DETECTING LOCUS-LOCUS INTERACTIONS USING MICROARRAY DATA

DETECTING LOCUS-LOCUS INTERACTIONS USING MICROARRAY DATA

By

YanFei Gao, BSc

A Project

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2006) McMaster University

(Statistics) Hamilton, Ontario

TITLE: Detecting Locus-Locus Interaction Using Microarray Data

AUTHOR: YanFei Gao, B.Sc

(McMaster University, Canada)

SUPERVISOR: Professor Angelo Canty

NUMBER OF PAGES: viii, 60

Abstract

In this report we explore how to find the locus-locus interaction using microarray data. Our analysis makes use of a dataset from an experiment with Affymetrix GeneChip MGU74Av2 for mice. In Chapter 1 we give the genetics background, an introduction to microarray methodology and the preprocessing of microarray data, and a review of SAM (Significance Analysis of Microarrays) method for finding differentially expressed genes in microarray data. In Chapter 2 we describe our dataset and our objective of finding the genes with locus-locus interaction but with no main effect. We also show how to find the interaction in this chapter. In Chapter 3 we show the simulation study of detecting the locus-locus interaction without main effects and propose a two-stage method of doing that. In Chapter 4 we apply the two-stage method to the microarray data and focus on the second stage analysis. In Chapter 5 we examine an alternative method using bootstrap resampling in place of permutations. Chapter 6 contains our conclusion and some suggestions for future research.

Acknowledgements

I would like to thank my supervisor, Dr. Angelo Canty, for his great guidance, support and patience, and for giving me an opportunity to work on this project. I really appreciate Dr. Canty's supervision on my coding in R.

I would like to thank Dr. Jayne Danska and Dr. Tanya Prasolava to supply the data used in this project.

I would like to thank Dr. Peter Macdonald and Dr. Roman Viveros-Aguilera for being the examiners of this project. I appreciate all the comments and criticism. I also would like to thank them for their help and support during my two-year study.

I would like to thank my parents and friends for their constant support and encouragement throughout my academic career.

Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Genetics Background

A *cell* is the minimal unit of life. The life process involves a wide array of molecules ranging from water to small organic compounds, and macromolecules (DNA, proteins, and polysaccharide) that define the structure of the cells. Macromolecules control and govern most of the activities of life. *Deoxyribonucleic acid* (DNA) molecules store information about the structure of macromolecules, allowing them to be made precisely according to cells' specification and needs (Lee, 2004).

The chemical components of the deoxyribonucleic acid (DNA) molecule dictate the inherent properties of a species. A DNA is a double-stranded helix of nucleotides which carries the genetic information of a cell. It encodes the information for the proteins and is able to self-replicate. A nucleotide consists of three components: a five-carbon sugar called deoxyribose, one or more phosphate group(s), and one of four nitrogen bases. The four nitrogen bases include purines consisting of adenine (A) and guanine

(G), and pyrimidines consisting of cytosine (C) and thymine (T). Those two strands constructing a DNA are joined together by binding of the complementary bases, and can be separated by heating. The process of *hybridization* is the fundamental basis of DNA microarrays. Two DNA strands hybridize if they are complementary to each other. Complementarity follows the Watson-Crick rule that adenine (A) binds to thymine (T) and cytosine (C) binds to guanine (G). When the DNA is copied by the processes of replication and transcription, the double helical structure of the DNA is opened up and a copy is made on the specificity of the base pairing.

Genes are the units of the DNA sequence that control the identifiable hereditary traits of an organism. A *gene* can be defined as a segment of DNA that specifies a functional RNA. The total set of genes carried by an individual or a cell is called its *genome*. The genome defines the genetic construction of an organism or cell, or the *genotype*. The phenotype, on the other hand, is the total set of characteristics displayed by an organism under a particular set of environmental factors. The outward appearance of an organism (phenotype) may or may not directly reflect the genes that are present (genotype). With microarray technology we can study the expression of all the genes in an organism simultaneously (Lee, 2004).

The core biochemical flow of genetic information can be summarized as the process of RNA synthesis (transcription) and the process of protein synthesis (translation). One or both strands of the DNA hybrid can be replaced by RNA and hybridization will still occur as long as there is complementarity (Knudsen, 2004). The first step in making a protein is to copy, or transcribe, the information encoded in the DNA of the genes into a single-stranded molecule called *ribonucleic acid* (RNA). The process is similar to the process of copying written words, the synthesis of RNA from DNA

is called *transcription.* The DNA is said to be transcribed into RNA, and the RNA is called a *transcript.* There are two general classes of RNAs. Those that take a part in the process of decoding genes into proteins are referred to as *informational* RNAs called *messenger* RNA (mRNA). In the other class, the RNA itself is the final functional product. These RNAs are referred to as *functional* RNAs. Functional RNAs are the *transfer* RNAs (tRNA) and the *ribosomal* RNA (rRNA), which are both parts of the intricate protein synthesis machinery that translates the informational mRNA into protein (Lee, 2004).

The primary structure of a protein is a linear chain of building blocks called *amino acids.* There are 20 amino acids that commonly occur in proteins. Because the process of reading the mRNA sequence and converting it into an amino acid sequence is like converting one language into another, the process of protein synthesis is called *translation* (Lee, 2004).

Gene expression is the amount of mRNA that is produced for a gene in the process, by which mRNA (and eventually protein) is synthesized from the DNA template of each gene. The first stage of this process is *transcription*, when an RNA copy of one strand of the DNA is produced. And the next stage of the process is the *translation* of the mRNA into protein.

1.2 Microarray Methodology

Microarray technology allows measurement of the levels of thousands of different RNA molecules at a given point in the life of an organism, tissue, or cell. Comparisons of the levels of RNA molecules can be used to decipher the thousands of processes going

on simultaneously in living organisms. Also, comparing healthy and diseased cells can yield vital information on the causes of diseases. Microarrays have been successfully applied to several biological problems and, as arrays become more easily available to researchers, the popularity of these kinds of experiments will increase (Lee, 2004).

Although all of the cells in the human body contain identical genetic material, the same genes are not active in every cell. Studying which genes are active and which are inactive in different cell types helps scientists to understand both how these cells function normally and how they are affected when various genes do not perform properly. In the past, scientists have only been able to conduct these genetic analyses on a few genes at once. With the development of DNA microarray technology, however, scientists can now examine how active thousands of genes are at any given time.

DNA microarrays are created by robotic machines that arrange minuscule amounts of hundreds or thousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellular machinery begins to copy certain segments of that gene. The resulting product is known as messenger RNA (mRNA). The mRNA produced by the cell is complementary, and therefore will bind to the original portion of the DNA strand from which it was copied.

To determine which genes are turned on and which are turned off in a given cell, a researcher must first collect the messenger RNA molecules present in that cell. The researcher then labels each mRNA molecule by attaching a fluorescent dye. Next, the researcher places the labeled mRNA onto a DNA microarray slide. The messenger RNA that was present in the cell will then hybridize - or bind - to its complementary DNA on the microarray, leaving its fluorescent tag. A researcher must then use a

special scanner to measure the fluorescent areas on the microarray.

If a particular gene is very active, it produces many molecules of messenger RNA, which hybridize to the DNA on the microarray and generate a very bright fluorescent area. Genes that are somewhat active produce fewer mRNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of the messenger molecules have hybridized to the DNA, indicating that the gene is inactive. Researchers frequently use this technique to examine the activity of various genes at different times.

For the gene expression analysis the field has been dominated by two major technologies. The one we used is the Affymetrix, Inc. GeneChip system prefabricated oligonucleotide chips. Affymetrix uses equipment similar to that which is used for making silicon chips for computers, and thus allows mass production of very large chips at reasonable cost. While computer chips are made by creating masks that control a photolithographic process for removal or deposition of silicon material on the chip surface, Affymetrix uses masks to control synthesis of oligonucleotide on the surface of a chip. The masks control the synthesis of several hundred thousand squares, each containing many copies of an oligonucleotide. For expression analysis, up to 40 oligonucleotides are used for the detection of each gene. Affymetrix has chosen a region of each gene that (presumably) has the least similarity to other genes. These oligonucleotides are referred to as probes. From this region 11 to 20 probes are chosen as perfect matches (PM) (i.e., perfectly complementary to the mRNA of that gene). In addition, they have generated 11 to 20 mismatch probes (MM), which are identical to the PM probes except for the central position 13, where one nucleotide has been changed to its complementary nucleotide. Affymetrix claims that the MM probes will be able to detect nonspecific and background hybridization, which is important for

quantifying weekly expressed mRNAs. The hybridization of each probes to its target depends on its sequence. All 11 to 20 PM probes for each gene have a different sequence, so the hybridization will not be uniform. That is of limited consequences as long as we wish to detect only *changes* in mRNA concentration between experiments (Knudsen, 2004).

The RNA samples are prepared, labeled, and gybridized with arrays. Arrays are scanned and images are produced and analyzed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide. The probe set intensities, called probe-level data, should be summarized to form one expression measure for each gene.

1.3 RMA Normalization of Microarray Data

In many of the applications of high-density oligonucleotide arrays, the goal is to learn how RNA population differs in expression in response to genetic and environmental differences (Irizarry *et al.*, 2003b). Observed expression levels also include variation introduced during the sample preparation, during manufacture of the arrays, and during the processing of the arrays (labeling, hybridization, and scanning) (Irizarry *et al.*, 2003b). Therefore, normalization at the probe-level is necessary.

Usually, statisticians prefer to take a look at the raw data of microarray data, because they want to understand the processing of the raw data and how it might influence the result of future analysis based on the raw data. The raw data from an Affymetrix microarray chips, called Affy chips, is in a .CEL file, and the useful information about the layout of the Affy chips is stored in a .CDF file. One approach

of normalization of Affy chips is called Robust Multi-array Average (RMA).

The RMA method for computing an expression measure is firstly to compute background-corrected perfect match intensities for each perfect match cell on every GeneChip. There is a description in Irizarry *et al.* (2003a). After background correction, the base-2 logarithm of each background-corrected perfect match intensity is obtained. These background-corrected and log-transformed perfect match intensities are normalized using the quantile normalization method developed by Bolstad *et al.* (2003). In the quantile normalization method, the highest background-corrected and log-transformed perfect-match intensity on each GeneChip is determined. These values are averaged, and the individual values are replaced by the average. This process is repeated with what were originally the second highest background-corrected and log-transformed perfect-match intensities on each GeneChip, the third highest, etc.

Following quantile normalization, an additive linear model is fit to the normalized data to obtain an expression measure for each probe on each GeneChip. The linear model for a particular probe set can be written as

$$Y_{ij} = m_i + a_j + e_{ij} \tag{1.1}$$

where Y_{ij} denotes the normalized probe value corresponding to the ith GeneChip and the jth probe within the probe set, m_i denotes the log-scale expression for the probe set in the sample hybridized to the ith GeneChip, a_j denotes the probe affinity effect for the jth probe within the probe set, and e_{ij} denotes a random error term. The estimated GeneChip-specific log-scale expression values would be reported as the RMA measures of expression for dataset.

1.4 Significance Analysis of Microarrays

Significance Analysis of Microarrays (SAM) methodology references in the context of a general approach to detecting differential gene expression in DNA microarrays. Some recently developed methodology for estimating false discovery rates and q-values has been included in the SAM.

SAM aims to identify differentially expressed genes from a set of microarray experiments, which falls under the heading of *multiple hypothesis testing* statistically. In other words, we must perform hypothesis tests on all genes simultaneously to determine whether each one is differentially expressed or not. In statistical hypothesis testing, we test a null hypothesis versus an alternative hypothesis. The null hypothesis is that there is no change in expression level between experimental conditions. The alternative hypothesis is that there is some change. We reject the null hypothesis if there is enough evidence in favor of the alternative. This amounts to rejecting the null hypothesis if its corresponding statistic falls into some predetermined rejection region. Hypothesis testing is also concerned with measuring the probability of rejecting the null hypothesis when it is really true (called a false positive) and the probability of rejecting the null hypothesis when the alternative hypothesis is really true (called power).

There are four important steps one must take in testing for differential gene expression, which are given in Irizarry *et al.* (2003b).

1. A statistic must be formed for each gene. The choice of this statistic is important in that one wants to make sure that no relevant information is lost with respect to the test of interest, yet all measurements on the gene are condensed into one

number.

2. Calculate the null distributions for the statistics.

3. Choose the rejection regions. One can take a priori symmetric or one-sided rejection regions, or one can choose them adaptively. This involves comparing the original statistics to null versions of the statistics. Asymmetric rejection regions are most appropriate because we do not know beforehand what proportion of differentially expressed genes are in the positive or negative direction.

4. Assess or control the number of false positives at the traditional 5% level, then the false positives would be large under null hypothesis if we were testing large number of genes. This is not acceptable, so some procedure must be performed to control the false positive rate in a reasonable manner.

1.5 Forming Test Statistics and Determining s_0

A reasonable test statistic for assessing differential gene expression is the standard (unpaired) t-statistic:

$$t_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j} \tag{1.2}$$

where $\bar{x}_{j1}$ and $\bar{x}_{j2}$ is the average gene expression for gene j under conditions 1 and 2, and s_j is the pooled standard error for gene j defined as

$$s_j = \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{\sum (x_{ji} - \bar{x}_{j1})^2 + \sum (x_{ji} - \bar{x}_{j2})^2}{n_1 + n_2 - 2}} \tag{1.3}$$

Here, n_k is the number of arrays in condition k, and each summation is taken over its respective group (Storey and Tibshirani, 2003).

Each of these statistics is formed using only information from the gene itself. It is possible to model the data in such a way that one can *borrow strength* across genes. For example, if we view the s_j as coming from some overall random process, then they can be jointly modeled. This can lead to reduce the overall variance of the s_j, give the tests more power on average. Tusher et al. (2001) take a nonparametric approach to this and shrink the s_j toward an adaptively chosen s_0. The modified t-statistic is then

$$d_j = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s_0} \tag{1.4}$$

Specifically, s_0 is chosen as the percentile of the s_j values that makes the coefficient of variation of d_j approximately constant as a function of s_j. This has the added effect of dampening large values of d_j that arise from genes whose expression variability is near zero.

The procedure for computing s_0 is given in Chu *et al.* (2005) on page 20:

1. Let s^α be the α percentile of the s_i values. Let

$$d_i^\alpha = \frac{\bar{x}_{j2} - \bar{x}_{j1}}{s_j + s^\alpha} \tag{1.5}$$

2. Compute the 100 quantiles of the s_i values, denoted by $q_1 < q_2 ... < q_{100}$.

3. For $\alpha \in (0, .05, .10...1.0)$

   a. Compute $v_j = \text{mad}(d_j^\alpha | s_i \in [q_j, q_{j+1}))$, $j = 1, 2, ...n$, where mad is the median absolute deviation from the median, divided by .64.

   b. Compute $cv(\alpha)$ = coefficient of variation of the v_j values.

4. Choose $\hat{\alpha} = \text{argmin}[cv(\alpha)]$. Finally, compute $\hat{s}_0 = s^{\hat{\alpha}}$. s_0 is henceforth fixed at the value $\hat{s}_0$.

1.6 Permutations

To estimate the p-value for a test of significance, estimate the sampling distribution of the test statistic when the null hypothesis is true by resampling in a manner that is consistent with the null hypothesis. Actually we just have one real sample, then we resample it R times without replacement. It is called permutation resampling.

For our microarray data, permutations can make no assumption about the distribution of the statistics. To calculate the d statistics d_j for each gene, $j = 1, ..., G$, the estimates and the standard errors of the estimates of factor β are needed. Therefore, we just need to permute the vector of factor labels when applying the usual ANOVA and then recalculate the statistics for new labeling. By SAM method, the plot of the average order statistics from the permutations $(\bar{d}_{(1)}, ..., \bar{d}_{(G)})$ against the observed $(d_{(1)}, ..., d_{(G)})$ is important to find differentially expressed genes.

1.7 False Discovery Rate (FDR) and q-values

When multiple hypotheses are being tested simultaneously, we need to consider some suitable measures of the error. The focus will be on the rate of false positives with respect to the number of rejected hypotheses N_r.

	Accepted Null	Rejected Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

Table 1.1: *The table for defining FDR*

Table 1.1 (McLachlan *et al.*, 2004 on page 140) describes the various outcomes when applying some significance test to perform N hypothesis tests. The specific N hypotheses are assumed to be known in advance, but the number N_0 and N_1 of true and false null hypotheses are unknown parameters. The number of rejected null hypotheses N_r is observable, while the number of false positives N_{01}, the number of false negative N_{10}, the number of true negatives N_{00}, and the number of true positives N_{11} are unobservable random variables (McLachlan *et al.*, 2004 on page 141).

The false discovery rate (FDR) was introduced by Benjamini and Hochberg (1995) as a new multiple hypothesis testing error measure, which is defined as

$$FDR = E(\frac{N_{01}}{N_r \vee 1}) \tag{1.6}$$

where $N_r \vee 1 = \max(N_r, 1)$ (McLachlan *et al.*, 2004 on page 141).

For our microarray dataset, the plot of the average order statistics from the permutations $(\bar{d}_{(1)}, ..., \bar{d}_{(G)})$ against the observed $(d_{(1)}, ..., d_{(G)})$ is shown in Figure 1.1.

Here we select a value of Δ, and draw two lines with slope 1 and intercepts $-\Delta$ and Δ. Then we can find the points $t_1(\Delta)$ and $t_2(\Delta)$ where the plot first crosses these lines. The observations further from the center than these are declared significant. By the values of $t_1(\Delta)$ and $t_2(\Delta)$, we can find the numbers of so-called significant genes in each permutation, and the average of these numbers is the average number of falsely detected differences for the give value of Δ. Then the False Discovery Rate $FDR(\Delta)$ for the given value of Δ is

$$FDR(\Delta) = \frac{\text{the average number of falsely detected differences}}{\text{number detected in the original sample}} \tag{1.7}$$

Usually the FDR is multiplied by $\hat{\pi}_0$, an estimate of the proportion π_0. π_0 is the proportion of true null (unaffected) genes in the dataset, and the algorithm for calculating

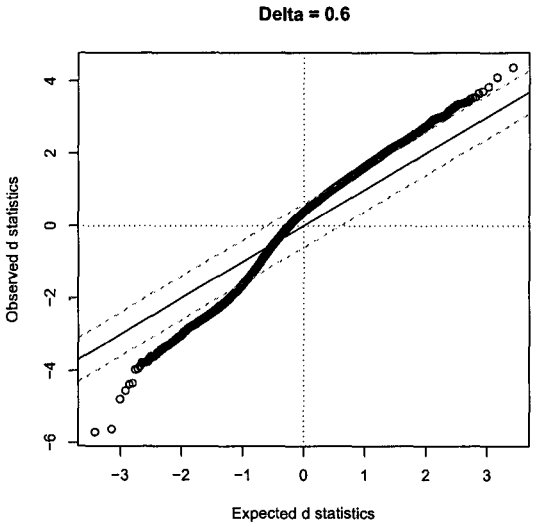

Figure 1.1: *The plot of the expected d statistics against the observed d statistics: given* $\Delta = 0.6$

$\hat{\pi}_0$, given in Storey and Tibshirani (2003)., is described in Section 1.7.

The SAM gives an algorithm to find FDR:

1. Plot the average order d statistics from the permutations (called expected d statistics) against the observed d statistics. Format the pair data $(d_{(1)}, \bar{d}_{(1)}), ..., (d_{(G)}, \bar{d}_{(G)})$ and divide this dataset into two parts from the median of the expected d statistic, called M. The part of $(d_{(i)}, \bar{d}_{(i)})$ with $\bar{d}_{(i)} \geq M$ is called the upper part, while the other part is called lower part.

2. For a fixed threshold Δ, starting at the dividing point, and moving up the right find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_1 are called *significant positive.* Similarly, start at dividing point, move down to the left and find the first $i = i_2$ such that $\bar{d}_{(i)} - d_{(i)} > \Delta$. All genes past i_2 are called *significant negative.*

For each Δ define the upper cut-point $cut_{up}(\Delta)$ as the smallest d_i among the significant positive genes, and similarly define the lower cut-point $cut_{low}(\Delta)$.

3. For a grid of Δ compute the total number of significant genes (from the previous step), and the mean of falsely called genes, by computing the mean of values among each of the B sets of $d^{*b}_{(i)}$, $i = 1, 2, ..., G$, that fall above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$.

4. Estimate π_0, the proportion of true null (unaffected) genes in the dataset, as follows:

   (a) Compute q25, q75 = 25% and 75% points of the permuted d values (if G = # genes, B = # permutations, there are GB such d values).

   (b) Compute $\hat{\pi}_0 = \#\{d_i \in (q25, q75)\}/(.5p)$ (the d_i are the values for the original dataset: there are p such values.)

5. The False Discovery Rate (FDR) is computed as the mean of the number of genes detected to be significant from permutation divided by the number of genes detected to be significant from observation.

The so-called q-value is the FDR analogue of the p-value (Storey, 2003). It gives the scientist a hypothesis testing error measure for each observed statistic with respect to the pFDR (Storey, 2003). For each gene g we can find the largest value of Δ, called Δ_g, for which that gene is significant. The q-value is then defined to be the

$$q\text{-value}(g) = \min_{\Delta \leq \Delta_g} FDR(\Delta) \tag{1.8}$$

An empirical Bayes interpretation of the q-value is the probability that the gene is actually significant given that it is called significant.

Chapter 2

Detecting Locus-Locus Interaction using Microarray Data

2.1 Objective

Microarray experiments are carried out to compare the relative abundance of specific RNA species in two or more biological samples. There may be many samples involved in an experiment, and they may have been derived from sources with their own experimental design structure (Wu *et at.*, 2003). Our experiment aims to do the research for Type 1 Diabetes, which is usually diagnosed in children and young adults, and was previously known as juvenile diabetes. To identify disease mechanisms and etiology, their genetic dissection may be assisted by evaluation of linkage in mouse models of human disease (Cordell *et al.*, 2001).

There are a number of genetic regions which are believed to play a part in Type 1

Diabetes susceptibility in mice. Two of these are called Idd5 and Idd13. Here we have two parental strains of mice Non-Obese Resistant (NOR) and Non-Obese Diabetic (NOD). These are identical by descent in 88% of the genome but NOD mice get Type 1 Diabetes at much higher rates than NOR mice (82-85% compared to 3-5% by age 6 months). We can construct *Congenic Strains* by selective multi-generational inbreeding of these mice. The NOR.NOD-Idd5 strain is identical to the parental NOR except in region Idd5 which it inherits from the NOD mice. Similarly, the NOR.NOD-Idd13 strain is identical to the parental NOR except in region Idd13 which it inherits from the NOD mice. The *Double Congenic* NOR.NOD-Idd5/13 strain is identical to the parental NOR except in regions Idd5 and Idd13 which it inherits from the NOD mice.

Our objective is to detect the locus-locus interaction using microarray dataset, and then to detect the genes not effected by Idd5 or Idd13 alone from the genes with interaction of two loci. In other words, our objective is to detect a special type of locus-locus interaction. We will describe that statistically in later sections.

2.2 Description of Dataset

In order to develop a new methodology to detecting the special case of locus-locus interaction, we start with a real dataset coming from an experiment where 4 strains of mice are used. Two of genetic regions suspected to play a part in susceptibility are Idd5 and Idd13. Then the 4 strains consist of the NOR, Single Congenic NOR-NOD.Idd5, Single Congenic NOR-NOD.Idd13, and Double Congenic NOR-NOD.Idd5/13. The day that an experiment is completed introduces non-biological variation into the process, called

Day (Blocking) Effect. There are 3 replicates for each strain on day 1 and 2 replicates for each strain on day 2 . The Affymetrix GeneChip MGU74Av2 arrays with 12,488 probe sets are used.

Our dataset are given in 20 CEL file and 1 CDF file. The .CEL file includes the mean and standard deviation (SD) stored in two matrices. The x,y entry in these matrices contains the probe intensity and SD in position x,y on the array (Irizarry, *et al.*, 2003b). The information relating genes and probe numbers to location on the chip are stored in a file with extension CDF. This implies that a CDF file is necessary to decode each CEL file (Irizarry, *et al.*, 2003b). Each chip type has a unique CDF file, while each hybridization has its unique CEL file.

2.3 Initial Analysis: Detecting Locus-Locus Interaction

The first consideration is to detect the locus-locus interaction. To look for interactions between these two loci, we use two-way ANOVA model

$$Y_{g,i} = \mu_g + \alpha_g D_i + \beta_{1,g} Idd5_i + \beta_{2,g} Idd13_i + \beta_{3,g} Idd5_i \times Idd13_i + \varepsilon_{g,i} \qquad (2.1)$$

where D_i is the day indicator, $Idd5_i = 1$ whenever locus Idd5 is NOD derived and 0 otherwise. Similarly, $Idd13_i = 1$ whenever Idd13 is NOD derived and 0 otherwise. And $\varepsilon_{g,i}$ is a random error term.

When the ANOVA model is fitted to data, we obtain estimates for each of the individual terms. Our interest is looking for interactions between that two loci. We can use the modified Wald Statistic, simply called d statistics in the report, for the

interaction term

$$d_g = \frac{\hat{\beta}_{3,g}}{se(\hat{\beta}_{3,g}) + s_0} \tag{2.2}$$

where $\hat{\beta}_{3,g}$ is the estimate of β_3, $se(\hat{\beta}_{3,g})$ is the standard error of β_3 and s_0 is the fudge factor we described in Chapter 1. Then we compute the ordered statistics $d_{(1)} \leq d_{(2)} ... \leq d_{(G)}$. We permute the 4 strain labels within days to recalculate the d statistics. There are almost 1 billion possible permutation and we take 500 at random. From the set of $B = 500$ permutations, estimate the expected order statistics by $\bar{d}_{(i)} = (1/B) \sum_1^B d_{(i)}^{*b}$ for $i = 1, 2, ..., G$ and $B = 500$, which is the mean of the order statistics from the permutations.

We plot the average of order d statistics from permutations against the observed order d statistic using our microarray data, which is the similar plot to Figure 1.1. It gives us some visual impression that how many genes can be selected as significant genes. Then we calculate the FDR and q-values for each gene using the method introduced in Chapter 1.

We can use SAM method to calculate FDR and then calculate the q-value of the genes. The formula of calculating FDR and q-value are both introduced in Chapter 1. If we state that, the gene shows evidence of an interaction if the q-value is smaller than 0.05. Then there are 14 genes that detected to have significant interaction terms. The information of 14 genes with interaction is given by ascending order of q-value and descending order of d statistics in Table 2.1.

Figure 2.1 shows the interaction plot of the gene with the smallest q-value. Although we have 14 genes detected to have significant interaction between two loci, we just illustrate the one with smallest q-value. Because the smaller q-value shows

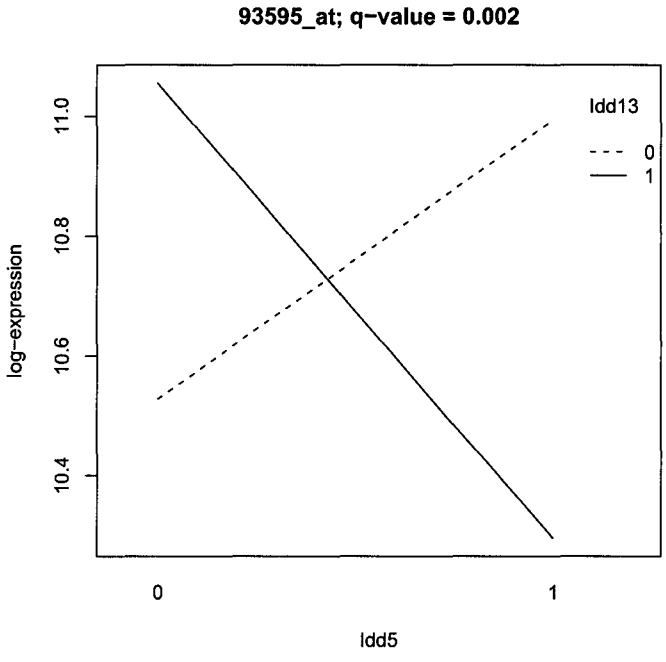

Figure 2.1: *The interaction plot of gene 93595_at*

stronger evidence that the gene has interaction. In Figure 2.2, the character 0 and 1 on the x-axis represent the locus Idd5 is NOR derived and NOD derived, respectively. The point and solid lines represent the locus Idd13 is NOR derived and NOD derived, respectively. We can observe that, the mean log-expressions of strains NORsingle Congenic NOR.NOD-Idd5, single Congenic NOR.NOD-Idd13 and double Congenic NOR.NOD-Idd5/13 are about 10.5, 11, 11, and 10.2, respectively. That indicates that when locus Idd5 is NOD derived only, compared with the mean log-expression of NOR, the mean log-expression of NOR.NOD-Idd5 increases. Similarly, when locus Idd13 is NOD derived only, compared with the mean log-expression of NOR, the mean log-expression of NOR.NOD-Idd13 also increases. However, when both of the loci

Idd5 and Idd13 are NOD derived, compared with the mean log-expression of NOR, the mean log-expression of NOR.NOD-Idd5/13 decreases. That means the interaction effect between two loci occurs. The X shape of the plot indicates the locus-locus interaction.

Index	Gene ID	NOR mean	Idd5 mean	Idd13 mean	Idd5/13 mean	q-value	absolute d stat
1	93595_at	10.529	10.993	11.0555	10.295	0.002	5.706
2	92942_at	8.212	8.332	8.363	8.0158	0.002	5.625
3	160118_at	11.0812	11.483	11.452	11.229	0.01	4.793
4	93881_i_at	5.738	6.0468	6.0507	5.917	0.0177	4.557
5	93097_at	12.362	12.608	12.528	11.860	0.0177	4.379
6	160714_at	7.375	7.823	7.846	7.714	0.0177	4.347
7	93498_s_at	9.069	9.529	9.512	9.229	0.0453	3.970
8	102983_at	7.244	7.547	7.566	7.466	0.045	3.943
9	160811_at	7.704	8.0384	8.025	7.719	0.048	3.878
10	97497_at	10.170	10.420	10.418	10.275	0.0479	3.776
11	100320_at	5.575	5.945	5.825	5.632	0.0479	3.773
12	99338_at	7.107	7.328	7.321	7.224	0.048	3.768
13	160617_at	9.027	9.288	9.292	9.175	0.048	3.755
14	96680_at	6.353	6.867	6.877	6.501	0.048	3.750

Table 2.1: *The 14 genes with significant interaction terms*

Chapter 3

Simulation

Our goal is to find the genes with interaction of Idd5 and Idd13 but not affected by Idd5 or Idd13 alone. In other words, we look for when the Double Congenic NOR.NOD-Idd5/13 is different from the other three including NOR, NOR.NOD-Idd5 and NOR.NOD-Idd13. We need to find a statistic which works well in usual ANOVA case of finding the different Double Congenic NOR.NOD-Idd5/13. Before applying the statistics to the microarray data, we try some simulated data first.

We divided our purpose into two steps: the first step is to find the interaction as described in Chapter 2 and the second step is to identify the genes not effected by Idd5 or Idd13 alone. We look for a statistic which detects the interaction of loci, and a statistic that detects when there is no main effect from either of the loci. In other words, we are interested in when Double Congenic NOR.NOD-Idd5/13 is different from the other three including NOR, NOR.NOD-Idd5 and NOR.NOD-Idd13. That means the main effects for each locus are zero but the interaction effect is non-zero. Therefore, we divided this detection into two stages. The first stage is to detect the

simulated data with non-zero interaction from all the simulated data, called stage 1. If a data is detected to have significant interaction term at a specific significance level, then we declare that it passes the first stage. Then the second stage is to detect the data with zero main effect from the data which passed stage 1, called stage 2. If a data is detected not to have any main effect at a specific significance level, then we declare that it passes the second stage.

There are 6 groups of simulation data with different values of β_1, β_2 and β_3 of model

$$G_{ij} = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \beta_3 Idd5_i \times Idd13_i + \varepsilon_{ij} \tag{3.1}$$

where G_{ij} is the simulated data of ith strain and jth replicate, ε_i is normally distributed with mean 0 and variance σ^2. And μ and σ^2 should be properly determined.

We set up 6 groups of simulated data with different values of β_1, β_2 and β_3, and generate the 6 groups of data from normal distribution.

Group	Values of			Model for simulation
	β_1	β_2	β_3	
1	$\neq 0$	$\neq 0$	$\neq 0$	$G_{ij} = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \beta_3 Idd5_i \times Idd13_i + \varepsilon_{ij}$
2	$= 0$	$= 0$	$= 0$	$G_{ij} = \mu + \varepsilon_{ij}$
3	$\neq 0$	$= 0$	$\neq 0$	$G_{ij} = \mu + \beta_1 Idd5_i + \beta_3 Idd5_i \times Idd13_i + \varepsilon_{ij}$
4	$\neq 0$	$= 0$	$= 0$	$G_{ij} = \mu + \beta_1 Idd5_i + \varepsilon_{ij}$
5	$\neq 0$	$\neq 0$	$= 0$	$G_{ij} = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \varepsilon_{ij}$
6	$= 0$	$= 0$	$\neq 0$	$G_{ij} = \mu + \beta_3 Idd5_i \times Idd13_i + \varepsilon_{ij}$

Table 3.1: *6 groups of simulated data with different values of parameters*

We generate 500 simulated datasets of each group then there are 3000 simulated datasets in total. Each dataset from each group contains 12 data point with 3 data points for each strain. The day effect is not considered in our simulation study. Our goal is to detect all the 500 datasets in group 6 generated by model $G_{ij} = \mu + \beta_3 Idd5_i * Idd13_i + \varepsilon_{ij}$. For our two-stage method, we expect that all the data in groups 1, 3, and 6 will pass the first stage since they have interaction effect ($\beta_3 \neq 0$). However only the data in group 6 will pass the second stage since they do not have main effects ($\beta_1 = 0$ and $\beta_2 = 0$).

For the first stage, we set the test hypothesis to be

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_a : \beta_3 \neq 0$$

We use F statistics and p-values to identify the non-zero interactions. For different values of β_1, β_2 and β_3, we set significance level to be 0.05. That means if p-value is smaller than 0.05, then we declare the data passes the first stage. Table 3.2 shows the power probabilities of that the data in group 1, 3 and 6 pass the first stage. Table 3.3 shows the size probabilities of that the data in group 2, 4, and 5 pass the first stage. We simply set the values of β_1, β_2 and β_3 are equal. From previous discussion

Group	Non-Zero Values of β_1, β_2 andβ_3			
	0.7	0.5	0.35	0.2
1	94%	74%	52%	14%
3	93%	74%	55%	18%
6	93%	71%	52%	12%

Table 3.2: *The power probability of simulated data that passed the first stage*

Group	Non-Zero Values of β_1 and β_2			
	0.7	0.5	0.35	0.2
2	3.8%	5.2%	5.2%	4.8%
4	5.6%	4.6%	5.6%	6%
5	4.6%	5%	4.8%	6.4%

Table 3.3: *The size probability of simulated data that passed the first stage*

of identifying data with interaction, we declared that most of such data would be detected if the value of β_3 is greater than 0.75. Then we explore more about the cases with small value of β_3. We can observe from Table 3.2 that, when the value of β_3 is 0.7, there are 94%, 93%, and 93% of data in group 1, 3, and 6 passed the first stage, respectively. The result is very close to our expectation, so it is good. Compared with the result when value of $\beta_3 = 0.7$, the result when value of $\beta_3 = 0.5$ is worse but still acceptable. When the value of β_3 is no less then 0.35, the result is still acceptable. Most data from group 1, 3, and 6 passed stage 1, while only 4 to 5 percent of data from each of group 2, 4 and 5 passed the first stage. It is an acceptable result, so we move forward to the stage 2.

For stage 2, we set the test hypothesis to be

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

For the null hypothesis, the model is

$$Y_{i,g} = \mu_g + \beta_3 Idd5_i \times Idd13_i + \varepsilon_i \tag{3.2}$$

and for the alternative hypothesis, the model is

$$Y_{i,g} = \mu_g + \beta_1 Idd5_i + \beta_2 Idd13_i + \beta_3 Idd5_i \times Idd13_i + \varepsilon_i \tag{3.3}$$

We tried to use test statistics F for this hypothesis testing. It is based on the gene-specific residual sums of squares, denoted by RSS_g, and the residual degrees of freedom, denoted by df. Hypothesis testing involves the comparison of two models, and test statistics are computed on a gene-by-gene basis. Thus, we can suppress the subscript g and use the notation RSS_0, df_0 for the null model and RSS_1, df_1 for the alternative model residual sums of squares and degrees of freedom, respectively (Wu *et al.*, 2003). The statistic F is the usual F statistic that one would compute if data were available for only a single gene,

$$F_1 = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1}$$

It generalizes the t-test approach that is widely used in microarray analysis (Dudoit *et al.*, 2002). Significance levels can be established by reference to the standard F distribution or by permutation analysis. This test does not require the assumption of common error variance. However, it has low power in typical microarray experiments because of small sample sizes and it can be sensitive to variations in the estimates of residual variance, RSS_1 (Wu *et al.*, 2003).

For the simulation data which passed stage 1, we use F statistics to detect the data with $\beta_1 = \beta_2 = 0$. We only illustrate the result of the data with $\beta_3 = 0.7$. We set the significance level to be 0.05, then calculate the p-values for each data passed the first stage. For any data with p-value smaller than 0.05, we declare that it passes the second stage. The result is that, 92% of data from group 6, 2.4% of data from group 2 passed stage 2, and no data from other group passed stage 2. From the Figure 3.1, we can observe that it is the case of interaction we look for. The data is with locus-locus interaction and the main effect of each locus is almost 0. Due to this acceptable result

then we apply this method to the microarray data.

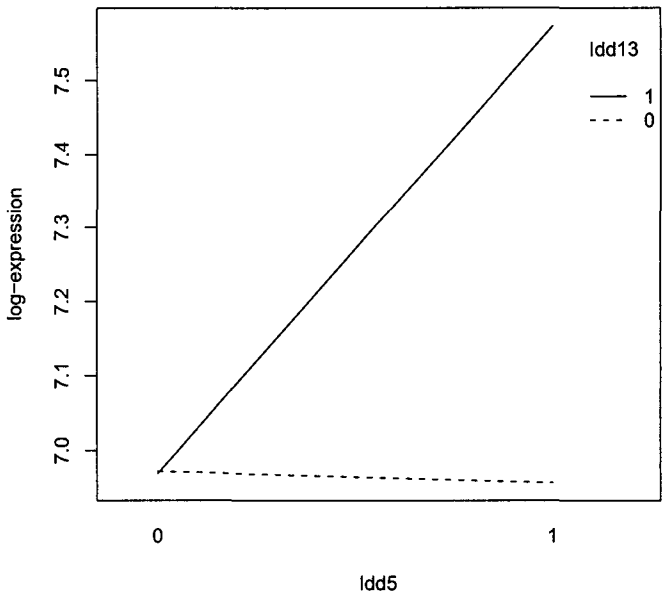

Figure 3.1: *The simulated data with special case of interaction that passed the second stage*

Chapter 4

Application of the Two-stage Method

In Chapter 2 we gave the initial analysis to detect interaction using SAM method, then we start with the stage 2 analysis in Chapter 4. There are 14 genes detected with significant interaction in Chapter 2 if we set the significance level to be 0.05. However, for better detection that determine which genes can pass the second stage we need to allow more genes to pass the first stage. Then we set a more liberal significance level of 0.2 to let more genes be detected. For our real microarray data, we have 1921 genes passed the first stage if the significance level is set to be 0.2. Next we look for the genes not affected by Idd5 or Idd13, say $\beta_1 = \beta_2 = 0$. We set the test hypothesis to be

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

The null model and the alternative model are similar to the models (3.4) and (3.5) in Chapter 3. The difference is that, we add the Day Effect into the models as a factor because the day an experiment completed introduces non-biological variation. For the

null hypothesis, the model is

$$Y_{g,i} = \mu_g + \alpha_g D_i + \beta_{3,g} Idd5_i \times Idd13_i + \varepsilon_{g,i} \tag{4.1}$$

and for the alternative hypothesis, the model is

$$Y_{g,i} = \mu_g + \alpha_g D_i + \beta_{1,g} Idd5_i + \beta_{2,g} Idd13_i + \beta_{3,g} Idd5_i \times Idd13_i + \varepsilon_{g,i} \tag{4.2}$$

where D_i is the day indicator, $Idd5_i = 1$ whenever locus Idd5 is NOD derived and 0 otherwise. Similarly, $Idd13_i = 1$ whenever Idd13 is NOD derived and 0 otherwise. $\varepsilon_{g,i}$ is the random error term.

Hypothesis testing involves the comparison of two models, thus, we can use the notation RSS_0, df_0 for the null model and RSS_1, df_1 for the alternative model residual sums of squares and degrees of freedom, respectively. We used F statistics to detected the genes

$$F = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1} \tag{4.3}$$

We calculated the F statistic for each gene that passed the first stage, and the information of the top genes with smallest F statistic are shown in Table 4.1. From Table 4.1, we obtain two genes with very small F statistics. The genes with small F statistics might has no main effect by locus Idd5 or Idd13; however, we cannot determine which genes pass the second stage based on the F statistic only. We need to have those genes with q-values smaller than the significance level. Because the small q-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. It indicates the strength of evidence for say, rejecting the null hypothesis H_0.

Figure 4.1 shows the interaction plot of gene with the smallest second stage F statistic. We observe that it is like the special case of interaction we look for. The log-

Index	Gene ID	NOR mean	Idd5 mean	Idd13 mean	Idd5/13 mean	F stat
1	92411_at	9.867	9.864	9.867	10.0592	0.00150
2	160096_at	8.955	8.969	8.958	8.823	0.0608
3	161410_r_at	3.873	3.919	3.849	3.557	0.206
4	94865_at	6.604	6.640	6.631	6.507	0.294
5	160925_at	8.374	8.419	8.438	8.216	0.363
6	97896_r_at	5.334	5.383	5.408	5.192	0.365
7	160328_at	8.327	8.304	8.285	8.446	0.392
8	160930_at	7.272	7.302	7.383	7.055	0.405

Table 4.1: *The information of genes in stage 2 analysis.*

expression stays almost the same when neither or one of Idd5 and Idd13 is from NOD mice, but it increases when both of Idd5 and Idd13 are from NOD mice. In other words, the Double Congenic NOR-NOD.Idd5/13 is different from the three others NOR, NOR-NOD.Idd5 and NOR-NOD.Idd13. Figure 4.1 shows some evidence that gene 92411_*at* has the type of interaction we look for, but we cannot declare that it passes the second stage because it may look like one straight line in a large scale plot. The q-values for the genes passed the first stage can be interpreted as the probability that the main effect is present if we declare that it is present given the interaction effect is present. There is no such small q-value to give us confidence to declare that gene 92411_*at* can pass the second stage.

Similarly to the first stage analysis, we permute the labels B ($B = 500$) times. Different from that, we permute the only 2 strain labels within days to recalculate F statistic 500 times. Our goal is to detect the genes without main effect; in other words,

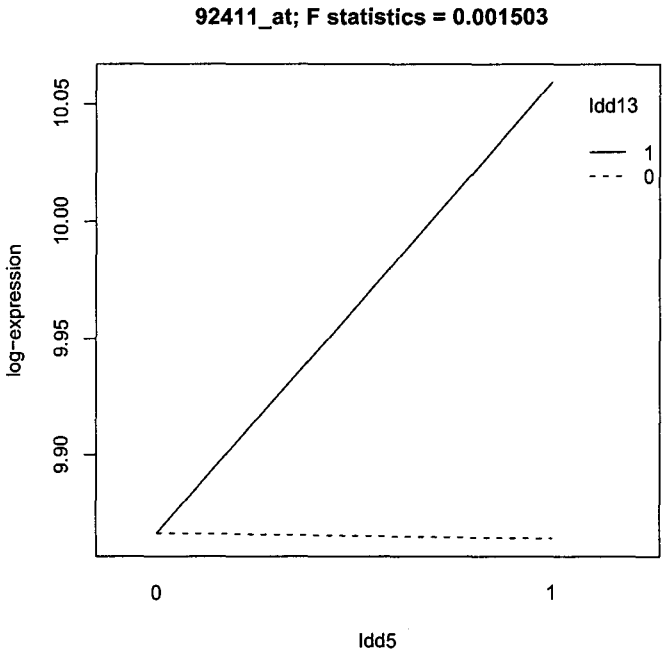

Figure 4.1: *The interaction plot of gene with smallest F statistics in stage 2 analysis*

our goal is to detect the genes which have the interaction effect by Double Congenic NOR.NOD-Idd5/13 different from the effects of other three strains. Therefore, we consider the three strain labels for NOR, single congenic NOR.NOD-Idd5 and NOR.NOD-Idd13 to be one group, and the strain labels for double congenic NOR.NOR-Idd5/13 to be another group. Then we permute the strain labels within each group and each day. Estimate the expected order statistics by $\bar{F}_{(i)} = (1/B)\sum_1^B F_{(i)}^{*b}$ for $i = 1, 2, ..., G$, which is the mean of the order statistic from the permutations. Then we plot the average order statistic (exptected F statistic) from the permutations $(\bar{F}_{(1)}, ..., \bar{F}_{(G)})$ against the observed $(F_{(1)}, ..., F_{(G)})$. We found that the expected F statistic is much smaller than the observed F statistic for each gene. By this observation we realized

that the q-values must be very large for almost all of the genes.

We are not confident to declare that any gene pass the second stage, even though the interaction plot visually shows some evidence that the interaction is of the type for which we are looking. It might be because the permutations are not simulated from the correct null distribution. The null hypothesis we wish to test is $H_0 : \beta_1 \neq 0$ or $\beta_2 \neq 0$, but not $H_0 : \beta_1 = \beta_2 = 0$. However, this kind of test of hypothesis is hard to realize, then we compromise to test the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. The incorrect null hypothesis may have led to the wrong permutation therefore, we cannot calculate the relevant q-values. To avoid the situation, we switch to bootstrap method.

Chapter 5

Bootstrap Approach

The motivation for using bootstrap instead of permutation is that we do not have the permutations from correct null distribution, then we cannot get confidence to declare any gene passed the second stage. To avoid that, we propose to use bootstrap method instead of permutations to resample the real 12488 gene expression data, called Y_g. For stage 1, we need to test hypothesis

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_a : \beta_3 \neq 0$$

therefore the null model is

$$M_0 : Y_i = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \varepsilon_i$$

. Then the estimated model is

$$\hat{M}_0 : \hat{Y}_i = \hat{\mu} + \hat{\beta}_1 Idd5_i + \hat{\beta}_2 Idd13_i$$

where $i = 1, 2, ..., 20$.

Let $r_i = Y_i - \hat{Y}_i, i = 1, ..., n$, then resample $(r_1, r_2, ..., r_n)$ with replacement R times, say $(r_1, r_2, ..., r_n) \rightarrow (r_1^*, r_2^*, ..., r_n^*)$. Y_i^* is the resampled data generated from bootstrap. Let $Y_i^* = \hat{Y}_i + r_i^*$, and fit the R corresponding models

$$Y_i = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \varepsilon_i \tag{5.1}$$

For each model, we calculate the d statistic

$$d_i^* = \frac{\hat{\beta}_{3,i}^*}{se(\hat{\beta}_{3,i}^*)} \tag{5.2}$$

We ranked each set of d^* by ascending order as $(d_1^*, ..., d_G^*)$,then we calculate the expected d statistic from bootstrap method by $\bar{d} = (1/R) \sum_1^R d^*$, and plot the expected d statistics against observed d statistics as with permutation method. The calculation of FDR and q-values are the same to the permutation method. The bootstrap approach analysis of stage 1 obtain the similar result as permutation approach. We declare that, all the genes with q-values smaller than 0.2 pass the first stage. Then there are 1250 genes that passed the first stage by bootstrap approach. The genes passed the first stage are almost the same as those listed in Table 2.1 especially the top genes. The top 10 genes passed the first stage is shown in Table 5.1.

For the second stage, we have the null model is

$$M_0 : Y_i = \mu + \beta_1 Idd5_i + \beta_2 Idd13_i + \beta_3 Idd5_i * Idd13_i + \varepsilon_i$$

, and the alternative model is

$$M_a : Y_i = \mu + \beta_3 Idd5_i * Idd13_i + \varepsilon_i$$

. The estimated model is

$$\hat{M}_0 : \hat{Y}_i = \hat{\mu} + \hat{\beta}_1 Idd5_i + \hat{\beta}_2 Idd13_i + \hat{\beta}_3 Idd5_i * Idd13_i$$

Index	Gene ID	NOR mean	Idd5 mean	Idd13 mean	Idd5/13 mean	q-value	absolute d stat
1	93595_at	10.529	10.993	11.0555	10.295	0.002	5.706
2	92942_at	8.212	8.332	8.363	8.0158	0.002	5.625
3	160118_at	11.0812	11.483	11.452	11.229	0.01	4.793
4	93881_i_at	5.738	6.0468	6.0507	5.917	0.0177	4.557
5	93097_at	12.362	12.608	12.528	11.860	0.0177	4.379
6	160714_at	7.375	7.823	7.846	7.714	0.0177	4.347
7	93498_s_at	9.069	9.529	9.512	9.229	0.0453	3.970
8	102983_at	7.244	7.547	7.566	7.466	0.045	3.943
9	160811_at	7.704	8.0384	8.025	7.719	0.048	3.878

Table 5.1: *The top genes passed the first stage (Bootstrap approach).*

, where $i = 1, 2, ..., n$.

Let $r_i = Y_i - \hat{Y}_i$, then resample $(r_1, r_2, ..., r_n)$ with replacement R times, say $(r_1, r_2, ..., r_n) \rightarrow (r_1^*, r_2^*, ..., r_n^*)$. Y_i^* is the resampled data generated from bootstrap, and let $Y_i^* = \hat{Y}_i + r_i^*$, then fit M_0 to Y_i^*. Calculate the F statistic

$$F_i^* = \frac{(RSS_{0,i} - RSS_{1,i})/(df_0 - df_1)}{RSS_{1,i}/df_1} \tag{5.3}$$

where RSS_0 and $RSS_1$is the sum square of residual of the null model and alternative model, respectively. The df_0 and df_1 are the degree of freedom null model and alternative model, respectively. Then we calculate a set of F_i^*, $i = 1, ..., N$ in each bootstrap, so there are $R \times N$ values of F_*. We calculate the bootstrap based p-values for each

gene passed the first stage by formula

$$p_i = \#(F^* <= F_i)/RG \tag{5.4}$$

where R is the number of bootstraps and N is the number of genes passed the first stage. Sort the bootstrap based p-values to obtain the order p-values such as $p(1) \leq p(2) \leq ... \leq p(N)$. Storey and Tibshirani (2003) gives a method to convert such p-values to q-values by following formulas:

$$\hat{q}(p_{(N)}) = \hat{\pi}_0 p_{(N)} \tag{5.5}$$

$$\hat{q}(p_{(i)}) = \min(\frac{\hat{\pi}_0 N p_{(g)}}{i}, \hat{q}(p_{(i+1)})) \tag{5.6}$$

where $i = N-1, ..., 1$. They also give a method to estimate π_0 but the method we introduced in Chapter 1 is fine. The q-values calculated are the second stage q-values.

let q_1 be the vector of the first stage q-values for those genes selected to go forward to the second stage and q_2 is the vector of the second stage q-values for those genes with small F statistic, we then we can calculate an overall q-value by

$$q_i = 1 - (1 - q_{1,i}) * (1 - q_{2,i}) \tag{5.7}$$

where $i = 1, ..., N$. Therefore, the overall q-value must be higher than each of q_1 and q_2. The first stage q-value q_1 can be interpreted as the probability of that the interaction is present if we declare that it is present; and the second stage q-value q_2 can be interpreted as the probability of that the main effect is present if we declare that it is present given that interaction effect is present. q_2 represents a conditional probability given that the interaction effect is present. Then the formula can give a set of overall q-values. The overall q-values can be interpreted as the probability that

the effect of interest is truly present if we declare that it is present for this gene, then the inference is based on the overall q-values.

The top genes with small F^* statistics and q-values are shown in Table 5.2. From this table, we can observe that even though the genes have small F statistics; however, both of the second stage q-values and the overall q-values are large. Figure 5.1 shows that the mean log-expressions of strain NOR, NOR.NOD-Idd5 and NOR.NOD-Idd13 are about the same values; while the mean log-expression of strain NOR.NOD-Idd5/13 is quite smaller than any one of those three strains. That indicates that the interaction effect between to loci are quite greater than the main effects by either locus alone. However, we can not declare that the genes can pass the second stage at a reasonable significance level due to the large q-values even though Figure 5.1 visually shows some evidence that there is the interaction we look for.

Index	Gene ID	NOR mean	Idd5 mean	Idd13 mean	Idd5/13 mean	F stat	First stage q-value	overall q-value
1	160096_at	8.955	8.969	8.958	8.823	0.0608	0.531	0.619
2	161410_r_at	3.873	3.919	3.849	3.557	0.206	0.531	0.608
3	94865_at	6.604	6.640	6.631	6.507	0.294	0.531	0.624
4	160925_at	8.374	8.419	8.438	8.216	0.363	0.531	0.595
5	97896_r_at	5.334	5.383	5.408	5.192	0.365	0.531	0.618
6	160930_at	7.272	7.302	7.383	7.055	0.405	0.531	0.623
7	94893_at	7.080	7.106	7.087	6.978	0.421	0.531	0.590
8	160114_at	10.707	10.851	10.820	10.497	0.441	0.531	0.612
9	104056_at	7.994	8.118	8.160	7.697	b 0.449	0.531	0.591
10	104572_at	8.540	8.588	8.579	8.438	0.456	0.531	0.605

Table 5.2: *The top genes in the second stage analysis.*

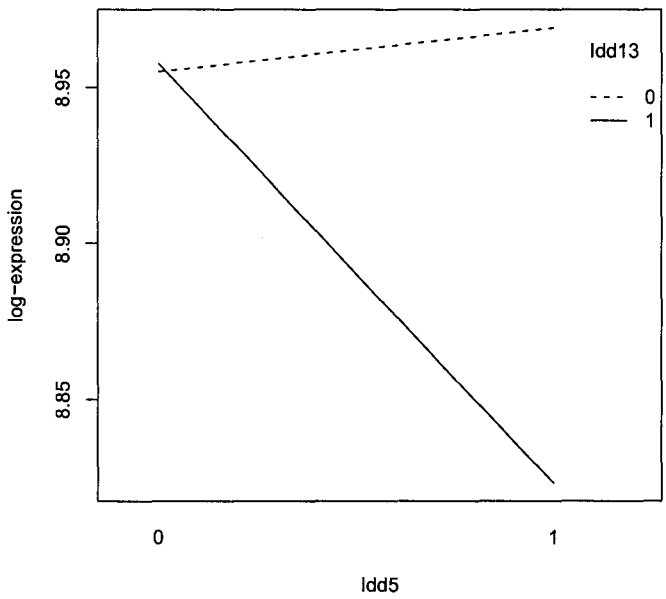

Figure 5.1: *The interaction plot of gene with smallest F statistics in the second stage analysis (Bootstrap approach).*

Chapter 6

Discussion and Future Work

In this project, we have researched to develop a methodology for detecting the special case of locus-locus interaction using microarray data. We have detected the interaction using our dataset and then selected genes to go forward to the second stage. The interaction plot of gene with smallest F-statistics visually showed some evidence that the interaction of interest is detected; however, they have very high q-values in the second stage analysis. We calculate the overall q-values for the genes, but they are even greater than the second stage q-values. Because overall q-value is calculated by $q = 1 - (1 - q_1) * (1 - q_2)$, it must be greater than either of q_1 and q_2. Due to the high overall q-values we are not confident to declare any gene passed the second stage.

A further question we may ask is how to calculate reasonable q-values. In our analysis, the difficulty of detecting genes that passed the second stage is that genes have high q-values. Even for those gene with small F statistic they still have large q-values. Due to this we cannot get enough statistical evidence to determine which genes can pass the second stage. In future work, the method of calculating q-values

could be of concern.

In future work it is necessary to consider some other test statistics. It might be helpful to have a simulation study and try to find an appropriate test statistic. We used modified Wald t-statistics (we call d statistics) and F statistic in our analysis. The result of detecting locus-locus interaction in stage 1 analysis is reasonable but the result of the second stage analysis does not meet our expectation. Someone may concern more about the test statistics in the second stage analysis and seek a better statistic.

We also suggest someone think about the null distribution we used to produce permutation. The null hypothesis the permutations coming from in the second stage is not correct so that we may be not confident to declare genes passed the second stage even though we can have reasonably smaller q-values.

Appendix A

Partial R Codes for Simulation

```
# The function generate is the to generate the random data using
#model.
#The parameter n is the number of simulated genes.
#The parameter replicate is the number of replicates of the same experiments.


#The function gnt is to generate the random data of model using
#model Y_ij = mu + beta_1 *Idd5_ +beta_2*Idd13_i +
#beta_3*Idd5_i*Idd13_i + e_ij


gnt<-function(n,mu,sigma,replicate,beta1,beta2,beta3) {
   group1=rnorm(n*replicate,mu,sigma)
   group2=rnorm(n*replicate,mu,sigma)+beta1
```

```
  group3=rnorm(n*replicate,mu,sigma)+beta2
  group4=rnorm(n*replicate,mu,sigma)+beta1+beta2+beta3
  matrix(c(group1,group2,group3,group4),ncol=4*replicate)
}

#group1 is the generated data with Idd5=0 and Idd13=0
#group2 is the generated data with Idd5 is not equal to 0 but Idd13=0
#group1 is the generated data with Idd13 is not equal to 0 but Idd5=0
#group1 is the generated data with neither Idd5 nor Idd13 equal to 0
#Then combined these 4 groups into one matrix
#The matrix is a n* (4*replicate) matrix
```

Appendix B

Partial R codes for calculating s_0

```
#The function Cal.S_0 is used to calculate s_0

Cal.S_0<-function(s,no.gene) {
   #Sort the r_i and s_i by ascending order of r_i
   #The s_i is sorted with the corresponding r_i
   o<-order(s[,2])
   s.order<-s[o,]
   r.rk<-s.order[,1]
   s.rk<-s.order[,2]

   #Let s.alpha to be the 0 to 20 quantile of s_i.
   s.alpha<-quantile(s.rk,(0:20)/20)

   #q is the 100 quantiles of s_i values
```

```
    q<-quantile(s[,2],probs=seq(1:100)/100)

    index=ceiling(no.gene*(seq(0,100)/100))
    alpha=seq(0,1,by=0.05)
    v.i<-rep(NA,100) cv.alpha=rep(NA,21)
    for (j in 1:21) {
       for (i in 1:100) {
          d.alpha=r.rk[(index[i]+1):index[i+1]]/(s.rk[(index[i]+1):index[i+1]]
          +s.alpha[j])
          #Function mad calculate the median absolute deviation
          #from the median, divided by .64
          v.i[i]=mad(d.alpha)
          }
       #cv.alpha is the coefficient of variation of the v_j values
       cv.alpha[j]=sd(v.i,na.rm=TRUE)/mean(v.i,na.rm=TRUE)
     }

    #min.position is the position of the smallest element in vector cv.alpha
    min.position<-which.min(cv.alpha)
    s.alpha[min.position]
}
```

Appendix C

R Codes for Calculating d Statistics

```
#Load affy library into R
library(affy)

#Read in the raw data of microarray data
exp1.data<-ReadAffy(widget=TRUE)


#rma() normalize the raw data and then y1 is the gene expression

#The columns of y1 represent the cases and the rows represent the gene
exp1.rma<-rma(exp1.data)
y1<-exprs(exp1.rma)

#Get the number of genes and the number of groups
```

```
no.gene<-nrow(y1)
no.group<-ncol(y1)

#Day is the label of day factor
Day<-factor(rep(c(0,1),c(12,8)))


#I5 and I13 is the labels of Idd5 and Idd13 factors, respectively.
I5<-c(rep(0,3),rep(1,3),rep(0,3),rep(1,3),rep(0,2),rep(1,2),rep(0,2),rep(1,2))
I13<-c(rep(0,6),rep(1,6),rep(0,4),rep(1,4))

#Get the estimate of beta3 and the standard error of estimate of beta3
est.beta3<-array(0,c(no.gene,2))
for (i in 1:no.gene) {
   est.beta3[i,]<-summary(lm(y1[i,]~Day+I5*I13))$coefficients[5,1:2]
}

#Calculate the d statistics of the origial data
r.i=est.beta3[,1];s.i=est.beta3[,2]
s_0<-Cal.S_0(est.beta3,no.gene)
d.i<-r.i/(s.i+s_0)
#ordered.d.i is the observed d statistics for each gene
ordered.d.i=sort(d.i)
```

Appendix D

R Codes for Calculating d Statistics from Permutation

```
#4 groups and 5 replicates for each
group<-c(rep(1:4,5))
#Number of permutations is 500
no.pmt<-500

#Estimates and standard deviations of beta3 for permutations
est.beta3.pmt<-array(0,c(no.gene,no.pmt,2))

for (i in 1:no.pmt) {
   #Permute the labels of the group
   pmt.group<-c(sample(group[1:12]),sample(group[13:20]))
```

```
   I5.pmt<-1*(pmt.group==2|pmt.group==4)
   I13.pmt<-1*(pmt.group>=3)
   for (j in 1:no.gene) {
      #Calcuate the estimate of beta3 and standard error of estimate of beta3
      est.beta3.pmt[j,i,]<-summary(lm(y1[j,]~Day+I5.pmt*I13.pmt))$coefficients[5,1
      }
}

#d.g.pmt is the ranked d statistics from each permutation
d.g.pmt<-array(NA,c(no.gene,no.pmt)) for (i in 1:no.pmt) {
   d.g.pmt[,i]<-sort(est.beta3.pmt[,i,1]/(est.beta3.pmt[,i,2]+s_0))
}

#Find the mean of the rows, which is the expected d statistics
mean.d.g<-rowMeans(d.g.pmt)
```

Appendix E

Partial R Codes for FDR and q-value for Stage 1

```
#Exp is the expected d statistics
#Obs is the observed d statistics
#M is the median of the expected d statistics
Exp<-mean.d.g;Obs<-ordered.d.i
M<-median(Exp)

#Divide the Obs and Exp into two parts: upper part and lower part
Obs.u<-Obs[which(Exp>=M)];Exp.u<-Exp[which(Exp>=M)]
Obs.l<-Obs[which(Exp<M)];Exp.l<-Exp[which(Exp<M)]


perms=d.g.pmt
```

```
#Set the cut-off points
cuts=quantile(perms,c(0.25,0.75))
pi.0=min(1,sum(Obs>=cuts[2]|Obs<=cuts[1])/(0.5*no.gene))

#The function to calculate FDR
findFDR<-function(delta,Obs,Exp,perms) {
   diffs=sort(Obs)-Exp
   upper=min(Obs[Obs>=median(Obs) & diffs>=delta])
   lower=max(Obs[Obs<=median(Obs) & diffs<=-delta])
   no.detect.o=sum(Obs>=upper|Obs<=lower)
   no.detect.p=colSums(perms>=upper|perms<=lower)
   mean(no.detect.p)/no.detect.o*pi.0
}

ind.u<-which(Obs>=median(Obs))
delta.u<-rep(NA,length(ind.u))
FDR.u<-rep(NA,length(ind.u))
qvalue.u<-rep(NA,length(ind.u))


#First value of the FDR and q-value
delta.u[1]<-max(0,Obs[ind.u[1]]-Exp[ind.u[1]])
FDR.u[1]<-findFDR(delta.u[1],Obs,Exp,perms)
qvalue.u[1]<-FDR.u[1]
```

```
for (i in 2:length(ind.u)) {
   delta.u[i]<-max(delta.u[1:(i-1)],Obs[ind.u[i]]-Exp[ind.u[i]])
   FDR.u[i]<-findFDR(delta.u[i],Obs,Exp,perms)
   qvalue.u[i]<-min(FDR.u[1:i])
}
```

Appendix F

Partial R Codes for Calculating d Statistics from Bootstrap

```
residual=y_hat=matrix(NA,ncol=no.group,nrow=no.gene)

for (i in 1:no.gene) {
   #mod is the linear model
   #y-hat is fit to the model and residual
   #contains the residuals of the meodel
   mod<-lm(y1[i,]~Day+I5+I13)
   y_hat[i,]<-fitted(mod)
   residual[i,]<-residuals(mod)
}

#Resample residual R times
```

```
R<-500

d_star<-matrix(NA,ncol=R,nrow=no.gene)
for (i in 1:R) {
   #Resample residual and then fit the corresponding model
   i1<-sample(1:20,replace=T)
   r_star<-residual[,i1]
   y_star<-y_hat+r_star
   beta_star<-matrix(NA,ncol=2,nrow=no.gene)
   for (j in 1:12488) {
      beta_star[j,]<-summary(lm(y_star[j,]~Day+I5*I13))$coefficients[5,1:2]
   }
   #d_star is the d statistiscs from each bootstrap
   d_star[,i]<-sort(beta_star[,1]/beta_star[,2])

}

#Find the mean of the rows, which is the expected d statistics
mean.d.star<-rowMeans(d_star)
```

Appendix G

Partial R Codes for Calculating q-values (Bootstrap Approach)

```
#Exp2 and Obs2 are the expected F statistics and observed F
#statistcs, respectively.
Exp2<-f.star.exp;Obs2<-f.obs


boots=f.star.sort
#cuts2 is the cut-off point
cuts2<-quantile(boots,0.75)
pi.0.2<-min(1,sum(Obs>=cuts2)/(0.5*no.gene))


#findFDR is the function to calculate false discovery rate
findFDR<-function(delta2,Obs2,Exp2,boots) {
```

```
   diffs2<-Obs2-Exp2
   upper2<-min(Obs2[diffs2>=delta2])
   #Detecting significant genes for calculating FDR
   no.detect.o2<-sum(Obs2<=upper2)
   no.detect.p2<-colSums(boots<=upper2)
   mean(no.detect.p2)/no.detect.o2*pi.0.2
}



delta.u2<-rep(NA,length(Obs2))
FDR.u2<-rep(NA,length(Obs2))
qvalue.u2<-rep(NA,length(Obs2))

#Calculating q-values
delta.u2[1]<-max(0,Obs2[1]-Exp2[1])
FDR.u2[1]<-findFDR(delta.u2[1],Obs2,Exp2,boots)
qvalue.u2[1]<-FDR.u2[1]

for (i in 2:length(Obs2)) {
   delta.u2[i]<-max(delta.u2[1:(i-1)],Obs2[i]-Exp2[i])
   FDR.u2[i]<-findFDR(delta.u2[i],Obs2,Exp2,boots)
   qvalue.u2[i]<-min(FDR.u2[1:i])
}
```

Bibliography

[1] Benjamini Y. and Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B,* 57: 289-300

[2] Bolstad B. M., Irizarry R. A., Astrand M. and Speed T. P. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19,2: 185-193

[3] Dudoit S., Fridlyand J., and Speed T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association,* 97(457): 77

[4] Chu G., Narasimhan B., Tibshirani B. and Tusher V. (2005) *SAM "Significance Analysis of Microarrays" Users guide and technical document*

[5] Cordell H. J., Todd J. A., Hill N. J., Lord C. J., Lyons P. A., Peterson L. B., Wicker L. S., and Clayton D. G. (2001). Statistical Modeling of Interlocus Interactions in a Complex Disease: Rejection of the Multiplicative Model of Epistasis in Type 1 Diabetes. *Genetics*, Vol. 158: 357-367.

[6] Knudsen S. (2004) *Guide to Analysis of DNA Microarray Data*, Second Edition. John Wiley and Sons, New York.

[7] Irizarray R. A., Bolstad B. M., Collin F., Cope L. M., Hobbs B. and Speed T. P. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research 2003*: 31(4): 15.

[8] Irizarray R. A., Gautier L. and Cope L. M. (2003b) An R Package for Analyses of Affymetrix Oligonucleotide Arrays. In *The Analysis of Gene Expression Data: Methods and Software* (Parmigiani G., Garrett E. S., Irizarry R. A. and Zeger S. L. editors) Springer-Verlag New York.

[9] Lee M. T. (2004) *Analysis of Gene Expression*, Kluwer Academic Publishers.

[10] McLachlan G. J., Do K. and Ambroise C. (2004). *Analysis of Gene Expression*, John Wiley & Sons, Inc., Hoboken, New Jersey.

[11] Storey J. D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31: 2013-2035.

[12] Storey J. D. and Tibshirani R. (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100: 9440-9445.

[13] Tusher V., Tibshirani R. and Chu, G. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences,* 98: 5116-5121.

[14] Wu H., Kerr M. K., Cui X. and Churchill G. A. (2003) MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. *The Analysis*

of Gene Expression Data: Methods and Software (Parmigiani G., Garrett E. S., Irizarry R. A. and Zeger S. L. editors) Springer-Verlag New York.