이 학 석 사 학 위 논 문

Multivariate approach to the analysis of correlated RNA-seq data

상호연관된 리보 핵산 시퀀싱 자료 분석의 다변량적 접근법

2017 년 2 월

서울대학교 대학원

자연과학대학 통계학과

박 현 진

Multivariate approach to the analysis of correlated RNA-seq data

by

Hyunjin Park

A thesis
submitted in fulfillment of the requirement
for the degree of Master
in
Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
Feb, 2017

Multivariate approach to the analysis of correlated RNA-seq data

지도교수 박 태 성

이 논문을 이학석사 학위논문으로 제출함

2017 년 2 월

서울대학교 대학원

자연과학대학 통계학과

박 현 진

박현진의 이학석사 학위논문을 인준함

2017 년 2 월

위 원 장 조 신 섭 (인)

부위원장 박 태 성 (인)

위 원 이 영 조 (인)

Abstract

Multivariate approach to the analysis of correlated RNA-seq data

Hyunjin Park

Department of Statistics

The Graduate School

Seoul National University

High-throughput RNA-seq technology has emerged as a powerful tool for understanding the molecular basis of phenotype variation in biology, including disease. Recently, some correlated RNA-seq datasets started to be generated. While there have been several approaches proposed for identifying the differentially expressed genes (DEGs), not many methods can analyze correlated RNA-seq data. We expect the simultaneous analysis of correlated RNA-seq data to increase of power of detecting DEGs. In this paper, we propose a multivariate method to find DEGs on correlated RNA-seq data based on the Generalized Estimating Equations (GEE) approach. The

advantage of the proposed method is to consider correlated RNA-seq data simultaneously while accounting for correlations. Through real data analysis and simulation studies, we show that our multivariate approach has higher power of detecting DEGs than the existing methods.

Key words: RNA-seq, Differentially Expressed Gene (DEG), simultaneously, correlation, multivariate, Generalized Estimating Equations (GEE)

Student number: 2015-20300

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Microarrays have been widely used to perform gene expression analysis for more than 10 years. Recently, high-throughput sequencing of RNA sample (RNA-seq) has become an attractive method [1] to study the molecular basis of phenotype variation in biology, including disease [2].

A common purpose in analyzing RNA-seq data is to determine which genes are differentially expressed under several different experimental conditions. Recently, several statistical methods such as edgeR [3], DESeq [4], LPEseq [5], limma+voom [6] have been proposed to accomplish this goal. These methods use Poisson model to be extended via quasi-likelihood or negative binomial distribution to account for over-dispersion. These methods were reported that they are powerful to analyze the independent RNA-seq data [7] in several review papers.

Recently, some correlated RNA-seq datasets started to be generated. For example, the Genotype-Tissue Expression (GTEx; See : http://www.gtexportal.org/) project examined pattern of gene expression levels across tissues. It is reported that multiple brain regions are strongly correlated as a single unit [8]. However, there are also exist a distinct regions so that they are weakly correlated each other. For the case of the correlated RNA-seq data, the performances of pre-existing univariate method are not examined.

1.2 Purpose

The present paper describes statistical method to find DEGs in correlated RNA-seq dataset. We applied the Generalized Estimating Equations (GEE) method to consider correlations between RNA-seq data. The advantages of the GEE method are that it does not require a specification of a joint distribution and it uses more information from given dataset by considering correlation. Through various real data analysis and simulation studies, we compare the results obtained from the GEE and other univariate methods. To compare the results of GEE, we used edgeR, DESeq, limma+voom, LPEseq methods, because comprehensive review papers reported that these methods perform much better than other univariate methods [7].

2. Materials and Methods

In this part commonly used univariate methods and proposed multivariate method are briefly reviewed. Two different real datasets are also described in detail.

2.1 Real RNA-seq datasets

We used two RNA-seq datasets to investigate the performance of each method containing GEE. The characteristics of these datasets are described in detail below.

2.1.1 Diet data

We used the RNA-seq data generated by Kyungpook National University (KNU) to check the Korean traditional drug's effect across

two organs (adipose tissue, liver). This dataset consisted of four groups. Each dataset contains three or four samples of A/J mouse. The dataset provide a case in which a small number of correlated datasets are available.

2.1.2 Toxicity data

We studied RNA-seq data generated by Ying Yu et al. [9] to investigate the chemical toxicities effect across developmental stages (juvenile, adolescent, adult and aged). We analyzed the DEGs between juvenile and aged of male rats. Within each developmental stages, this dataset consisted of eight rats (four male and four female rats) from 11 organs (Adrenal Gland, Brain, Heart, Kidney, Liver, Lung, Muscle, Spleen, Thymus and Testes for males and Uterus for female rats). For analysis between any other two developmental stages, we can do in the same manner. The dataset provide a case in which a large number of correlated datasets are available.

2.2 Review of commonly used approach

The following univariate methods were considered: edgeR, DESeq, and limma+voom. The main characteristics of these univariate methods are summarized in TABLE 1.

2.2.1 edgeR [3]

edgeR was developed for performing differential expression test using count data under a negative binomial model especially on experiments with small numbers of replicates. Trimmed Mean of M values (TMM), Relative Log Expression (RLE), or Upper quantile normalization method can be used to calculate normalization factors between samples. We use empirical Bayes procedure to adjust over-dispersion across genes. Finally, exact test or Generalized Linear Model (GLM) is used to find DEGs. It is reported that edgeR has a relatively high power in general. However, edgeR suffers from high false discovery rate in many cases [7].

2.2.2 DESeq [4]

DESeq was extension of edgeR method by allowing more general relationships of variance and mean and using DESeq

sizeFactors to adjust for different sequencing depth. Local regression between mean and variance is used to find these relationships and estimate over-dispersion. Finally, exact test adapted to over-dispersed data is used to find DEGs. DESeq is reported to be so conservative that it has relatively lower power in general [7].

2.2.3 limma+voom [6]

limma was widely used method to analyze microarray data. Nowadays, the 'voom' transformation gives immediate access to RNA-seq analysts. limma is based on gene-wise linear model and uses TMM normalization to adjust for different sequencing depth. By using empirical Bayes method, limma can detect DEGs. This method has a great ability to control type1 error. However, limma has a lower power for small sample size [7].

Table 1. Characteristic of the commonly used univariate methods.

Method	Normalization method	Differential expression test	Other characteristic
edgeR	TMM/RLE/Upper quantile	Exact test, GLM	Generally high TPR Poor FDR control in many cases
DESeq	DESeq sizeFactors	Exact test	Generally low TPR Good FDR control for larger sample sizes
limma+voom	TMM	Empirical Bayes method	Good type I error control Low TPR for small sample sizes.

2.3 Proposed approach : GEE method

The GEE method is the extension of the Generalized Linear Model (GLM) with a proper correlation structure via quasi-likelihood approach [10]. We assume that Poisson distribution as the marginal distribution to account for count data. We assume that the correlation among $\{Y_t\}$ is unstructured, where $\{Y_t\}$ represents an expression of the $\{t^{th}\}$ gene. We can assume other correlation working structures. It is well known that choosing a working correlation structure well approximating the true correlations can pay benefits regarding efficiency of estimation of the model parameters. After that we adjusted the over-dispersion dividing the Wald type statistics by the scale parameter which can be calculated by gee package in R [11].

To adapt the GEE method for correlated RNA-seq data, we relate the marginal response $\mu_{ij} = E(y_{ij})$ to a linear combination of the covariates using link function $g(\mu_{ij}) = \log(\mu_{ij})$. For simplicity, the number of correlated RNA-seq datasets is set to *k* and the number of groups is set to two. It can be easily extended to more than two groups by using dummy variables

$$\begin{bmatrix} g(\mu_{i1}) \\ \vdots \\ g(\mu_{ik}) \end{bmatrix} = \begin{bmatrix} 1 & x_{i1} & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 & x_{ik} \end{bmatrix} \begin{bmatrix} \mu_{i1} \\ \alpha_{i1} \\ \vdots \\ \mu_{ik} \\ \alpha_{ik} \end{bmatrix}$$

where α_{ij} are coefficients of group effects and y_{ij} are gene expression which follows a Poisson distribution with mean μ_{ij} for i^{th} gene and j^{th} correlated RNA-seq data. Other types of design matrixes are also possible. For example, if we can assume that the covariate effects to genes are same in all of the correlated dataset, we can just use 1 covariate effect variable to increase power.

$$\begin{bmatrix} g(\mu_{i1}) \\ \vdots \\ g(\mu_{ik}) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & x_{i1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & x_{ik} \end{bmatrix} \begin{bmatrix} \mu_{i1} \\ \vdots \\ \mu_{ik} \\ \alpha_i \end{bmatrix}$$

So this method is flexibly applicable to correlated RNA-seq dataset.

3. Simulations

To compare the GEE approach with other existing univariate methods, we generated correlated and over-dispersed count data using copula package in R [12]. Each marginal distribution is assumed to follow a negative binomial distribution with mean and over-dispersion parameter estimated from real data. The purpose of this simulation study was to investigate power. So we generated 10,000 genes in total for each dataset. The number of group was set to two and the number of sample in each group was set to five. We randomly selected κ% of the total genes as DEGs and added effect size δ in one of randomly selected group. φ% of DEGs in each correlated dataset was set to DEGs in all of the organs simultaneously. The number of correlated datasets was set to n and the correlation between the correlated datasets was set to 0.7. In our various simulation setting, we varied the value of κ, δ, φ, and n as stated below and investigated the power.

3.1 Simulation Settings

3.1.1 Different number of DEGs

The number of DEGs can vary due to the biological phenomena of interest. We set 0%, 5%, 10%, 20% and 30% of total genes as to be DEGs and observed each method's type 1 error rate and power. In this scenario, δ was set to 400. These values were estimated from real dataset and φ was set to 10 [8].

3.1.2 Different value of φ

The proportion of DEGs in all of the correlated datasets simultaneously can vary due to the environment of the data and relationship between datasets. We set 10%, 20%, 30% and 40% as the value of φ. In this scenario, $\kappa, \delta,$ and n was set to 20, 400 and 2 respectively.

3.1.3 Different number of correlated datasets

The number of correlated datasets can vary due to the limitation of budget. We varied the number of correlated datasets as 2,

3, 5, and 10. In this scenario, $\kappa, \delta,$ and φ was set to 20, 400 and 10 respectively.

3.2 Results of Simulation

We investigated the performance of GEE with other univariate methods for finding DEGs in all of the correlated datasets simultaneously. We apply four univariate methods (edgeR, DESeq, limma+voom, LPEseq) to each organ separately and select DEGs in all correlated datasets simultaneously with multiple comparison adjustment such as FDR [13] and Bonferroni correction.

Table 2. Type 1 Error rate of each methods.

First column shows the number of DEGs in each datasets. However, there are no DEGs which are set to DEGs in all of the correlated datasets simultaneously.

# of DEGs in each datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
0	0.0087	0.0043	0.0063	0.0038	0.0135
500	0.0157	0.0098	0.0119	0.0090	0.0184
1,000	0.0260	0.0182	0.0215	0.0164	0.0241
2,000	0.0766	0.0601	0.0632	0.0448	0.0537

First, we investigated type1 error rate of each methods at a nominal p-value threshold 0.05. Type 1 error rate is presented in Table2. The type1 error rate of GEE method is greater than other univariate methods. However, the type1 error rate of GEE is also small.

Next, we compare the power and false discovery rate of each method according to the several scenario mentioned above. When the number of DEGs is small, edgeR is the most powerful but the power is not significantly different in each method. However, when the number of DEGs is large, the GEE method find much more DEGs compared with other univariate methods. DESeq and limma+voom methods didn't identify enough number of DEGs because these methods are so conservative. In other simulation scenario, we can find that edgeR is the most powerful to detect DEGs. However, edgeR has a problem to controlling false positive rate. Although DESeq and limma+voom has relatively low false positive rate, these methods also has low power to detect DEGs. LPEseq and GEE have moderately high power controlling false positive rate.

Table 3. Power of each method in scenario 1

First column shows the number of DEGs in all of the correlated datasets simultaneously. Each datasets has 10 times more DEGs than the number of DEGs presented in first column.

# of DEGs in all of the datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
50	0.9155	0.2113	0.6761	0.8592	0.6620
100	0.8967	0.4510	0.4185	0.8750	0.7500
200	0.8952	0.5588	0.3971	0.8824	0.8640
300	0.8230	0.5438	0.5621	0.8778	0.8932

Table 4. False Discovery Rate (FDR) of each method in scenario 1

First column shows the number of DEGs in all of the correlated datasets simultaneously. Each datasets has 10 times more DEGs than the number of DEGs presented in first column.

# of DEGs in all of the datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
50	0.0299	0.0000	0.0000	0.0000	0.1455
100	0.0462	0.0119	0.0128	0.0183	0.0980
200	0.1081	0.0380	0.0270	0.0123	0.0562
300	0.2373	0.1227	0.1487	0.0098	0.0472

Table 5. Power of each method in scenario 2

First column shows the proportion of DEGs in all of the correlated datasets simultaneously in each datasets.

proportion of DEGs in all of the datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
10 %	0.8952	0.5588	0.3971	0.8824	0.8640
20 %	0.8922	0.5790	0.4267	0.8836	0.8650
30 %	0.8896	0.5888	0.4353	0.8820	0.8617
40 %	0.8869	0.5895	0.4346	0.8806	0.8576

Table 6. FDR of each method in scenario 2

First column shows the proportion of DEGs in all of the correlated datasets simultaneously in each datasets.

proportion of DEGs in all of the datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
10 %	0.1081	0.0380	0.0270	0.0123	0.0562
20 %	0.0854	0.0290	0.0198	0.0081	0.0429
30 %	0.0691	0.0252	0.0144	0.0071	0.0369
40 %	0.0515	0.0192	0.0095	0.0036	0.0308

Table 7. Power of each method in scenario 3

First column shows the number of correlated datasets.

# of correlated datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
2	0.8952	0.5588	0.3971	0.8824	0.8640
3	0.8504	0.4961	0.2992	0.8898	0.8150
5	0.7610	0.2927	0.1171	0.7951	0.7317
10	0.7000	0.2350	0.0850	0.7150	0.6100

Table 8. FDR of each method in scenario 3

First column shows the number of correlated datasets.

# of correlated datasets	edgeR	DESeq	limma+voom	LPEseq	GEE
2	0.1081	0.0380	0.0270	0.0123	0.0562
3	0.0809	0.0233	0.0130	0.0044	0.0282
5	0.0250	0.0164	0.0400	0.0000	0.0066
10	0.0000	0.0000	0.0000	0.0000	0.0000

4. Application to Real Data

We analyzed two different real RNA-seq datasets by using edgeR, limma+voom, and GEE methods: diet dataset with a small number of correlated organs, and toxicity dataset with a large number of correlated organs.

First, we examined the diet dataset. This dataset has only two organs. The Pearson correlation between organ's gene expression levels are about 0.7 which indicate that the two datasets are highly correlated. We compared the number of DEGs found by each method. Figure2 shows the result using venn diagram. limma+voom, edgeR and GEE method identified 937, 161 and 1900 genes as DEGs after FDR correction[13], respectively. Among these, 62 genes were commonly identified in all methods.

Second, we examined the toxicity dataset. This dataset has ten organs (Adrenal Gland, Brain, Heart, Kidney, Liver, Lung, Muscle,

Spleen and Testes for males and Uterus for female mouse).The Pearson correlation between organs is presented in Figure3. This figure indicates that the ten organs are correlated with each other. We compared the result of GEE with those of edgeR and limma+voom. Figure2 shows the result of venn diagram for limma+voom, edgeR and GEE method which identified 41, 9 and 44 genes as DEGs after FDR correction, respectively. Among these, two genes were commonly identified by all methods and 30 genes were additionally called as DEGs by GEE method. This additional set of genes contains Ect2 and Ndc80 DEGs. Ect2 is reported to be an oncogene in multiple human cancers [14]. Ndc80 is also reported to be related in benign tumor cells [15]. However, other genes in the list need to be investigated.

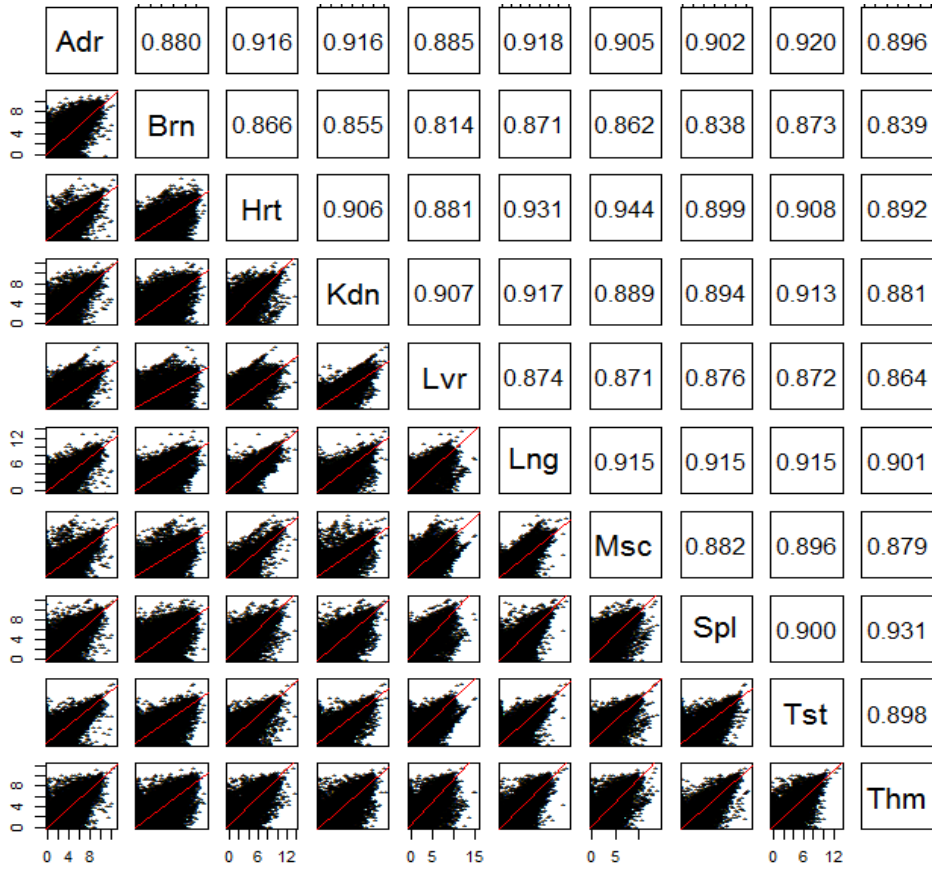

Figure 1. Correlation between organs in toxicity dataset.

The plot represents the correlation between ten organs. The right upper corners represent correlations between each organ's gene expression levels and the left bottom corners represent scatter plots between gene expression levels.

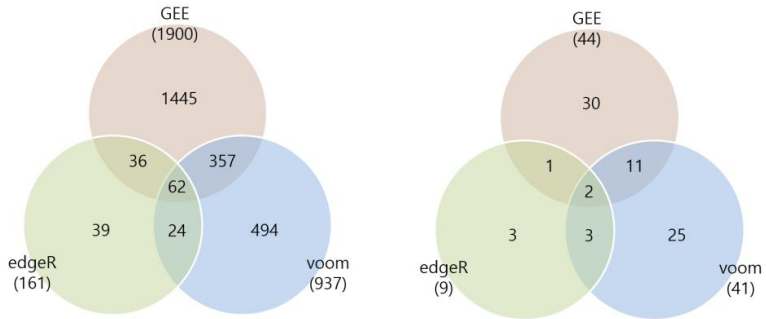

Figure 2. Venn-diagrams of the number of DEGs.

The venn-diagram represent the number of DEGs in all of organs in a significance threshold put at an adjusted p-value of 0.05. GEE (green), limma+voom (red), edgeR (blue) were used. The left venn-diagram represents the result of diet datasets and right venn-diagram represents the result of toxicity datasets.

5. Discussion

We proposed a method for testing DEGs in correlated RNA-seq data. By using GEE method, we can increase power to detect DEGs in all of the correlated datasets simultaneously.

Our comparison through real and simulation studies shows that the GEE method detects more DEGs compared with other widely used univariate methods. In toxicity dataset, we examined 30 genes which were additionally found by GEE method and found out that some of these genes are already reported as a significant genes. In our simulation studies, we compared the performance of these method using power and false discovery rate (FDR). Compared to the univariate methods, the GEE method was shown to provide a similar power for the datasets with a small number of DEGs, and outperformed the other univariate methods for the datasets with a large number of

DEGs. Compared with others, the GEE method finds much more true positive genes, while it also finds more false positive genes.

.

Bibliography

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNASeq. Nature methods, 5(7), 621-628.

2. Chen, G., Wang, C., & Shi, T. (2011). Overview of available methods for diverse RNA-Seq data analyses. Science China Life Sciences, 54(12), 1121-1128.

3. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.

4. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology, 11(10), 1.

5. Gim, J., Won, S., & Park, T. (2016). LPEseq: Local-Pooled-Error Test for RNA Sequencing Experiments with a Small Number of Replicates. PloS one, 11(8), e0159182.

6. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, gkv007.

7. Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. BMC bioinformatics, 14(1), 1

8. GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, 348(6235), 648-660.

9. Ying Yu, James C. Fuscoe, Chen Zhao, Chao Guo, Meiwen Jia, Tao Qing, Desmond I. Bannon, Lee Lancashire, Wenjun Bao, Tingting Du, Heng Luo, Zhenqiang Su, Wendell D. Jones, Carrie L. Moland, William S. Branham, Feng Qian, Baitang Ning, Yan Li, Huixiao Hong, Lei Guo, Nan Mei, Tieliu Shi, Kevin Y. Wang, Russell D. Wolfinger, Yuri Nikolsky, Stephen J. Walker, Penelope Duerksen-Hughes, Christopher E. Mason, Weida Tong, Jean Thierry-Mieg, Danielle Thierry-Mieg, Leming Shi, and Charles Wang, 'A Rat Rna-Seq Transcriptomic

Bodymap across 11 Organs and 4 Developmental Stages', Nat Commun, 5 (2014)

10. Agresti, A., & Kateri, M. (2011). Categorical data analysis (pp. 206-208). Springer Berlin Heidelberg

11. Carey VJ (2002). gee: Generalized Estimation Equation Solver. R package version 4.13-10; Ported from S-PLUS to R by Thomas Lumley (versions 3.13 and 4.4) and Brian Ripley (version 4.13).

12. Yan, J. (2007). Enjoy the joy of copulas: with a package copula. Journal of Statistical Software, 21(4), 1-21.

13. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

14. Fields, A. P., & Justilien, V. (2010). The guanine nucleotide exchange factor (GEF) Ect2 is an oncogene in human cancer. Advances in enzyme regulation, 50(1), 190.

15. Bièche, I., Vacher, S., Lallemand, F., Tozlu-Kara, S., Bennani, H., Beuzelin, M., ... & Cizeron-Clairac, G. (2011). Expression analysis of mitotic spindle checkpoint genes in breast carcinoma: role of NDC80/HEC1 in early breast tumorigenicity, and a two-gene signature for aneuploidy. Molecular cancer, 10(1), 1.

초 록

약 10 여 년이 넘는 기간 동안 리보 핵산 시퀀싱 기술은 질병을 비롯하여 생물학에서 분자기반의 표현형 변이를 이해하는 강력한 도구로 부상해 왔다. 또한, 최근에는 동일 사람의 다른 장기로부터 얻은 데이터, 형제자매로부터 얻은 데이터 등 상호 연관된 리보 핵산 시퀀싱 데이터들이 생성되기 시작했다.

리보 핵산 시퀀싱 데이터에 대해서 그룹에 따라 서로 상이하게 발현 유전자 (DEG)를 찾는 많은 방법들이 알려져 있지만, 이러한 상호 연관된 데이터들에 대한 분석 방법은 많이 알려져 있지 않다. 이러한 이유로 우리는 여러 데이터에서 동시에 상이 발현되는 유전자를 규명하기 위한 방법을 제시하고자 한다.

본 논문에서 우리는 일반화 추정 방정식 (Generalized Estimating Equation)을 통해 여러 장기에서 동시에 발현되는 유전자를 규명하는 모델을 만들고 이에 대한 성능평가를 하였다. 일반화 추정 방정식의 장점은 리보 핵산 시퀀싱 데이터들 사이의 상관관계를 고려한 분석을 실시하고 데이터로부터 더 많은 정보를 이용하고 이를 통하여 검정력을 높일 수 있다는 것이다. 다양한 시뮬레이션 및 실제 데이터 분석을 통하여 이러한 다변량적 방법이 기존에 제시된 방법에 비해 유사하거나 높은 검정력을 가진다는 사실을 확인할 수 있었다.

주요어: 리보 핵산 시퀀싱, 다변량, 일반화 추정 방정식 (GEE), 상이 발현 유전자 (DEG)

학 번: 2015-20300