이학석사학위논문

Joint identification of differentially expressed gene and phenotype associated genes

특정형질과 유의성과 차별 발현 유전자 공동 판별

2013년 2월

서울대학교 대학원

협동과정 생물정보학 전공

조성환

Joint identification of differentially expressed gene and phenotype associated genes

지도교수 박태성

이 논문을 이학석사학위논문으로 제출함

2013년 2월

서울대학교 대학원

협동과정 생물정보학 전공

조성환

조성환의 석사학위논문을 인준함

2013년 2월

위 원 장 김 선 (인)

부 위 원 장 박 태 성(인)

위 원 천 종 식(인)

Abstract

Samuel Sunghwan Cho
Interdisciplinary Program in Bioinformatics
The Graduate School
Seoul National University

The emergence of a wide variety of new techniques has led to the production of diverse types of biological data. Among them microarray technology has brought innovative changes in biological field and is still most commonly used in various research fields. For the last decade, many analytical methods and tools have been developed. In general, the detection of differentially expressed genes (DEGs) among different treatment groups is often a primary purpose of microarray data analysis. In addition, the association studies investigating the relationship between genes and the phenotype of interest such as survival time became also popular in microarray data analysis. Such association analysis provides the list of phenotype associated genes (PAGs). In this study, I consider a joint identification of DEGs and PAGs in microarray data analyses. The first approach is a naïve approach which detects DEGs and PAGs separately, and then identifies the intersection genes of PAGs and DEGs. The second approach is a hierarchical approach which detects DEGs first and then chooses PAGs among DEGs, or visa versa. I propose a new model-based approach for a joint

identification of DEGs and PAGs simultaneously. Through a real microarray data analysis, I show that our model-based approach provides a more powerful result than the naïve approach and the hierarchical approach.

Keyword: Differential expression genes (DEGs), Phenotype associated genes (PAGs), Joint identification, association study, Linear regression model.

Student number: 2010-23163

Content

Abstract i
Content iii
List if Tables iv
List of Figures v
1. Introduction 1
1. 1. Microarray technology 1
1. 2. DEGs and PAGs 3
1. 3. Purpose of research 6
2. Materials and Methods 8
2. 1. Data 8
2. 2. DEG detection 9
2. 3. PAG detection 10
2. 4. Joint identification 13
3. Result 16
3. 1. DEGs 16
3. 2. PAGs 19
3. 3. Joint identification 23
4. Simulation study 26
5. Discussion 29
6. Reference 31
Abstract 37

List of table

Table 1. Top significant genes list for t-test.....18

Table 2. Top significant genes list for SAM.....18

Table 3. The number of significant genes from joint identification.....24

Table 4. The gene list from model-based approach.....25

List of figure

Figure 1. The problem of model M1.....12

Figure 2. Correlation plot for phenotype. 20

Figure 3. The venn diagram of PAG.....21

Figure 4. The PAG example plot. 22

Figure 5. Simulation result.....28

1. Introduction

1.1 Microarray technology

The development of various new technologies has greatly affected the biological field. Specifically, the advent of microarray technology provides a crucial turning point in biological research areas (Hal. et al., 2000; Schulze and Downward, 2001; Debouck and Goodfellow, 1999; Gershon, 2002). Microarray technology has been commonly used for identifying the gene expression patterns in cells for thousands of genes simultaneously. Additionally, microarray technology continues to improve in performance aspects regarding sensitivity and selectivity and in becoming a more economical research tool (Heller, 2002). An important emerging medical application domain for microarray technology is clinical decision support in the form of diagnosis of disease as well as the prediction of clinical outcomes in response to treatment (Statnikov. et al., 2005).

Recently, the improvement of microarray technology leads to development of various platforms. So far, many studies have tried to integrate various platforms. For example, the MicroArray Quality Control (MAQC) project provided gene expression levels that were measured from seven different platforms. MAQC study provided a resource representing an important first step toward establishing a

framework for the use of microarrays in clinical and regulatory settings (MAQC Consortium, 2006). In addition, microarray technology has been successfully commercialized, and as results, numerous microarray data have been generated. Several studies have been performed for the integration analysis of microarray data. Meta-analysis was shown to be very powerful in unifying various results of gene expression studies (for example, breast cancer, Wirapati. et al., 2008). Statistical models such as analysis of variance model were shown to be effective in integration analysis to identify genes that have different gene expression profiles in the presence of many controlling variables (Park. et al., 2006).

1.2 DEGs and PAGs

In general, the most common goal of microarray data analysis is to identify differentially expressed genes (DEGs). Microarray technology allows us to produce expression data of target genes more easily than other technologies. Thus, DEGs would become more easily detected by microarray technology than before. In real data application, causal genes of diseases can be easily obtained by discovering DEGs. For the last decade, many statistical methods have been extensively proposed such as t-test, significance analysis of microarray (SAM) (Tusher et al., 2001), the regression modeling approach, the mixture modeling approach (Pan, 2002), and local pooled error (LPE) test (Jain N. et al., 2003).

Among these approaches, a t-test is the most commonly used statistical test for comparing means between two groups. It is a parametric method which requires a normality assumption. However, microarray data rarely satisfy the normal distributional assumption. Therefore, a permutation test which does not require such a normality assumption is alternatively used to detect DEGs (Dudoit et al., 2002; Klebanov. et al., 2001). SAM (Tusher et al., 2001)

uses a t-type of statistics using a fudge factor to stabilize the variance and it controls the false discovery rate (FDR) (Benjamini and Hochberg, 1995). SAM is also a non-parametric analysis which does not require a normal distributional assumption.

The application of microarray technology has also led to diverse studies beyond identifying DEGs such as a study examining relation between phenotype and expression data. Various phenotypes have been used in microarray experiments. For example, the survival time was utilized as a phenotype for analyzing the recurrence of cancer in the survival analysis (Kantoff. et al., 2010; Newland. et al., 2006). Several genes associated with the survival time were identified. Microsatellite instability (MSI) was utilized as a phenotype for a microarray study of colorectal cancer. Since the CpG island methylator phenotype (CIMP) is associated with microsatellite instability (MSI) and BRAF mutation in colorectal cancer (Ogino.et al., 2009), MSI plays an important role in colorectal cancer studies. Additionally, subtype of tumor can also be an important phenotype. For example, Estrogen Receptor(ER), Progesterone Receptor (PR), and HER2 jointly define the subtypes of breast cancer. Triple-negative phenotype (ER-negative, PR-

negative, and HER2−negative) is most commonly used (Bauer. et al., 2007).

I call genes that are associated with a phenotype of interest as phenotype associated genes (PAGs). PAGs can be identified by regression analysis such as linear regression analysis for the continuous phenotypes and Cox regression model for the survival time phenotype (Wei, 1992; Lin, 1994). When the phenotype is a binary variable representing two groups, identification of PAGs becomes equivalent to identification of DEGs.

1. 3. Purpose of research

In this article, I focus on the joint identification of DEGs and PAGs in microarray data analyses. Our study was motivated from a microarray experiment consisting of high fat diet (HFD) and normal diet (ND) groups. 10 mice were assigned for each ND group and HFD group for microarray experiment. In addition, four phenotypes composing of leptin, adiponectin, insuline-like growth factor 1 (IGF-1) and insulin were extracted from each blood sample. I are interested in determining influential genes associated with obesity. Thus, I need to identify genes that are both DEGs between HFD and ND groups and PAGs for these phenotypes.

Although many approaches have been proposed for the separate identification of DEGs and PAGs, only a few approaches are available for the joint identification of DEGs and PAGs. In this article, I consider the methods for the joint identification and then propose a new method.

The first approach for the joint identification of DEGs and PAGs is a naïve approach which detects DEGs and PAGs separately, and then

identifies the intersection genes of PAGs and DEGs. The second approach is a hierarchical approach (Reiner-benaim.et.al, 2006) which detects DEGs first and then chooses PAGs among DEGs, or visa versa. Both approaches are two-stage analysis which requires separate testing of DEGs and PAGs, and as a result, are not easy to control false positive errors. Thus, I propose a new model-based approach for a joint identification of DEGs and PAGs simultaneously. The model-based approach uses a linear regression model. This method is one-stage analysis which has less computing time, is easier to control false positive errors, and has higher power than naïve and hierarchical approaches. Through a real mice microarray data analysis, I compare our model-based approach with naïve and hierarchical approaches.

2 Materials and Methods

2.1 Data

Microarray data consist of high fat diet (HFD) and normal diet (ND) groups to determine influential genes associated with obesity. ND is ingested normal fat diet which is made of 11% fat. HFD is ingested high fat diet which has 40% of fat. Therefore, I observed change in the gene expressions due to obesity and fat proportion. 10 mice were assigned for each ND group and HFD group. Then, 3 mice among ND group and 6 mice among HFD group were selected for microarray experiment. Finally, I got data which consists of 3 ND samples and 6 HFD samples, and each sample has 45281 probes.

Four phenotypes associated with regulating metabolism were extracted from the blood sample including leptin, adiponectin, insuline-like growth factor 1 (IGF-1) and insulin. Leptin is an adipocyte-secreted hormone with a key role in energy homeostasis (Brennan and Mantzoros, 2006). IGF-1 is similar in molecular structure to insulin. IGF-1 is the important hormone for childhood growth. Adiponectin take a role to control glucose levels as well as fatty acid breakdown. Insulin is one of the important hormones in the metabolism system.

2.1 DEG detection

First of all, I detect DEGs by two-sample t-test and permutation test. Since the permutation test needs no normality assumption, it has been used in many microarray fields and previous studies (Dudoit et al., 2002; Klebanov. et al., 2001ref xxx).

Secondly, Significant Analysis of Microarrays (SAM) (Tusher. et al., 2001) is performed to identify DEGs. The SAM method was proposed to identify differential expressions from microarray data. SAMThis method uses the modified t statistics by adding new a fudge factor (s_0) to common statistics (r_i/s_i) as one of the penalized methods. s_i is the estimated standard error from gene i, and s_0 is calculated as a percentile based on α. Then, the following test statistic is used:

$$d_i = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{s_i + s_0}, i = 1,2,\dots,p$$

In addition, the SAM method uses the permutation algorithm to control False Discovery Rate (FDR) (Benjamini and Hochberg, 1995). Therefore, I can control FDR more easily than other tests such aslike t-test.

2. 3. PAG detection

Linear regression analysis is utilized to determine PAGs. There are two treatment groups: ND and HFD. As mentioned earlier, the phenotypes of interest consist of leptin, adiponectin, IGF−1 and insulin. Then linear regression analysis is performed for each phenotype. Two linear regression models are applied to identify linear relationship between genes and phenotype:

$$\mathrm{M1: Phenotype} = \beta_0 + \beta_1 \cdot \mathrm{Expression_i} + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}(0,\ \sigma^2),$$

$$\mathrm{M2: Phenotype} = \beta_0 + \beta_1 \cdot \mathrm{Expression_i} + \beta_2 \cdot \mathrm{Group} +$$

$$\varepsilon_i,\ \varepsilon_i \sim \mathrm{N}(0,\ \sigma^2)\ ,$$

where $i(= 1,2, \dots, \mathrm{p})$ represents genes. Group information is denoted by Group. $\mathrm{Expression_i}$ indicates the expression value for the ith gene. The first model M1 is to identify the effect of expression on the phenotype. The second model M2 is an extension of M1 with an additional Group covariate. Since the significance of linear relationship between gene and phenotype may be affected by group effect, some genes may not have marginal effects on the phenotype but may have conditional effects given the group information. M1 is for detecting the marginal effect, while M2 is for

detecting conditional effect. For example, V1rh4 gene is a non−PAG by model M1. However, it is identified as a PAG by model M2 (Figure 1.). Model M2 is a more appropriate model than M1, when a group effect exists. However, model M1 provides PAGs that do not depend on the group effect. Thus, both M1 and M2 need to be fitted.

In model M1, the expression effect β_1 is the main interest explaining the high fat diet effect between the ND group and HFD group. In model M2, the group effect β_2 is added. The PAGs can be identified by testing the following hypotheses:

$$H_0: \beta_1 = 0 \; for \; M1$$

$$H_0: \beta_1 = 0 \; for \; M2$$

Significances of these hypotheses can be tested by calculating F−statistics for each gene.

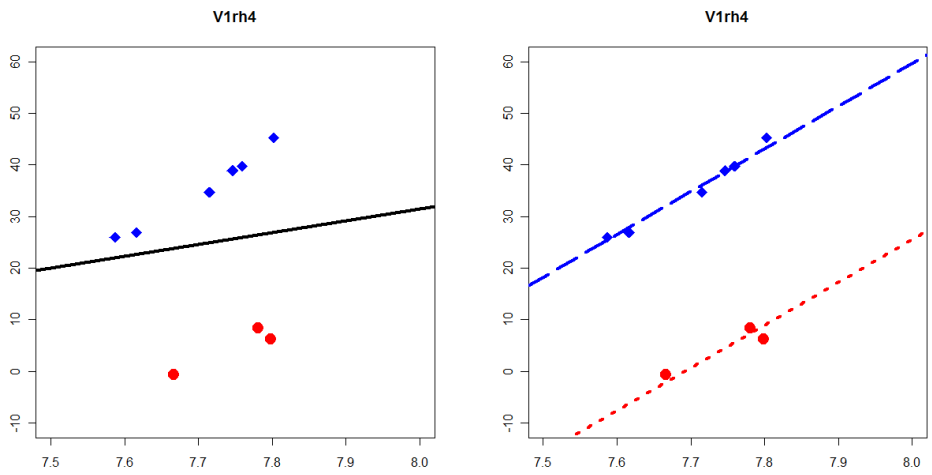

Figure1. The problem of model M1.

Model without consider a group effect cannot detect significant correlation between phenotype and gene V1rh4 (left). However, if I consider a group effect (2), I can identify novel significant interaction (right). Y-axis is phenotype value and x-axis is expression value. Blue is high fat diet (HFD) group and red is normal diet (ND) group

2. 4. Joint identification

Naïve approach

The naïve approach detects DEGs and PAGs separately, and then identifies the intersection genes of PAGs and DEGs. It consists of the following two steps:

Step1. Identifying DEGs and PAGs separately

Step2. Determining intersection gene sets of DEGs and PAGs

Hierarchical approach

The hierarchical approach (Reiner-benaim.et.al, 2006) detects DEGs first and then chooses PAGs among DEGs, or visa versa. The hierarchical approach consists of two steps: identifying DEGs and then detecting PAGs among DEG set, or alternatively, identifying PAGs and then detecting DEGs among PAG set.

Step1. Identifying DEGs

Step2. Detecting PAGs by linear regression with models M1 and M2 for the DEGs selected at Step 1

Both naïve and hierarchical approaches are two-stage analysis which requires separate testing of DEGs and PAGs, and as a result,

are not straightforward to control false positive errors. Thus, I propose a new model-based approach for a joint identification of DEGs and PAGs simultaneously. The model-based approach uses a linear regression model as follows.

Model-based approach

I propose a new model M3 to determine DEGs and PAGs simultaneously.

$$\text{M3: Expression}_i = \gamma_0 + \gamma_1 \cdot \text{Group} + \gamma_2 \cdot \text{Phenotype} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1,2, \dots, p$$

The proposed model M3 has a different structure from models M1 and M2. M3 regards Expression as a dependent variable and Group and Phenotype as covariates. On the other hand, M1 and M2 treat Phenotype as dependent variable. In M3, the group effect r_1 accounts for the differential expression between ND group and HFD group. The phenotype effect r_2 indicates linear relationship between gene i and phenotype. In model M3, I am interested in identifying DEGs and PAGs at the same time.

$$H_0: \gamma_1 = \gamma_2 = 0$$

I perform a significant test of this hypothesis by calculating F-test

for each gene. DEGs and PAGs are determined simultaneously by testing γ_1 and γ_2 in one model. Then, I can detect DEGs and PAGs simultaneously by determining significant genes from null hypotheses.

Model M3 provides significant genes as for DEGs and PAGs simultaneously. However, I cannot definitely determine genes which are identified by group or phenotype effects. Therefore, I need post-hoc analysis for phenotype and group. The t-value from linear regression analysis is utilized for testing for each effect. In post-hoc test, $\alpha/2$ is used to cut-off p-value as multiple comparison (Box et.al., 1978).

Unlike other joint identification methods, our proposed model does not need to identify DEGs and PAGs separately, but jointly identify DEGs and PAGs.

3. Result

3.1. DEGs

Two-sample t-test, permutation test and SAM method were utilized to identify DEGs. T-test didoes not provide any significant results after controlling FDR at the 5% level for the multiple comparison. Neither Bonferroni correction for the p-values provided any significant results at the 5% significance level. Next, I obtained significant results at the 5% significance level from the permutation test which used adjusted p-values to control family wise error rate (Westfall and Young, 1993). Finally, I obtain DEGs by nominal p-value from student t-test without multiple comparisons. Tthe SAM method provided significant results at the 0% median FDR=0. Although SAM provided only a few significant genes which only up regulated genes compared to other methods, the SAM result is expected to be very reliable because it didoes not contain any false positive results (FDR=0). Moreover I obtain significant result from permutation test that use adjusting p-value to control family wise error rate at α=0.05 (Westfall and Young, 1993). As a result, the top lists of genes are summarized in three statistical methods provide different result (Tables 1 and 2).

Gene Symbol	p-value	q-value
A230069A22Rik	0.000166	0.930077
2610018G03Rik	0.00019	0.930077
Pim3	0.000303	0.930077
D930042N17Rik	0.000326	0.930077
Ctns	0.000333	0.930077
Cib3	0.000374	0.930077

Table 1. Top significant genes list for t-test.

Gene Symbol	q-value
Tfrc	0
Tfrc	0
Sprr1a	0
Cyp4f16	0
9030605I04Rik	0
Tfrc	

Table 2. Top significant probes list for SAM.

3. 2. PAGs

Figure 2 shows the pairwise plot showing correlation coefficients among the four phenotypes. They are all highly positively correlated. Leptin and insulin have a high correlation coefficient (0.836). Both leptin and insulin are well known to be associated with body composition (Zoico. et al., 2008), and with BMI and type 2diabetes (Lacobellis. et al., 2005; Osuna. et al., 2006).

Models M1 and M2 were employed to detect PAGs for these phenotypes. Figure 3 shows the Vendiagram of the number of PAGs identified by M1 and M2 at the significance level 1%. Depending on the phenotypes, the numbers of overlapped and non-overlapeed PAGs differ greatly. Figure 4 shows examples of PAGs. The first one is detected only by M1, the second detected by both M1 and M2, and the third one detected only by M2.

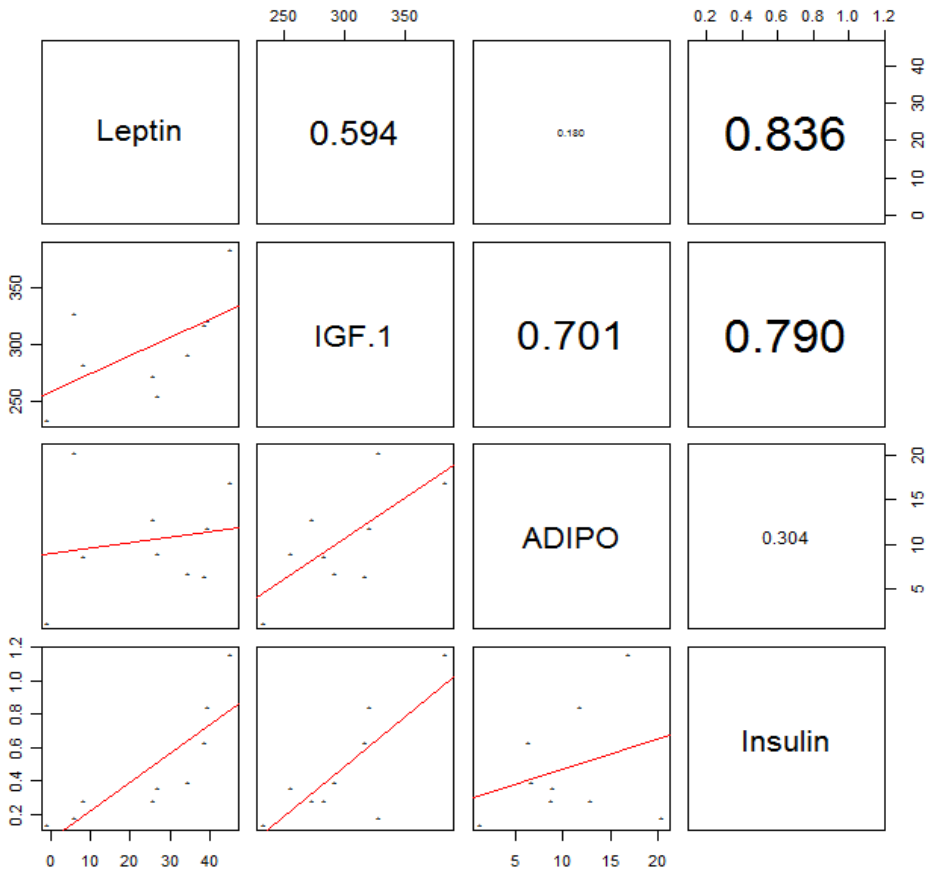

Figure 2. Correlation plot of phenotype.
Leptin and Insulin show the highest correlation value. IGF-1 and ADIPO represent low correlation values with reported to Leptin or Insulin.

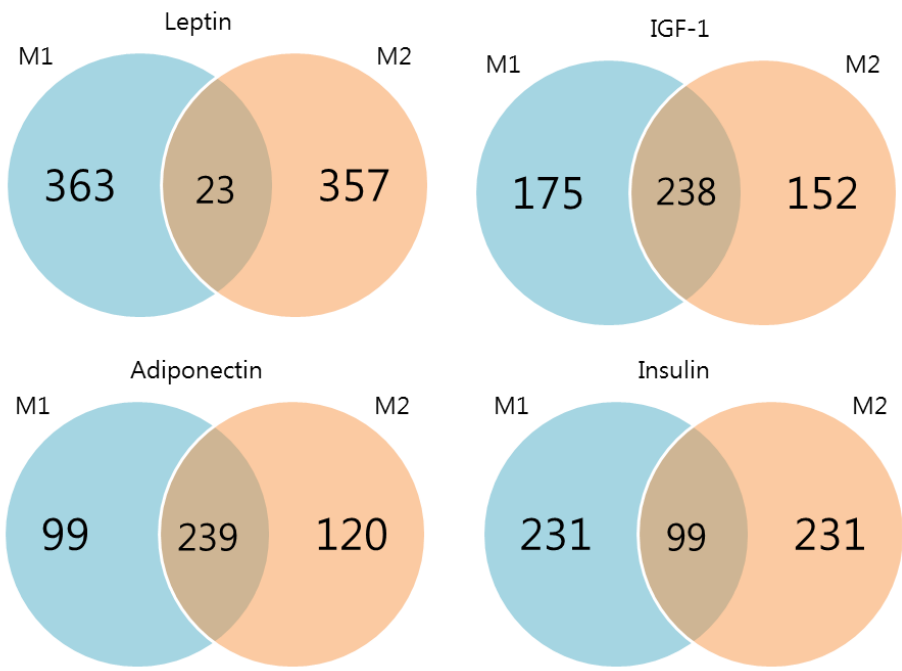

Figure3. The venn diagram of PAG.

I can found M1 reveals a lot of number of significant PAG. Therefore, model M1 should be considered statistical model to detect PAG. However, it may have high false positive rate and M2 is more appropriate model.

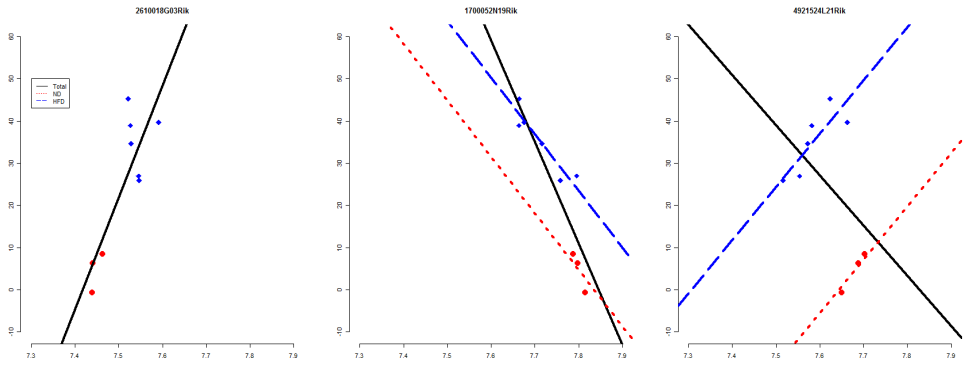

Figure 4. The PAG example plots

The y-axis is phenotype and the x-axis is the expression value. (a) detected only by model M1 (2610018G03Rik). (b) detected by models M1 and M2 at the same time (1700052N19Rik). (c) detected only by model M2 (4921524L21Rik). Blue is high fat diet (HFD) group and red is normal diet (ND) group T-test

3. 2. Joint identification

I applied three joint identification methods to the mice microarray data. I first applied the naïve approach by comparing the list of DEGs and PAGs. I then applied the hierarchical approach by detecting DEGs first and then PAGs. Finally, I applied the model-based approach using model M3. For the purpose of fair comparison, I fixed the same level of significance and FDR. Table 3 summarizes the numbers of significant genes that are both DEG and PAG. The number of significant genes that are identified by the model-based approach is greater than those obtained by other methods. However, it is difficult to determine whether or not the model-based approach produced a more false positive rate is unknown. The top lists of genes are summarized in Table 4.

	Joint identification method	the Naïve approach		the Hierarchical approach (DEG → PAGs)		model-based approach
	PAGs method	M1	M2	M1	M2	M3
Leptin	T-test	932	124	385	55	640
	SAM	0	0	4	0	
	Permutation test	92	90	28	33	
IGF.1	T-test	41	67	25	29	171
	SAM	0	0	0	0	
	Permutation test	95	87	30	42	
Adiponectin	T-test	0	13	0	1	124
	SAM	0	0	0	0	
	Permutation test	97	93	21	23	
Insulin	T-test	337	130	188	62	307
	SAM	0	0	0	0	
	Permutation test	81	88	35	42	

Table3. The number of significant genes from joint identification.

I can see model-based approach provide s shows better power than the naïve approach. Significance level is fixed for 5%

Leptin		Adiponectin	
Gene Symbol	P-value	Gene Symbol	P-value
Aps-pending	7.77E-05	LOC236170	1.43E-05
1700028I16Rik	7.89E-05	1700072E05Rik	1.68E-05
V1rh4	9.96E-05	Asb17	1.95E-05
Gnptg	0.000113278	Grb10	6.33E-05
LOC383443	0.000136373	Smyd4	9.62E-05
LOC231501	0.000172273	1700006J14Rik	0.000124993
IGF-1		Insulin	
Gene Symbol	P-value	Gene Symbol	P-value
4930564C03Rik	3.98E-05	Klhdc4	8.84E-05
Bivm	5.53E-05	9030421J09Rik	0.000103105
4933400A22Rik	0.000101201	Mphosph8	0.00015379
D630037D12Rik	0.000108573	D930042N17Rik	0.000155715
Sfxn3	0.000135788	LOC381996	0.000191544
Glmn	0.000172726	4930404F20Rik	0.000226355

Table4. Top gene list from model-based approach.

4. Simulation study

In the previous section, we applied three joint identification methods to real microarray data. Although the model-based approach using model M3 showed a better performance than other methods, the comparison based on the number of significant genes is rather limited. For a more systematic comparison, we perform an extensive simulation study. We need biological validation to detect false positive rate in the result. We perform simulation study instead of biological validation for comparing joint identification methods.

First, we generated phenotype data from normal distribution with the same mean and variance of real phenotype data. Since adiponectin most well followed the normal distribution as determined by the normality test (Shapiro and Wilks, 1965), the mean and variance of adiponectin were used to simulate phenotype data. Then, microarray data were generated from the normal distribution and were assumed to be log-transformed. We generated 10,000 genes data which consisted of 1,000 DEGs and 9,000 non-DEGs. In this simulation data, differentially group falls in both DEGs and PAGs criteria. To categorize differential expressed group and non-differential expressed group, we add effect size (i.e. fold-change) to the differentially expressed group. Then, we use

linear regression analysis with model M2 to produce PAGs group and non-PAGs group. Finally, we obtain the simulation data sets which consist of DEGs and non DEGs.

Simulation data is generated for each effect size from 1 to 5. Then we calculate true positive rate and false positive rate for each effect size. The model-based approach and the naïve approach with T-test and linear regression from models M1 and M2 are compared. Figure 5 (a) shows the simulation results, x-axis is the effect size, and y-axis is true positive rate. The model-based approach represents the best performance among the available methods. In Figure 5 (b), false positive rate is plotted. The model-based tend to have low false positive rate.

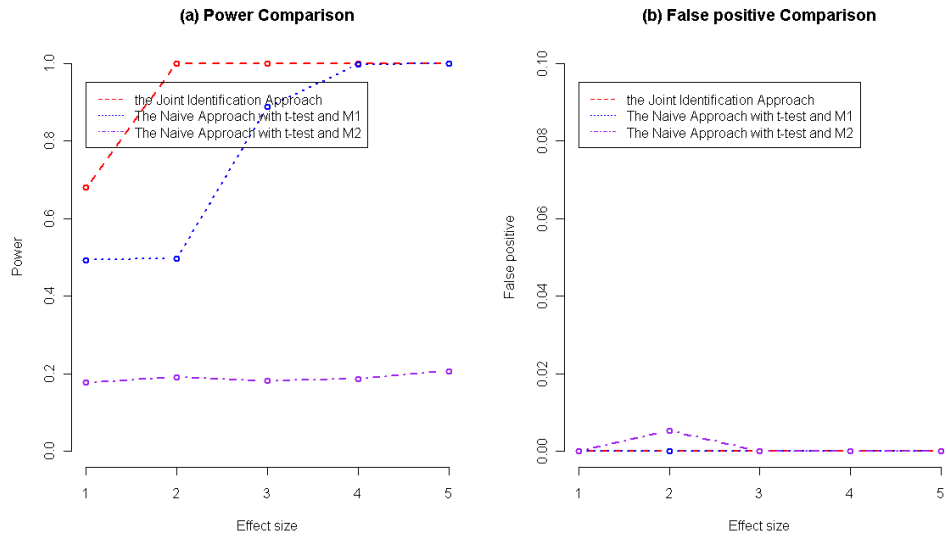

Figure5. Simulation result.

The model-based approach shows better power than the naïve approach. All joint identification approach tends to provide almost zero value of false positive rate.

5. Discussion

As the microarray experiment design becomes more complex, a more complicated analysis method needs to be developed. In the previous microarray studies, either DEGs or PAGs need to be identified. However, recent microarray design requires a more challenging method in order to detect the genes that are simultaneously DEGs and PAGs. Although various methods have been proposed for detecting DEGs and PAGs, most of them can identify DEGs or PAGs separately. Then, integrating DEGs and PAGs tend to have low statistical power.

In this paper, we propose a statistical model for detecting DEGs and PAGs simultaneously. The proposed model is more efficient than other naïve methods for the simultaneous identification of DEGs and PAGs. Through a real microarray data and simulation studies, the proposed model was compared to the other methods and was shown to have larger power. In other words, the proposed model provided more significant genes than other approaches at the same condition (Table 4.).

Additionally, the proposed approach was flexible and easy to extend.

Since our model is linear regression model, it can be extended to analysis when there are more than two factors. For example, our model can be applied for analyzing of variety clinical covariate at the same time.

Also, four phenotypes were well known that associated with regulating metabolism. Thus, a lot of significant genes may be associated with regulation function. For example, the Olfr137 gene Olfactory receptors interact with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell (Young et al., 2003; Amadou et al., 2003). The Cntnap4 gene which was identified significant gene with adiponectin product belongs to the neurexin family, members of which function in the vertebrate nervous system as cell adhesion molecules and receptors (Spiegel et al., 2002). The Grb10 gene encodes a growth factor receptor-binding protein that interacts with insulin receptors and insulin-like growth-factor receptors (Garfield. et al., 2011). In addition, we can found many genes had function of regulating systems by gene ontology analysis.

6. Reference

1. Adomas A. et al. (2008) Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. Tree Physio, 28 (6), 885–897.
2. Almut Schulze and Julian Downward (2001) Navigating gene expression using microarrays — a technology review. Nature Cell Biology, 3, 190 – 195
3. Amadou et al.(2003) Co-duplication of olfactory receptor and MHC class I genes in the mouse major histocompatibility complex. Hum. Mol. Genet. 12 (22):3025-3040.
4. Antigone S.Dimas. et al. (2009) Common Regulatory Variation Impacts Gene Expression in a Cell Type-Dependent Maner. Science 325, 1246-1250.
5. Aoife M Brennan & Christos S Mantzoros (2006). Drug Insight: the role of leptin in human physiology and pathophysiology— emerging clinical applications. Nature Reviews Endocrinology, 2, 318-327
6. Bauer KR. et al. (2007) Descriptive Analysis of Estrogen Receptor (ER)-Negative, Progesterone Receptor (PR)-Negative, and HER2-Negative Invasive Breast cancer, the So-

called Triple-Negative Phenotype. Cancer, 109, 1721-1728.

7. Benjamini,Y and Hochberg.Y. (1995) Controlling the false discovery rate: a pratical and powerful approach to multiple testing. J R Stat Soc Ser B, 57, 289-300.
8. Cheang MCU. et al. (2008) Basal-Like Breast Cancer Defined by Five Biomarkers Has Superior Prognostic Value than Triple-Negative Phenotype. Clin Cancer Res, 14, 1368
9. Christine Debouck & Peter N. Goodfellow (1999) DNA microarrays in drug discovery and development. Nature Genetics 21, 48 – 50
10. Diane Gershon (2002) Microarray technology: An array of opportunities. Nature 416, 885-891
11. Dudoit. et al. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12, 111-139
12. D.Y.Lin (1994) Cox regression analysis of multivariate failure time data: the marginal approach. Statistics in medicine, 13, 2233-2247
13. Garfield. et al.,(2011) Distinct physiological and behavioural functions for parental alleles of imprinted Grb10. Nature 469 (7331): 534-538
14. Gerhard et.al. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection

(MGC). Genome Res, 14(10B):2121-7.

15. Hacia JG. et al. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet, 22 (2), 164-167.
16. Hal. et.al. (2000) The application of DNA microarrays in gene expression analysis. Journal of Biotechnology, 78,271-280
17. Jain N. et.al. (2003) Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. BMC Bioinformatics, 19, 1945-1951
18. Kantoff. et al. (2010) Overall Survival Analysis of a Phase II Randomized Controlled Trial of a Poxviral-Based PSA-Targeted Immunotherapy in Metastatic Castration-Resistant Prostate Cancer. American Society of Clinical Oncology, 28, 1099-1105
19. Klebanov. et al. (2001) A permutation test motivated by microarray data analysis. Elsevier, 50, 3619-3628
20. Lacobellis. et al. (2005) Relationship of thyroid function with body mass index, leptin, insuline sensitivity and adiponectin in euthyroid obese women. Clinical Endocrinology, 62, 487-491.
21. L. J. Wei (1992) The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. Statistics in medicine, 11, 1871-1879
22. MAQC Consortium (2006) The MicroArray Quality Control

(MAQC) project shows inter- and intraplatform reproducibility of gene expression. Nature biotechnology, 24, 1151-1161

23. Michael J. Heller (2002) DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications. Annu. Rev. Biomed. Eng, 4, 129-153
24. Newland. et al. (2006) Pathologic determinants of survival associated with colorectal cancer with lymph node metastases. A multivariate analysis of 579 patients. Cancer, 73, 2076-2082
25. Ogino S. et al. (2009) CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. Gut, 58, 90-96
26. Osuna C. et al. (2006) Relationship between BMI, total testosterone, sex hormone-binding-globulin, LEPTIN, Insulin and Insulin resistance in obese men. Informa Healthcare, 52, 355-361
27. Park T. et al. (2006) Combining multiple microarrays in the presence of controlling variables. BMC Bioinformatics, 22, 1682-1689.
28. Patrick C A. Dubios. et al. (2010) Multiple common variants for celiac disease influencing immine gene expression. Nat Genet, 42, 295-302.
29. Reiner-Benaim.A.et.al. (2007) Associating quantitative behavioral traits with gene expression in the brain: searching

for diamonds in the hay. BMC Bioinformatics, 23, 2239–2246.

30. Schena M. et al. (1995) Quantitative Monitering of Gene Expression Patterns with a Complementary DNA Microarray. Science 270, 467–469.
31. Shapiro,S. S. and Wilk,M.B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika 52 (3–4), 591–611.
32. Spiegel et al. (2002) Caspr3 and caspr4, two novel members of the caspr family are expressed in the nervous system and interact with PDZ domains. Mol Cell Neurosci 20 (2): 283–97.
33. Statnikov A. et.al. (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. BMC Bioinformatics, 21, 631–643.
34. Steven et. al. (1998) The Mammalian γ–Tubulin Complex Contains Homologues of the Yeast Spindle Pole Body Components Spc97p and Spc98p. The Journal of Cell Biology, 141, 663–674
35. Tusher, V.G. et al. (2001) Significance analysis of microarrays applied to the ionizing radiation response. PNAS, 98(9), 5116–5121.
36. Wei Pan (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated

microarray experiments. BMC Bioinformatics, 18, 546-554.

37. Westfall, P.H. and Young, S.S. (1993). Resampling-based multiple testing: examples and methods for p-value adjustment. John Wiley & Sons, Inc., NY, USA..
38. Wirapati. et.al. (2008) Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Research, 10, R65
39. Young et. al.(2003) Odorant receptor expressed sequence tags demonstrate olfactory expression of over 400 genes, extensive alternate splicing and unequal expression levels. Genome Biol. 4(11): R71.
40. Zoico E. et al. (2008) Relation between adiponectin and bone mineral density in elderly post-menopausal women: role of body composition, leptin, insulin resistance, and dehydroepiandrosterone sulfate. Journal of Endocrinological Investigation, 31(4), 297-302

Abstract (Korean)

특정형질과 집단간
유의한 유전자 공동 판별

다양한 기술의 등장에 따라서 다양한 종류의 생물학적 데이터가 만들어졌다. 그 중에서도 마이크로어레이 기술을 이용한 데이터는 생물학분야에 중요한 전환점이 되었고, 마이크로어레이 기술이 등장한지 10년이 넘어가지만 아직도 생물학분야에서는 활발하게 사용되고 있는 기술이다. 10년동안 이런 마이크로어레이 데이터를 분석하기 위한 다양한 분석방법과 도구가 개발되어왔다.

일반적으로 DEGs는 마이크로어레이 데이터를 이용한 연구에서 가장 주된 연구 주제로 다루어졌다. 게다가 특정형질과 유전자의 발현간의 간계를 연구하는 연관분석이 성행하면서 다양한 형질들이 연구에서 사용되었다. 이런 연관분석을 통해서 우리는 특정형질과 관련되어 있는 유전다, 즉 PAGs(phenotype associated genes)를 찾을 수 있다.

우리 연구는 마이크로어레이 데이터와 특정형질의 데이터를 이용하여 DEGs와 PAGs를 동시에 찾는 작업을 수행하였다. 첫번째로 DEGs와 PAGs를 독립적으로 전체 데이터에서 분별한 후에 이들의 교집합을 찾는 the niave approach 방법을 사용하였다. 두 번째로는 DEGs를 먼저 찾고 그 다음 DEGs 중에서 PAGs를 고르는 hierarchical approach 를 사용하였다.

마지막으로 우리는 통계적인 모델을 기반으로 DEGs와 PAGs를 동시에

찾는model-based approach 방법을 제안하였다. 실제 마이크로어레이 데이터와 simulation study를 통해서 우리가 제안한 model-based approach 방법과 다른 방법들을 비교 분석 하였다.

Keyword: Differential expression genes (DEGs), Phenotype associated genes (PAGs), Joint identification, association study, Linear regression model.
Student number: 2010-23163