 306-13.

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

Khatri P, Draghici S (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587-95.

Kobayashi H, Takemura Y, Ohnuma T (1998). Variable expression of RFC1 in human leukemia cell lines resistant

to antifolates. *Cancer Lett*, **124**, 135-42.
Kong SW, Pu WT, Park PJ (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373-80.
Koppen IJ, Hermans FJ, Kaspers GJ (2010). Folate related gene polymorphisms and susceptibility to develop childhood acute lymphoblastic leukaemia. *Br J Haematol*, **148**, 3-14.
le Cessie S, van Houwelingen H (1995). Testing the fit of a regression model via score tests in random effects models. *Biometrics*, **51**, 600-14.
Li X (2009). ALL: A data package [cited; Available from: www.bioconductor.org].
Liu Q, Irina Dinu I, Adewaleet AJ, et al (2007). Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.
Man MZ, Wang X, Wang Y (2000). POWER SAGE: comparing statistical test for SAGE experiments. *Bioinformatics*, **16**, 953-9.
Matheson EC, Hall AG (1999). Expression of DNA mismatch repair proteins in acute lymphoblastic leukaemia and normal bone marrow. *Adv Exp Med Biol*, **457**, 579-83.
Merritt WD, Sztein MB, Reaman GH (1988). Detection of GD3 ganglioside in childhood acute lymphoblastic leukemia with monoclonal antibody to GD3: restriction to immunophenotypically defined T-cell disease. *J Cell Biochem*, **37**, 11-9.
Mootha VK, Lindgren CM, Eriksson KF, et al (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **3**4, 267-73.
Nettleton,D, Recknor J, Reecy JM (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192-201.
Shah KC, Rajshekhar V (2004). Glioblastoma multiforme in a child with acute lymphoblastic leukemia: case report and review of literature. *Neurol India*, **52**, 375-7.
Song S, Black MA (2008). Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, **9**, 502.
Subramanian A, Tamayo P, Mootha VK, et al (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**, 15545-50.
Teresa Gomez CM, de la Iglesia S, Perera M, et al (2002). Renin expression in hematological malignancies and its role in the regulation of hematopoiesis. *Leuk Lymphoma*, **43**, 2377-81.
Tian L, Greenberg SA, Kong SW, et al (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA*, 102, 13544-9.
Tsai CA, Chen JJ (2009). Multivariate analysis of variance test for gene set analysis. *Bioinformatics*, **25**, 897-903.
van Laarhoven JP, Spierenburg GT, Bakkeren JA, et al (1983). Purine metabolism in childhood acute lymphoblastic leukemia: biochemical markers for diagnosis and chemotherapy. *Leuk Res*, **7**, 407-20.
Zölzer F, Basu O, Devi PU, et al (2010). Chromatin-bound PCNA as S-phase marker in mononuclear blood cells of patients with acute lymphoblastic leukaemia or multiple myeloma. *Cell Prolif*, **43**, 579-83.