Evaluating the False Positive Rates of Gene Expression Profile Analysis Approaches

Cao Yiqun

A0105560R

Undergraduate Research Opportunity Program

Project Report

Computational Biology

Faculty of Science

National University of Singapore

AY 2013/2014 Semester 2

Module code: ZB3288

Project number: 13245

Supervisor: Professor Wong Limsoon

Number of words: 1805

Abstract

The study of microarray experiments allows description of genome-wide expression changes in health and disease. Many different gene expression profile analysis methods have been applied to identify the significant genes in the microarray experiments. This report attempts to evaluate the reliability of p-values provided by two popular permutation involved analytical technics. Two main methods were designed to check the reliability of each profile technic individually.

1. Introduction

The study of microarray experiments allows description of genome-wide expression changes in health and disease. Different mechanisms are used to monitor expression levels for thousands of genes simultaneously. The selection of differentially expressed genes is a very important stage of microarray data analysis and involves the use of methods that can be used when the number of features is much larger than the number of samples. Two analytical techniques are chosen to conduct the experiments of evaluating the reliability of the provided p-value.

One technique is Significance Analysis of Microarrays (SAM) (Tusher et al. 2001). It identifies genes with statistically significant changes in expression by gene-specific t tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are deemed potentially significant. Permutations of the measurements are used to estimate the false discovery rate (FDR), which is defined as the expected percentage of false positives among all the claimed positives.

Another technique, Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), is a computational method that determines whether a prior defined gene set shows statistically significant, concordant differences between two biological states. Its focus is gene sets, which are groups of genes that share common biological function, chromosomal location, or regulation. The goal of GSEA is to determine whether members of a gene set tend to occur toward the top (or bottom) of the list, in which case the gene set is correlated with the phenotypic class distinction. Phenotype label permutation is used by GSEA to compute statistical significance. Leading edge analysis of GSEA gives the subset of genes that contributes to score the maximum enrichment score for the high-scoring gene sets.

This study mainly focuses on the false positive outcomes of individual gene expression analytical technique and does not consider the issue of proper overlapping genes returned by different technics.

2. Experimental design

Two control experimental methods were designed. In general, the first method was designed to control all the returned significant genes to be false positive, while the second method assessed the agreement between two resulting lists.

The reference data was from a study of lung cancer in Boston (Bhattacharjee et al., 2001). The original dataset contained a total of 203 specimens, including histologically defined lung adenocarcinomas (n = 127), other related adenocarcinomas (n = 12), squamous cell lung carcinomas (n = 21), pulmonary carcinoids (n = 20), SCLC (n = 6) cases, and normal lung (n = 17) specimens. In order to obtain clear and reliable

outcomes, only squamous cell lung carcinomas (SCC) and normal lung (NL) specimens were selected to design the datasets for two control experimental methods.

Method 1.

Studies showed that the disease types were usually classified into some subtypes. For example, primary lung adenocarcinoma was observed with four subclasses (Bhattacharjee et al., 2001), and SCC consisted of high and low risk clusters. Compared with these disease types, control type samples were much less diverse and hence, more suitable to ensure that the claimed significant genes were false positives.

Therefore, 17 NL specimens were assumed to have no significant gene expression difference. They were randomly separated into two classes and artificially assigned with different phenotypes Normal_1 (n=8) and Normal_2 (n=9). Based on the pervious assumption, the expected significant genes return by two technics would be controlled within the number of given false positives.

Method 2.

SCC was discovered to have some gene markers, such as CCND1, encoding cyclin D1 and TP73L, encoding p63 ("Lung: Non-small cell carcinoma," n.d.). This suggested that it would be easy to detect the significant genes between SCC and NL. Although SCC samples could contain subtypes, the differences between subtypes can be ignored when compared to the control samples.

Based on this, 21 SCC samples were first assumed to have the same gene expression levels. 10 out of 21 SCC and 8 out of 17 NL samples were randomly selected and combined together to generate the first dataset Data_2(1). Another 10 SCC and 8 NL were then chosen randomly from the rest samples in the similar manner to generate the

second dataset Data_2(2). The samples construction of these two datasets was in the same pattern but the samples were not overlapping across the two datasets. This helped to control all the factors to be the same except the expression data of individual samples. After running the profile method using these two datasets independently, two lists of significant genes would be generated. The difference between the two lists was expected to within set FDR. Jaccard index was computed to check the agreement degree of these two gene lists.

3. Results and Discussions

3.1. SAM results

For each gene i, the relative difference d(i) is a value that incorporates the change in expression between conditions. And the expected relative difference $d_E(i)$ is derived from controls generated by permutations of data. When the difference between d(i) and $d_E(i)$ is greater than a fixed threshold delta, gene i is considered to be significant. In plot d(i) vs $d_E(i)$, the more a gene deviates from the d(i) = $d_E(i)$ line, the more likely it is to be significant. The mean number of genes exceeding cutoffs defined by delta in the permuted data gives an estimate for FDR. Larger delta will usually give fewer significant genes and lower FDR.

SAM analysis was carried out using unpaired (2 class) option, with number of permutations 1500. 12600 native features of input dataset were collapsed into 9078 genes with gene symbols before run using R studio.

Method 1

The corresponding false positive table was shown in table 1. For a medium delta = 0.4 (plot d(i) vs d_E(i) shown in Figure 1), 64 significant genes were called with claimed 22 false positives. However, according to the previous assumption, called genes should be all the false positives. Table 2 presented the significant genes list with setting FDR 5% and 9 genes were returned, while the real FDR should be 100% under the assumption Hence, the provided false positive and FDR both underestimated the true values.

Method 2.

As shown in Table 3, the numbers of called genes of two sets of data with same delta value were not equal, even though the score plots (delta = 0.3, Figure 2) appeared to have similar patterns for two datasets.

There were large numbers of significant genes returned for both datasets. In order to focus on the most different expression genes, FDR<1% was used to call the significant genes. The numbers of significant genes returned for Data_2(1) and Data_2(2) were 1445 and 1121 respectively (data not shown). 848 common genes were identified using python programming. Hence, the Jaccard index was computed as 848/1718 = 49.36%. The similarity between two were relatively high. However, with FDR < 1%, the largest number of total different genes between two lists should be 14.15+11.21 = 25.66, under the assumption that the same true positives were returned in two lists. The influence of false negatives should have small effect on above assumption, since the data patterns were similar and the same mechanism (SAM) was used to select genes. Whereas, the actual number of different genes were (1445-848) + (1121-848) = 870, which was much greater than the value derived from the provided FDR.

3.2. GSEA results

Enrichment score (ES) reflects the degree to which a gene set is overrepresented at the extremes of the entire ranked gene list. Normalized enrichment score (NES) adjusts ES for multiple hypotheses testing through normalizing the ES for each gene set to account for the size of the set. The higher NES with smaller FDR gives more significant gene sets.

GSEA analysis was conduced using curated gene sets C2 collection with number of permutations 1000. C2 collection contains 4722 gene sets and collected from various curated sources such as online pathway databases. The results showed that in total of 3403 gene sets were used in the analysis.

Method 1

The numbers of significant gene sets were shown in table 4. For phenotype Normal_1, 946 gene sets were considered to be significantly up-regulated with FDR<25%, which also underestimated the expected 100% FDR. With p-value < 5%, there would be at most 170 false positives, and it was relatively close to the result 182. While for phenotype Normal_2, the returned genes were well controlled under the given false positive rate.

Method 2

Set the common threshold FDR < 5%, 223 and 648 gene sets were considered to be significant for Data_2(1) and Data_2(2) respectively (data not shown), with 207 common gene sets. Hence, Jaccard index for gene sets was 31.2%.

In order to detect the significant genes, 6 common gene sets with high NES absolute values (greater than 2) for both datasets were selected to conduct leading edge analysis. For the two datasets, there were respectively 250 and 274 gene members from these 6

gene sets that appeared on the ranked gene list. Among these leading edge subset genes, 240 of them were shared, which gave Jaccard index 84.5% and were higher than that given by SAM.

4. Conclusions

In conclusion, both SAM and GSEA tend to underestimate the real false positive rate. Generally, GSEA gives relatively more reliable false positive values. It is based on the gene sets from the same gene ontology, which is different from pure statistical analysis. The genes returned by GSEA would be more biologically significant. In addition, permuting the phenotype labels of GSEA instead of the genes maintains the complex correlation structure of the gene expression data and hence produces fewer false positives and provides a more stringent assessment of significance.

More studies should be done for further evaluation of false positive rate given by different gene expression methods. For example, MAPPFinder program involving permutation of gene labels, and GSA uses both gene and sample permutations and rotation tests to estimate p-values. In addition, different datasets with larger sample size need to be accessed to come to a generalizing conclusion.

References

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., … Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(24), 13790–13795. doi:10.1073/pnas.191502998

Lung: Non-small cell carcinoma. (n.d.). Retrieved April 16, 2014, from http://atlasgeneticsoncology.org/Tumors/LungNonSmallCellID5141.html

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Lander, E. S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation respone. Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5116-21. Epub 2001 Apr 17. Erratum in: Proc Natl Acad Sci U S A 2001 Aug 28;98(18):10515. PubMed PMID: 11309499; PubMed Central PMCID: PMC33173.

Tables and Figures

3.Results

3.1. SAM

Method 1

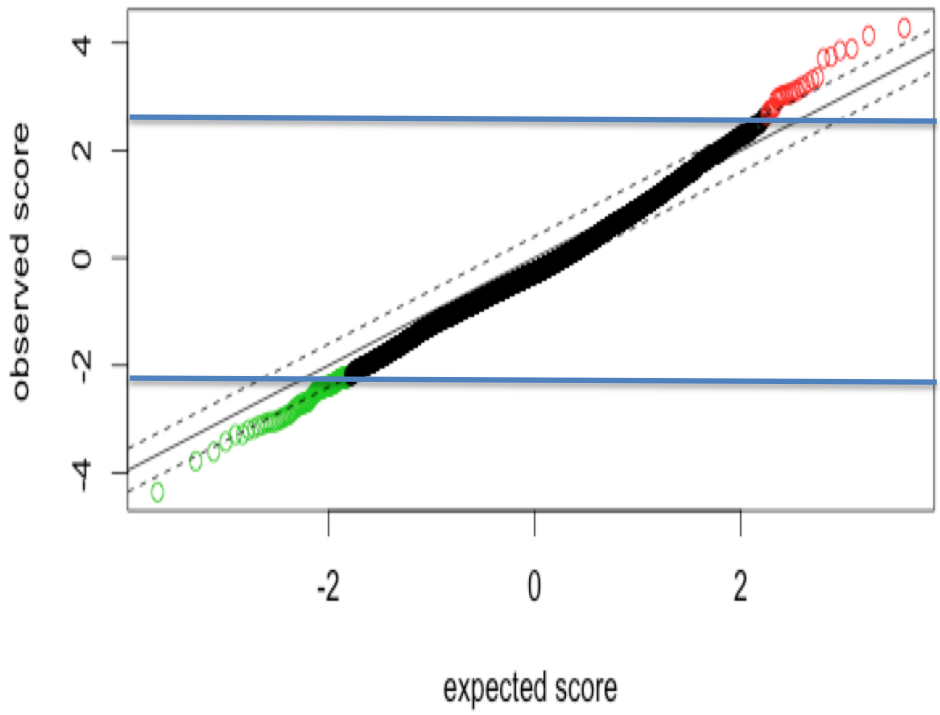
Table 1. SAM false positive results for designed datasets

delta	#false pos	#called	FDR
0.2	2967.076686	4600	0.6450167
0.3	195.687528	378	0.5176919
0.4	22.688409	64	0.3545064
0.5	10.209784	42	0.2430901
0.6	3.403261	22	0.1546937
0.7	0	6	0
0.8	0	6	0

Table 2. SAM Significant gene list for designed datasets.

	Row	Gene ID	Gene Name	Score(d)	Fold Change	q-value(%)
1	5429	g5428	TOMM34	4.2776694	7.58E+85	0
2	6263	g6262	ZNF10	4.1355352	1.479E+18	0
3	4393	g4392	FAT2	3.8922434	9.54E+59	0
4	6084	g6083	RHOBTB2	3.8578756	1.027E+14	0
5	1396	g1395	KLF7	3.740733	3.39E+63	0
6	7920	g7919	LST1	3.7092689	2.37E+126	0
7	5175	g5174	RPS6KA4	3.358908	1.418E+17	0
8	3421	g3420	USP9X	-4.357455	1.39E-13	0
9	1523	g1522	SNX4	-3.781932	8.46E-29	0

Figure 1. SAM plots for two designed datasets.



Method 2

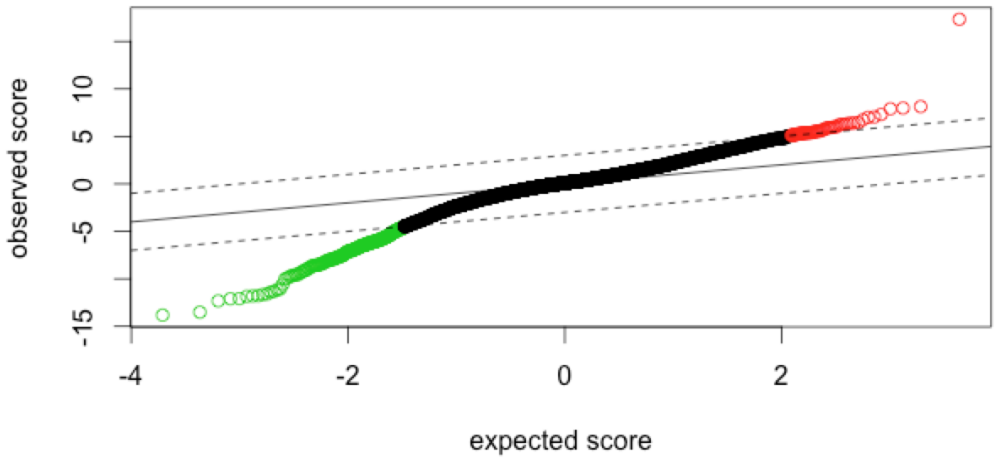
Table 3. SAM false positive results for designed datasets

Data_2(1)			
delta	# false pos	# called	FDR
0.5	862.1871	4083	0.211165
1.0	77.56479	2194	0.035353
2.0	1.2820626	884	0.00145
3.0	0	376	0
4.0	0	191	0

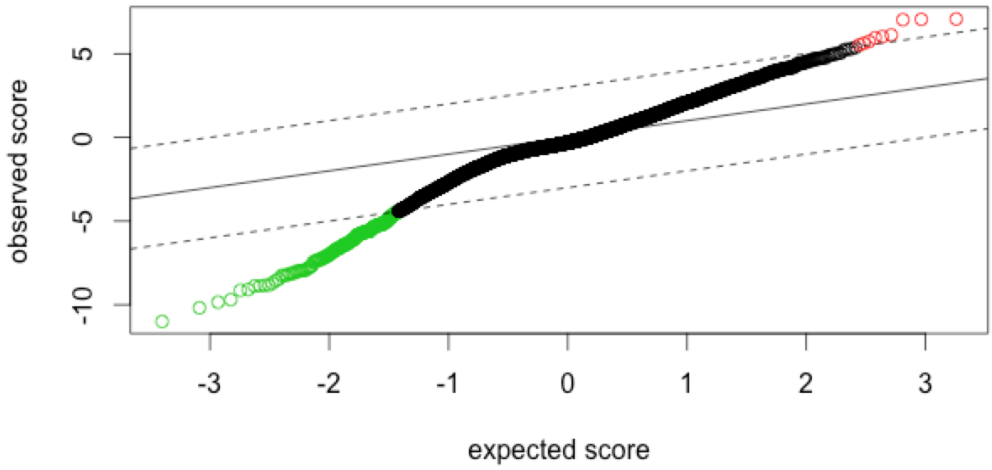
Data_2(2)			
delta	# false pos	# called	FDR
0.5	801.8171	3345	0.23971
1.0	84.513001	1865	0.045315
2.0	0.704275	658	0.00107
3.0	0	239	0
4.0	0	92	0

Figure 2. SAM plots for two designed datasets.

(a) Data 2(1)



(b) Data_2(2)



2. GSEA

Table 4. GSEA results for designed datasets in Method 1.		
Phenotype	Normal_1	Normal_2
# Up-regulated gene sets	2898	505
# Significant gene sets (FDR < 25%)	946	0
# Significant gene sets (nominal p < 5%)	182	2

Acknowledgements

The project could not have been performed without the help and contribution of many individuals. I would like to extend my gratitude to my supervisor, Professor Limsoon Wong for offering me this project and guiding me through the research process. I would like to thank the coordinator Professor Zhang Louxing and Professor Greg Tucker-Kellogg for necessary help and suggestions.