STATISTICAL APPROACHES FOR THE ANALYSIS OF MATCHED MRNA MICROARRAY DATA FROM DEGRADED TISSUES WITH APPLICATION TO UNFROZEN ARCHIVED NEWBORN BLOOD SPOTS FROM A CASE-CONTROL STUDY OF CHILDREN WITH CEREBRAL PALSY

By

Nhan Thi Ho

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment for the requirements
for the degree of

DOCTOR OF PHILOSHOPHY

Epidemiology

2012

ABSTRACT

STATISTICAL APPROACHES FOR THE ANALYSIS OF MATCHED MRNA MICROARRAY DATA FROM DEGRADED TISSUES WITH APPLICATION TO UNFROZEN ARCHIVED NEWBORN BLOOD SPOTS FROM A CASE-CONTROL STUDY OF CHILDREN WITH CEREBRAL PALSY

By

Nhan Thi Ho

Cerebral palsy (CP) describes a group of defects that are caused by damage to the motor-controlling centers of the brain. This damage occurs either during pregnancy, during childbirth, or in early infancy. Currently the etiology of CP is unclear but has been speculated to arise from hypoxia, infection and other influences. In this matched case-control study in children aged from 2-16 years, we examined the mRNA expression patterns in blood for evidence of exposure to agents that have been associated with the development of CP. The prospective collection of newborn blood samples derived from CP cases and matched controls is not practical while archived unfrozen dried neonatal blood spots (uDNBS) have been showed to preserve a sufficient amount of mRNA to perform mRNA microarray analysis. Therefore, we utilized previously collected uDNBS for genome-wide expression profiling.

mRNA expression data was derived from a set of 106 uDNBS, which represented 53 subjects that subsequently developed CP and 53 age, gestational- age and gender- matched control subjects. Established methods for processing and analyzing of microarray data were used to study evidence of changes in gene expression between cases and controls. The analysis focused on a gene set-based approach prioritizing seven pre-selected gene sets representing four major hypothesized pathophysiologic pathways of CP, i.e. inflammation, thyroid disorders, hypoxia/asphyxia, and coagulation disorders. The empirical inflammatory and hypoxic gene sets

were significantly down-regulated while the empirical thyroidal gene set appears significantly up-regulated. The analysis of gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database also revealed some significant inflammatory related gene sets. Gestational age and CP type had interactive effects on the expression pattern of the three significant empirical gene sets.

Several important technical and theoretical concepts were also evaluated in detail. First, the time-dependent degradation of mRNA, or the difficulty in extracting mRNA from uDNBS over time, is inevitable, and this may affect the technical quality of microarray data produced from uDNBS. Thus, the quality issues of microarray data need to be taken into account when processing and analyzing microarray data from uDNBS. Further evaluation of the quality of microarray data over time showed that differential expression at individual gene and gene set level could be seen better in uDNBS of less than six years old. The proposed approach for selecting housekeeping genes helped pick up six potential housekeeping genes which can be used for quantitative polymerase chain reaction (qPCR) assays to validate microarray data.

Second, the published literature for gene set analysis of matched case-control study design is meager, and existing microarray analysis methods may not function properly. Thus, the performance of existing methods was evaluated and new approaches have been developed to address many methodological aspects of gene set analysis of matched microarray data. Both the published GAGE (generally applicable gene set enrichment for pathway analysis) method and the proposed ZZ-GSA (two stage z-test for gene set analysis) approach can be used for gene set analysis of matched microarray data although each has some strengths and limitations especially in term of power and type I error.

ACKNOWLEDGMENTS

I would like to thank the Vietnam Education Foundation for giving me the opportunity and partially funding my PhD program. I would like to thank the OWL team who have worked hard to provide me with the data used in this dissertation and to provide me with an inspiration for my research ideas. I would like to thank Dr. Wenjiang Fu for his encouragement when I started working with microarray data and for his critique of my methodological approaches. I would like to thank Dr. Nigel Paneth, Dr Julia Busik, Dr. Kyle Furge, Dr. Qing Lu for their help, and for being on my dissertation committee. I especially would like to thank Dr Furge for having mentored me on many aspects such as programming, molecular biology, and methodology. I also would like to thank Dr Busik for having guided me on biological and laboratory issues. And certainly, I would like to thank Professor Paneth for being the advisor of my PhD program. I am grateful for his guidance, his patience and for all of the lessons he has taught me, either academic or non-academic, scientific or non-scientific, directly or indirectly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The original plan of this dissertation is to describe the results of pathway analysis of mRNA microarray data from uDNBS of CP cases vs. matched controls from an ongoing matched case-control study investigating the etiology of cerebral palsy (the Origins, Wellness and Life history of CP (OWL) study). Possible pathways to CP during the peri-partum period are examined by gene set analysis of mRNA microarray data and by evaluating the influence of clinical context during the peripartum period on gene expression pattern. The hypotheses and research aims for these analyses include:

Hypothesis 1: Inflammation, hypoxia/asphyxia, thyroid disorders in peri-partum period and coagulation disorders may be causal factors, or may contribute to the development of CP.

Aim 1: Use the best existing methods of gene set analysis for microarray data to evaluate the differential expression of empirical and canonical gene sets selected to represent the four pre-hypothesized pathways to CP.

Hypothesis 2: Gene sets selected for pre-hypothesized pathways may not fully represent the pathways of interest. In addition, the pre-hypothesized pathways themselves may not fully represent all possible pathways contributing to the development of CP. Other disorders, represented by different gene pathways during the peri-partum period may also contribute to the development of CP.

Aim 2: In addition to assessing gene sets of pre-hypothesized pathways, multiple gene sets representing various pathways from the KEGG database which includes clinically meaningful pathways should also be explored for their differences between cases and controls.

Hypothesis 3: The expression of genes in the causal pathways of CP may differ corresponding to covariates such as gestational age, CP type, or some newborn conditions.

Aim 3: Stratify case-control differences in gene expression on some clinical covariates such as CP type (hemiplegia, diplegia, quadriplegia), gestational age (term vs. preterm), etc.

The mRNA microarray data from uDNBS used in this dissertation are special due to the deterioration of microarray data quality over time of storage and the issues related to matching design. During the analytic process, some problems related to the methodology of processing, analyzing and validating of mRNA microarray data from heterogeneously degraded tissue of matched case-control study emerged. Published literature addressing these issues is meager and existing methods may not well handle different aspects of these issues. Therefore, these issues need to be discussed and possible methodological solutions need to be developed. As a result, an alternative structure of this dissertation was suggested as below.

ORGANIZATION OF THE DISSERTATION

Together with the introduction section for the dissertation, the main chapters of this dissertation include:

The first chapter covers a brief review about cerebral palsy, research on the etiology of cerebral palsy and the use of uDNBS. The OWL (Origins, Wellness and Life history of Cerebral Palsy) study, and laboratory techniques used for the OWL study will be described. This chapter also includes some review of overall procedures for analysis of mRNA microarray data.

The second chapter corresponds to the three hypotheses and specific aims for pathway analysis. This core chapter covers the results of pre-hypothesized gene sets and multiple gene sets from gene set database (KEGG) using existing statistical methods for matched microarray data. Pre-hypothesized gene sets emphasize the pathways of interest and reduce the need for

adjusting for multiple testing. Exploration of KEGG gene sets supplements and strengthens the results of pre-selected gene sets, and helps discover other potential pathways related to CP. Gene set analysis is also stratified by some important covariates to assess the effect of clinical context on gene expression patterns in CP cases vs. controls.

The third chapter covers processing and exploratory analysis of mRNA microarray data, including evaluation of overall distribution and other characteristics of microarray data before and after filtering, normalization, aggregating to gene level. This chapter helps understand the characteristics of the microarray data used, patterns of mRNA degradation, and thus, helps orient the subsequent processing and analysis approaches and their use for mRNA microarray data from uDNBS.

The fourth chapter covers qPCR validation for the results of mRNA microarray data. This chapter includes the approaches for selection of housekeeping genes, selection of genes for validation and the results of the comparison between qPCR and mRNA microarray data.

The fifth chapter provides an evaluation of existing methods (including the methods used in previous chapters) and proposed modified methods of gene set analysis of matched microarray data in term of power, type I error, and influence of missing value. This chapter will describe a rigorous simulation and permutation approach as well as an imputation procedure. This chapter helps evaluate the performance of gene set analysis methods of matched data based on the log fold change of expression between cases vs. controls and also helps interpret the results of gene set analysis described in previous chapters.

The sixth chapter is a conclusion chapter summarizing the results of all the above chapters, and discussing future research applications or directions that may come from the work of this dissertation.

CHAPTER 1. REVIEW

About CP

CP was first identified by an English surgeon named William Little in 1860. Initially, CP was known as "Cerebral Paralysis".[1 2] In earlier days, CP was believed to be mainly caused by asphyxia during birth or prematurity. However, in 1897, a neurologist named Sigmund Freud advocated the idea that difficult or premature birth was only a symptom of other effects on fetal development, not the cause.[3] Considerable research from the late 1980s up to date has shown that only a small percentage of CP cases results from lack of oxygen during birth.[4 5]

The diagnosis of CP can usually only be confirmed when the child reaches the age of 2 years. According to the report of CP committee consensus published in 2006,[6] "CP describes a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to non-progressive disturbances that occurred in the developing fetal or infant brain". "Motor disorders of CP are often accompanied by disturbances of sensation, perception, cognition, communication, and behavior, by epilepsy, and by secondary musculoskeletal problems". Also according to this report, CP may be classified by the nature and typology of the motor disorder (e.g. spasticity, dystonia, choreoathetosis, ataxia), anatomic distribution (e.g. quadriplegia, diplegia, hemiplegia), functional motor abilities (e.g. Gross Motor Function Classification System (GMFCS) with 5 levels), accompanying impairments (epilepsy, hearing, vision disorders, mental retardation, etc), cause and timing (e.g. CP of clear postnatal origins, CP as part of genetic syndrome, etc). CP classification by anatomic distribution is commonly used in clinical practice.

Epidemiology and economic impact of CP

According to some reviews published in 2006 and earlier, the prevalence of CP in the general population in Western countries is approximately 1-2 /1000 live births. Overall CP prevalence for the past 40 years is notably stable. Prevalence of CP may vary depending on how CP is defined, denominator used or real variation.[7 8 9] However, according to some recent US population-based studies, the prevalence of CP seems to be higher than 3/1000 school-age children.[10 11 12]

The prevalence of CP increases significantly when gestational age decreases. The prevalence of CP is 14.6% for children born at 22 to 27 weeks of gestation, 6.2% at 28 to 31 weeks, 0.7% at 32 to 36 weeks, and 0.1% in term infants. Prevalence of CP decreases significantly when GA is≥ 27 weeks. Spastic CP is predominant in preterm infants. Other non - spastic forms of CP are more common in term infants than in preterm infants.[13]

People with CP often require special and costly medical and educational services for their whole life. This makes CP an important health care burden for any country. According to a study in the US in 2003, economic cost, including lost income, for each CP individual is about $921,000.[14]

Review of studies investigating the etiology of CP

Although CP is a major public health problem, research on the etiology of CP is still rare. Most recent etiologic studies have used data from administrative or medical record databases,[15 16 17 18 19 20] and some studies have focused only on infants < 32 weeks gestation or < 1kg.[21 22 23 24] There has been only one recent US case-control study examining CP cases and controls that interviewed mothers.[25]

According to previous research, less than 20% of cerebral palsy can be attributed to intrapartum events and around 70-80% of CP cases are due to prenatal factors. Common prenatal factors that may contribute to the development of CP are intrauterine growth retardation, maternal thyroid abnormalities, intrauterine viral infections (e.g. CMV, Rubella), intrauterine infection/inflammation with a maternal response (consisting of chorioamnionitis), fetal inflammatory response, autoimmune and coagulation disorders, and some other factors such as cerebral dysgenesis, multiple birth, genetic factors. CP may result from a combination of antepartum and intrapartum insults. [26] [27] [28]

Although the role of difficult birth on the development of CP is smaller than it was believed in the early days, adverse obstetric events remain one of the leading causes of CP since around one-third of children with CP had one or more birth related event.[29] Intrapartum hypoxia-ischemia is found in around one-sixth of CP children while around one-fourth of term born infants with cord blood pH < 7.0 (an evidence of peri-partum asphyxia) develop neonatal neurologic morbidity and mortality.[30]

Inflammation has been shown to play in an important role in developing CP. Chorioamnionitis has proved to double the risk of CP. [31] [32] In preterm, especially in extremely preterm infants, the role of fetal inflammation on brain damage is more predominant. [33] [34] Microbial organisms are found in amniotic cavity of around one-fourth or more of all preterm births.[35] Fetal inflammation may be a major mechanism responsible for many complications during the perinatal period and infancy such as preterm birth, fetal periventricular leucomalacia, and CP. [36] [37] Prenatal infections and inflammation are often accompanied by asphyxia or peripartum hypoxia. This “double-hit effect” often occurs in neonatal brain damage.

Proinflammatory cytokines may serve as major mediators in brain injury in newborns with either or both perinatal asphyxia and bacterial infection. Understanding the balance between neurodamaging and neuroprotective effects of cytokines is essential to neutralize the pathologic effects of inflammation associated with brain damage.[38 39]

There are several situations in which thyroid hormone disorders may increase the risk of cerebral palsy. Transient hypothyroxinemia of prematurity (THOP) in which serum thyroid hormones T3, T4 is low but TSH is normal in premies till around 6 weeks after birth is common in extremely premature neonates born before 28 gestational weeks.[40 41 42 43 44 45 46] This condition has been shown to increase the risk of CP to 3.6-4.4 fold at age 2 and cognitive impairment in early childhood.[47 48] Thyroid hormone supplementation may improve mental development in premature newborns although this has not been proved.[49 50] Maternal iodine deficiency has been believed to be associated with endemic neurologic cretinism of the offspring. Other maternal thyroid disorders such as high TSH or low free T4 measured at 12 gestational week may associate with some adverse neurodevelopmental outcomes such as lower Intelligent Quotation (IQ), lower psychomotor development index (PDI),[51 52] lower neonatal behavioral assessment scale (NBAS).[53]

Perinatal coagulation disorders, commonly manifested by perinatal arterial stroke, are associated with all anatomic types of CP but much more strongly with hemiplegia.[54 55] More than one half of hemiplegic infants have at least one coagulation abnormality.[56] Neonatal ischemic stroke of different forms (including perinatal arterial ischemic stroke, presumed pre- or perinatal stroke, and cerebral sinovenous thrombosis) is a leading cause of congenital

hemiplegia[57] and can contribute to other long-term neurological impairment including seizures and cognitive disorders.[58]

Some studies have examined the genetic origin of CP. According a systematic review summarizing the results from more than 20 studies, CP has been found to be associated with some thrombophilic, cytokine, apolipoprotein E genes and some other SNPs. The most potential genes among these genes are factor V Leiden, methylenetetrahydrofolate reductase, lymphotoxin-a, tumour necrosis factor-a, eNOS and mannose binding lectin. However, metalysis has not confirmed these association.[59]

The use of uDNBS and gene expression from uDNBS

Gene expression in blood contains more than 80% overlap with the transcriptome of at least 9 organs, including the brain. [60] [61] uDNBS are available in many states and may be a rich resource for epidemiological research. [62] Previous pilot work from our team has shown that mRNA extracted from long-term unfrozen storage of archived unfrozen dried neonatal blood spots (uDNBS) leftover from newborn genetic screening is sufficient to perform mRNA microarray analysis for genome-wide expression profiling. [63] More recent follow-up studies have detected gender specific expression patterns although the dynamic range of the mRNA expression is severely compressed.[64]

There have been some studies in the US and in Australia using newborn blood spots to study CP, but only to examine proteins, human DNA polymorphisms or viral RNA/DNA in frozen spots. [65] [66] [67] [68] [69] [70] No study has yet used dried newborn blood spots (uDNBS) to examine gene expression to study causal pathways to CP. We hypothesize the RNA expression patterns present in blood may provide a molecular "snapshot" into the neonatal state at the time

of delivery. Comparisons between the white blood cell transcriptome between control and affected newborns may give insights into the epidemiology of CP.

Leveraging the existing OWL study to evaluate archived uDNBS expression patterns

This is an on-going matched case-control study investigating the etiology of CP. Study subjects are recruited mainly from three regions of Michigan including Lansing, Ann Arbor, and Grand Rapids. Cases are children ages 2-15 years with the diagnosis of CP assigned by a neurologist, physiatrist or family physician, and a Gross Motor Function Classification Score ≥1 (classifying severity of CP). For access to birth certificates and blood spots, cases and controls must be born in Michigan. Age is matched and restricted to births since 1994 to reduce variation from aging of newborn blood spots. Little is known about how gender, GA, multiple birth, storage conditions, and age of stored samples affect mRNA expression in dried blood spots. Thus, to reduce systematic differences due to these factors, cases and controls (other than siblings) are matched on gender, birth year and GA in four categories (<28 weeks, 28-32 weeks, 33-37 weeks, >37 weeks).

For all cases and controls, clinical data are collected from maternal interview, birth certificate, birth hospital discharge abstract for mother and infant, clinician's referral form indicating CP type, severity and associated conditions, parent reported form indicating level of gross motor (GMFCS) [71], manual (MACS),[72] and speech function. Human mRNA and viral DNA are extracted from archived unfrozen dried newborn blood spots (uDNBS) which are obtained from the *Michigan Biotrust for Health.* [73]. Saliva samples from the child and parents are also collected to extract human DNA.

Maternal interviews (conducted via telephone) provide data on demographics, family and maternal medical history, prenatal screening tests, labor and delivery, and on the infant's

perinatal period including feeding behavior and neurologic findings. Michigan birth certificates provide data on approximately 100 pregnancy and perinatal variables. Hospital discharge abstracts provide diagnoses and procedure codes and length of hospital stay.

Laboratory techniques used in OWL study

mRNA microarray techniques

Total RNA is extracted, purified and concentrated from three 3mm punches of uDNBS. Single-stranded cDNA is generated using the WT-Ovation Pico RNA Amplification System from NuGEN Technologies. The Agilent Whole Human Genome Gene Expression 8x60K Microarray assay platform is used to profile gene expression. Each array contains 60,000 oligonucleotide probes (60bp) covering 27,958 Entrez gene RNAs and 7,419 long intergenic non-coding RNAs.

qPCR technique

Masked cDNA samples synthesized from neonatal blood spot RNA are used for qPCR analysis. For the genes selected for qPCR validation, specific optimized Taqman probes and primers were obtained from Applied Biosystems by Life Technologies (Carlsbad, CA) and qPCR was performed using Applied Biosystems 7500 Fast Real-Time PCR System.

Summary of overall procedures for analysis of mRNA microarray data

The study will have a discovery phase to identify major confounding factors that affect RNA expression in uDNBS that are not addressed by the case-control study design. Next, expression patterns present in the uDNBS will be examined to identify expression patterns that associate with the CP disease state. The flow chart of common overall procedures for analysis of mRNA microarray data is shown in figure 1.1. Briefly, main stages of analysis include raw data processing, analysis of individual genes, analysis of gene sets, assessment of the influence of

covariates on analysis results, validation of results statistically, and validation of mRNA microarray data by qPCR.

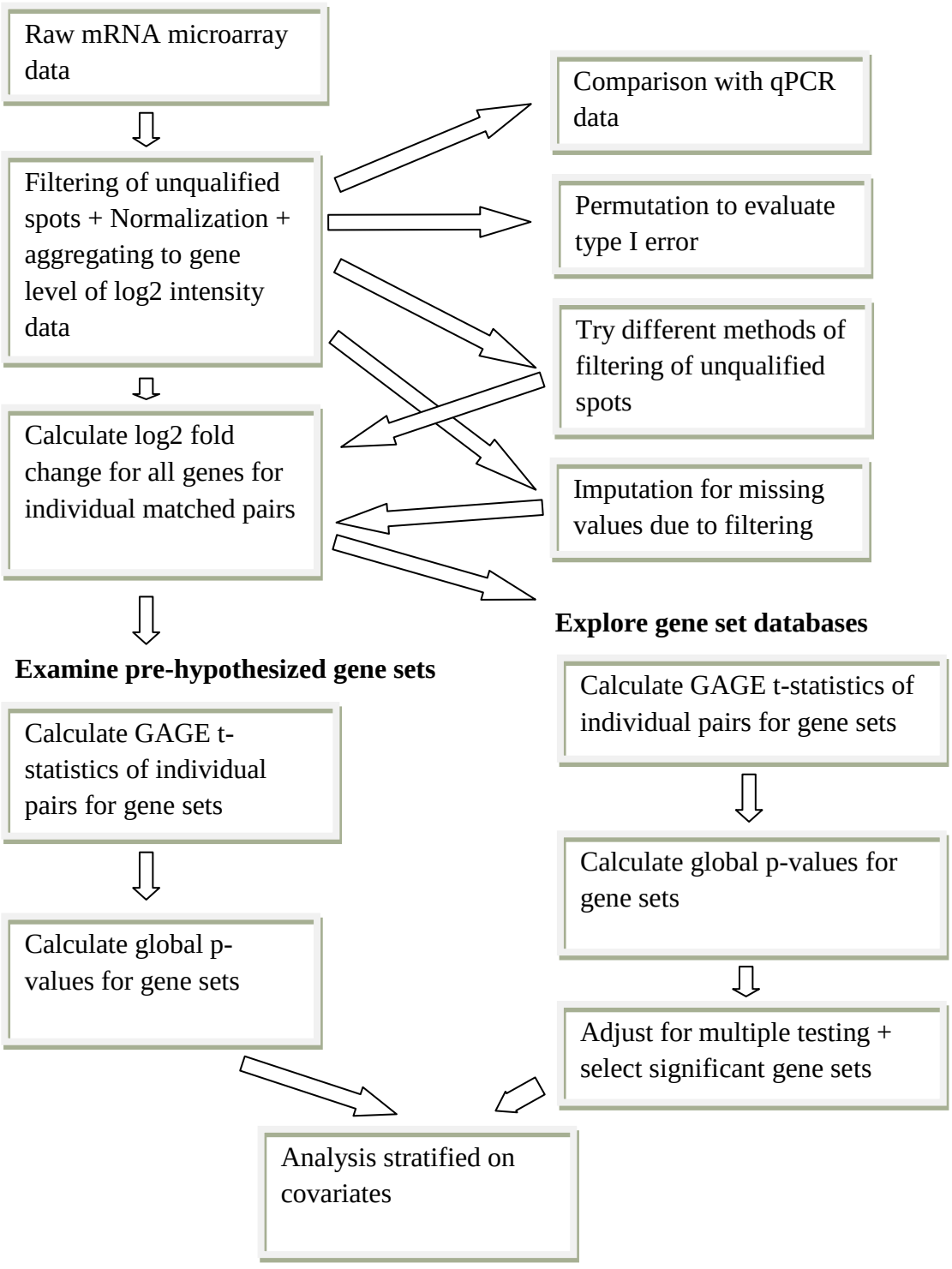

Figure 1.1. Analytic procedure.

CHAPTER 2. ANALYSIS OF DIFFERENTIAL EXPRESSION USING EXISTING METHODS

Abstract

Background: The causes of cerebral palsy (CP), the commonest major motor disability of childhood, are known to operate during pregnancy and the perinatal period, but are poorly understood. Many states archive residual filter paper blood after routine newborn genetic screening, and we have been shown that such specimens yield sufficient mRNA for gene expression profiling even after years of unfrozen storage. We thus undertook to examine causal pathways to CP by examining the newborn expression of gene sets representing potential causal pathways to CP in children with and without CP.

Methods: We selected one experimental and one curated gene set for each of three hypothesized pathways to CP (inflammatory, hypoxic and thyroidal) and one curated gene set reflecting coagulatory function for gene expression studies in unfrozen residual newborn blood spots archived by the Michigan Department of Community Health. mRNA expression of gene sets was assessed, using DNA microarray on an Agilent platform, in the newborn blood of 53 singleton children with CP and 53 control children without CP, individually matched on year of birth, gender and gestational age. These seven pre-hypothesized gene sets, and a further 205 exploratory gene sets from the Kyoto Encyclopedia of Genes and Genomes (KEGG), were analyzed for log2 fold change differences between cases and controls using the Generally Applicable Gene Set Enrichment (GAGE) Method.

Results: The empirical inflammatory and empirical hypoxic gene sets were significantly down regulated in term-born CP cases (N = 33) as compared to matched controls (p = 0.0007, 0.0009 respectively), while both pathways were significantly up-regulated (p =0.0055, 0.0223

respectively) in preterm-born CP cases (N = 20). The empirical thyroidal gene set was significantly up-regulated in preterm-born CP (p = 0.0023). Exploratory analysis of KEGG gene sets showed that the five most up-regulated gene sets (of 205) included a gene set with strong inflammatory signals (systemic lupus erythromatosus) as did the five most down-regulated gene sets (leukocyte transendothelial migration).

Conclusion: The newborn transcriptome as recorded on unfrozen archived filter paper blood spots can serve as a platform for investigating gene expression patterns in children who later develop CP or other developmental disorders. Genes of inflammatory, hypoxic and thyroidal pathways were differentially expressed in children with CP compared to matched controls, and the pattern of differential expression differed in term-born and preterm-born CP cases. Inflammatory processes operative during the peri-partum period appear to play an important role in the development of CP in both term and preterm infants.

Introduction

CP is a severe childhood neurological disorder that is characterized by impaired control of motor function and some other associated deficits such as mental retardation, epilepsy, learning disorders, visual and hearing impairment.[74] CP occurs in 1-3 infants for every 1000 live births making it one of the most common of the disabling childhood disorders.[75] The prevalence of CP has not been reduced over the past several decades. People with CP often require special and costly medical and educational services for their whole life. This makes CP an important health care burden for any country.

The etiology of cerebral palsy is not well known. No study has yet examined gene expression from newborn blood spots to study the patho-physiological pathways leading to cerebral palsy (CP). Previous pilot work from our team has shown that mRNA extracted from

long-term unfrozen storage of archived unfrozen dried neonatal blood spots (uDNBS) leftover from newborn genetic screening are sufficient to perform mRNA microarray analysis for genome-wide expression profiling. [76]

Our on-going matched case-control study investigating the etiology of cerebral palsy (CP) in children aged from 2-16 years uses uDNBS. The Agilent Whole Human Genome Gene Expression 8x 60 K Microarray platform is being used for the mRNA microarray assays. Clinical data are obtained from maternal interviews, birth certificates and maternal and newborn hospital discharge abstracts. Cases and controls are matched by age, sex, and gestational age in four categories.

The analysis of gene expression for groups of genes (gene sets) is employed to examine the co-effect of genes in pathways leading to the development of CP. Gene set analysis of pre-selected gene sets are done to specifically investigate the roles of the four pre-hypothesized pathways (including inflammation, thyroid disorders, hypoxia/asphyxia, and coagulation disorders). To discover other possible pathways contributing the development of CP, gene sets from gene set databases (such as KEGG) are explored. Clinical exposures during the prenatal and peri-partum period may play important roles in altering gene expression profiles. Thus, analysis integrating gene expression data and important clinical covariates is also drawn on to study the role of clinical exposures and gene expression patterns to the development of CP.

We anticipate that the findings from the data in this study will help produce a more complete understanding of the etiology of CP. CP is a condition that largely develops during pregnancy and the peri-partum period. Thus, understanding the exposures during pregnancy and peripartum period is essential to understanding potential causal factors for CP.

Methods

The mRNA microarray data and clinical data in this chapter is from archived unfrozen dried blood spots of 53 cerebral palsy case - matched control pairs of an on-going case-control study investigating the etiology of cerebral palsy. All 106 study subjects are singletons, 31 females and 75 males, aged from 2.9-16 years. CP cases and controls are matched by year of birth, and gestational ages by 4 categories (<28 weeks, 28-32 weeks, 33-37 weeks, >37 weeks). Microarray assays of Agilent platform were used to profile mRNA extracted from blood spots. Real time reverse transcriptase quantitative polymerase chain reaction (qPCR) were used to validate the expression of some selected genes. More details about this study, laboratory techniques and data from this study are described in previous chapters.

Statistical methods

All procedures for data processing and analysis were done using statistical software R (version 2.13.2). Unqualified spots were filtered using the method of Patterson et al (expression data were removed wherever gProcessed signal was less than twice the gProcessed signal error).[77] Gene expression data were normalized using a between-array quantile normalization method,[78] and further aggregated to the gene level using the mean of the expression signal of all probes of each gene. Differential expression of individual genes was examined with the moderated paired t-test (which is appropriate for matched pairs) of the linear model and an empirical Bayes method implemented in R package *limma*.[79] [80] The significance of gene expression was corrected for multiple testing using the false discovery rate (FDR) approach.[81] [82] Among several methods for gene set analysis,[83] [84] we chose the Generally Applicable Gene

Set Enrichment (GAGE) method, probably the only published method specifically applicable to a matched case-control study. [85]

Briefly, the GAGE method conducts a two-sample-like *t*-test to compare the expression of genes in the gene set of interest to the expression of all genes measured on the array of each matched pair:

$$t_{kl} = (m - M) / \sqrt{\frac{s^2}{n} + \frac{S^2}{n}}$$

where *m* and *M* are the mean log fold change of genes in the set and all genes in the array, respectively. s and S are the standard deviation of the log fold change of genes in the set and that of all genes in the array, respectively. *n* is the number of genes in the set. The p-values of the individual within-pair *t*-tests are further summarized using a meta-test for global significance:

$$\mathrm{x} = -\frac{1}{L}\sum_{kl} \log \mathrm{P_{kl}}$$

which follows *Gamma* (*k*,*l*). This method allows calculating a test statistic and p-value for assessing differential expression of the gene set for each individual matched pair before summarizing all pairs for global significance. Heterogeneity in differential expression among matched pairs can also be examined. The global test further allows detection of the significance of a group of pairs with small p-values, which is important when some pairs are differentially up-regulated while some others are down regulated. This is particularly relevant in a disorder such as CP, which is composed of sub-types likely to have distinct etiologies. Genes within a gene set may also be regulated in different directions (up-regulated versus down-regulated).The use of the

absolute value of the log2 fold change of genes avoids the cancelation of the significance of genes in different direction of regulation, where a upper one-sided test is appropriate to assess the significance of gene regulation away from the normal expression.

The analysis of gene sets were performed on pre-selected gene sets of the four pre-hypothesized pathways and on gene sets of the Kyoto Encyclopedia of Genes and Genomes. For each of the four pathways, one empirical gene set (genes indentified from experiments) and one canonical gene set (genes derived from expert opinion) are selected. The empirical inflammation gene set includes the genes differentially expressed in cord blood of prematures with and without markers of fetal inflammation.[86] The canonical inflammation gene set GO:005072; inflammatory response, and the canonical coagulation gene set, GO:0007596; blood clotting biological process, are obtained from the Gene Ontology (GO) database. No empirical gene set for coagulation is found. The empirical asphyxial gene set is derived from the experiments on responses of cells in tissue culture exposed to hypoxemia compared to normoxemia.[87] The canonical asphyxial gene set is based on the view that *hypoxia-inducible transcription factor (HIF)* binds a consensus DNA sequence termed the *hypoxia-responsive element* (HRE).[88] The canonical thyroid responsive element (TRE) gene set is also assembled using a similar approach.[89] The experimentally derived gene set was isolated following human exposure to thyroid hormone.[90]

Analysis of imputed data

The processed microarray dataset contains large percentage of missing values (20-70% for each array) after filtering unqualified spots. For simplicity, missing values were replaced by a value that approximate the smallest expression value of the remaining expression data after filtering (smallest log2 intensity was approximately 5 in this situation). The simple assumption

was that all missing values produced by filtering are low expression and are equally treated. This is a way to check for the robustness of the results. If the results of gene set analysis of imputed data (based on the above assumptions) are similar to those of the data with missing values, the results of gene set analysis may be robust (not sensitive to the effect of missing values). In other words, the purpose is to evaluate whether the loss of expression information due to missing values caused by filtering of unqualified spots, especially when taking fold change can distort the results or not, or to check whether the effect of missing values is too large to produce misleading or non-robust results for gene set analysis.

Results

Analysis of seven pre-selected gene sets representing four pre-hypothesized pathways to CP

Three of the seven gene sets, all empirical, showed significantly different regulation between cases and matched controls after adjusting for multiple testing when false discovery rate (FDR) was set at 0.05 (Table 2.1). The empirical inflammatory and asphyxial gene sets were both significantly down-regulated in CP cases, compared to controls (q-value is 0.0008 and 0.0059 respectively and approximate effect size is -0.19 SD units and -0.16 SD units respectively), while the thyroidal gene set was significantly up-regulated (q-value is 0.0273 and the approximate effect size is 0.13 SD units). For the empirical inflammatory and asphyxial gene sets, the global P-values for up-regulation reached marginally significance (p-value=0.0791 and 0.0983) but no longer significant after adjusting for multiple testing (q-value >0.1).

To describe the extent of gene set differences between individual case and control pairs, and the degree of heterogeneity of these differences across pairs, the distribution of GAGE- t-statistics for 7 gene sets, representing the difference in gene expression between cases and controls, is shown in Figure 2.1. The three significant gene sets (experimental inflammatory,

asphyxial and thyroidal) show many pairs with large differences in gene expression, while the canonical asphyxial and thyroidal gene sets show modest inter-pair differences. For the coagulation gene set and the canonical inflammatory gene set, virtually no differences within pairs are seen. Heterogeneity across pairs in differential expression is notable for all three significant gene sets. Pairs show large differences in either up or down regulation.

The Fetal Inflammatory Response Syndrome (FIRS) Gene Set

Among the hypothesized gene sets, the largest case-control differences in gene expression were seen for the FIRS gene set. Figure 2.2 shows the heatmap of the log2 fold change of all genes in FIRS gene set for all case-matched control pairs ordered by the magnitude of the GAGE t-statistics. Positive t-statistics (up-regulation of case compared to control) are seen as red, negative t-statistics (down regulation of case compared to control) are seen as blue. The heatmap also shows that a considerable percentage of genes in FIRS gene set have missing values (grey color). The up or down regulation of the FIRS gene set in individual pairs seem to be driven especially by the following genes: *S100A9* (S100 calcium binding protein A9), *S100A12* (S100 calcium binding protein A12), *ALOX5AP* (arachidonate 5-lipoxygenase-activating protein), *PGLYRP1* (peptidoglycan recognition protein 1), *HP* (haptoglobin), *FLOT1* (flotillin 1), and *FGR* (Gardner-Rasheed feline sarcoma viral oncogene homolog).

Table 2.1. Gamma GAGE analysis for seven gene sets representing four pre-hypothesized pathways.

Gene sets	Reference to the source of the gene set	Mean of GAGE *t*-statistics (SD units)$	*P*-values (q-values)# for up-regulation	*P*-values (q-values) for down# regulation	*P*-values (q-values)# for Bi-directional regulation
Coagulative					
Canonical (*n*=93; ne=92)	GO:0007596; blood clotting	-0.08	0.9749 (>0.1)	0.7737 (>0.1)	0.6048 (>0.1)
Inflammatory					
canonical (*n*=31; ne=31)	GO:0050727; regulation of inflammatory response	-0.10	0.9870 (>0.1)	0.7796 (>0.1)	0.9439 (>0.1)
empirical (*n*=36; ne=36)	Fetal inflammatory response	-0.19	0.0791 (>0.1)	**0.0001 (0.0008)**	0.2139 (>0.1)
Asphyxial					
canonical (*n*=37; ne=36)	Hypoxia	0.18	0.1656 (>0.1)	0.9401 (>0.1)	0.9620 (>0.1)
empirical (*n*=127; ne=126)	Hypoxia	-0.16	0.0983 (>0.1)	**0.0016 (0.0059)**	0.9749 (>0.1)
Thyroidal					
canonical (*n*=200; ne=198)	V$T3R_Q6; TRE consensus	-0.03	0.8183 (>0.1)	0.7344 (>0.1)	0.9993 (>0.1)
empirical (*n*=140; ne=139)	Thyroid hormone	0.13	**0.0039 (0.0273)**	0.2873 (>0.1)	0.9767(>0.1)

$ Since the number of genes in all seven selected gene sets is >30, the GAGE t-statistic approximates a z-statistic. Thus, the mean of GAGE t-statistics, which can be expressed in terms of standard deviation (SD) units, is an approximation of effect size.
# global *P* values of all pairs, with *q*-values in parentheses calculated to adjust for multiple testing using *q*-values R package with FDR set at 0.05. n Number of genes in the gene set. ne Number of genes of the gene set that are found in the array used in this study.

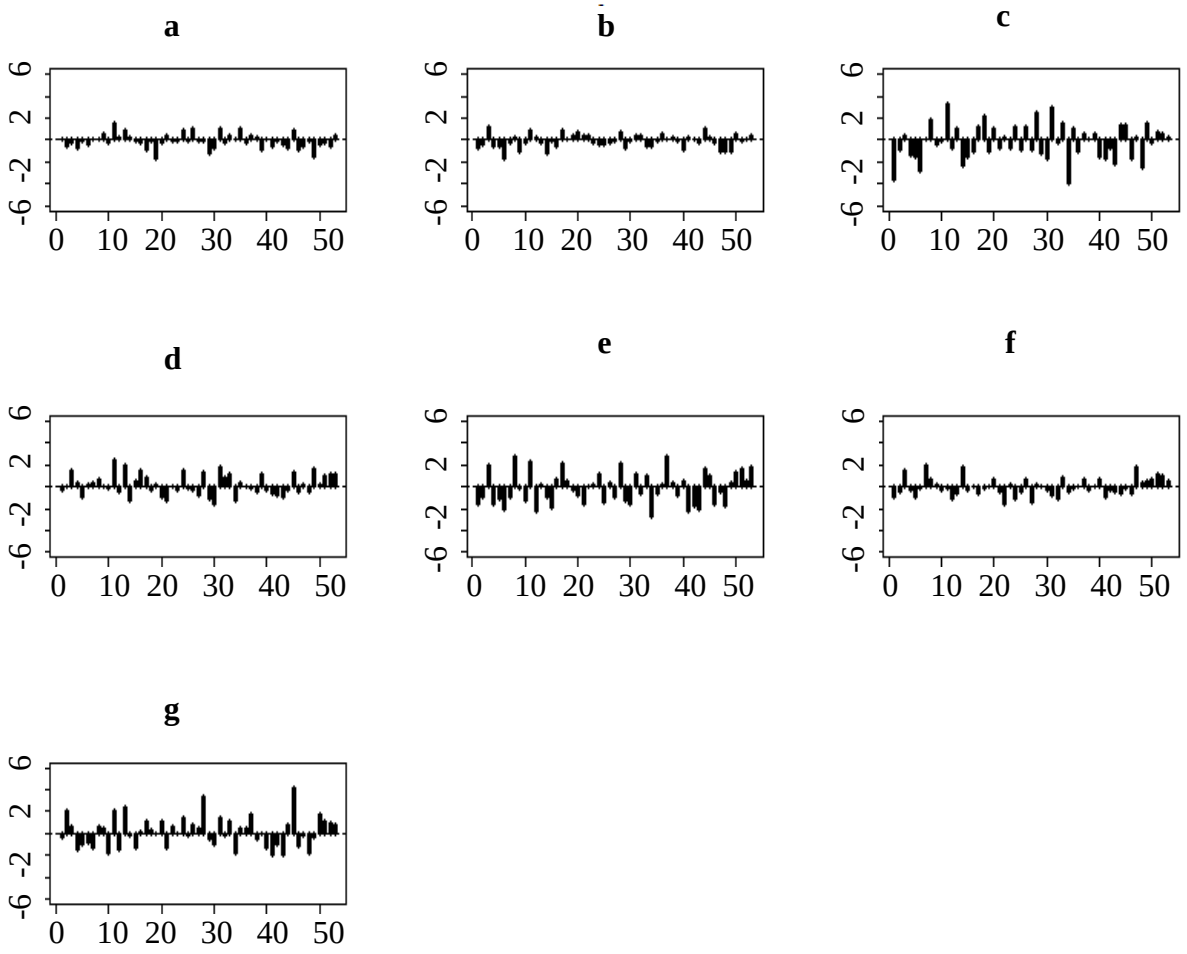

Figure 2.1. GAGE t-statistics for the seven pre-hypothesized gene sets.
a, b, c, d, e, f, g: canonical coagulation, canonical inflammation, empirical inflammation, canonical asphyxia, empirical asphyxia, canonical thyroid, empirical thyroid gene sets respectively. For each graph: X-axis: matched pair (total 53 pairs); Y-axis: scale of GAGE t-statistic; each bar within each graph: the GAGE t-statistic of the gene set for each pair.

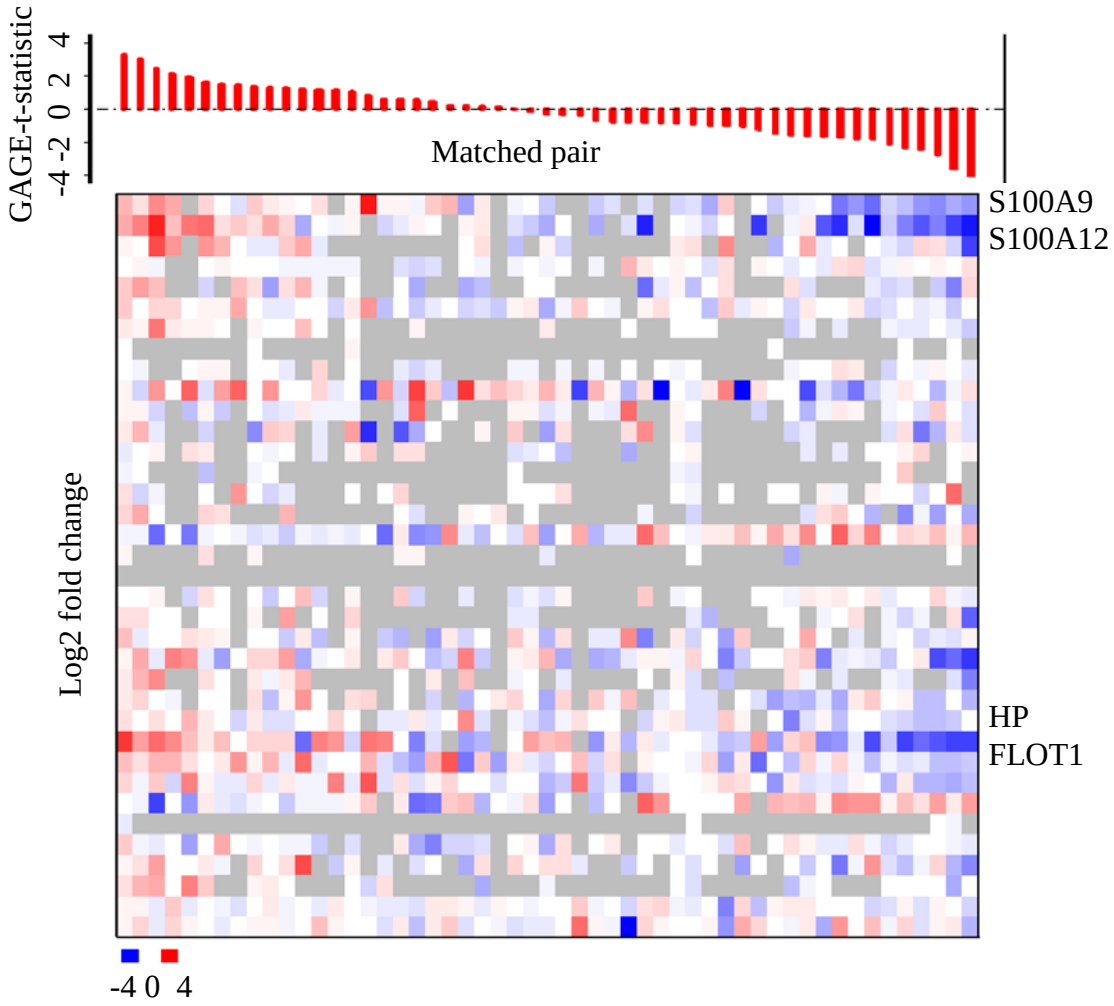

Figure 2.2. Heatmap of FIRS gene set with pairs ordered by magnitude of GAGE t-statistics.

(a) FIRS gene set in which the matched pairs are ordered by the values of the GAGE t-statistics of the pairs from most positive to most negative.

(b) Heatmap: X-axis: matched pairs in the same order as the upper graph; Y-axis: gene names. Each small square represents log2 fold change of each of all genes of FIRS gene set of each of all pairs. Gradient scale for color from bluest (most negative log2 fold change or the gene expresses lowest in case vs. control) to white (log2 fold change is zero or the gene expresses equally in case vs. control) to reddest (positive log2 fold change or the gene expresses highest in case vs. control): – 4 to 0 to + 4. Grey color: absence of data (missing values) due to unmet filtering criteria.

(For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation).

Table 2.2. Gene expression findings for three gene sets stratified by GA and CP type.

Gene sets	**Empirical inflammatory gene set**			**Empirical asphyxial gene set**			**Empirical thyroidal gene set**		
	Mean GAGE *t*-stat	*P*-values up	*P*-values down	Mean GAGE *t*-stat	*P*-values up	*P*-values down	Mean GAGE *t*-stat	*P*-values up	*P*-values down
Gestational age									
>=37 weeks (n=33)	-0.42	0.6567	0.00007	-0.36	0.5014	0.00089	-0.04	0.1345	0.1204
<37 weeks (n=20)	0.19	0.0055	0.1537	0.16	0.0223	0.2381	0.42	0.0023	0.7353
CP type$									
Quadriplegia (n=24)	-0.26	0.60869	0.02830	-0.18	0.11721	0.01024	0.08	0.01865	0.16745
Diplegia (n=15)	0.16	0.03753	0.23559	0.03	0.27462	0.30122	0.32	0.10591	0.84832
Hemiplegia (n=13)	-0.45	0.11210	0.00009	-0.32	0.30896	0.01476	0.01	0.09332	0.19450

$CP type missing for 1 case

Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets

We explored case-control differences for the 205 gene sets archived by KEGG (2009 version). The five most up-regulated gene sets in CP cases compared to controls were: Ribosome, Systemic lupus erythematosus (SLE), Olfactory transduction (OT), Cell cycle (CC) and Oxidative phosphorylation (OP). Down-regulation was seen most strongly for three of the above five - Ribosome, CC and SLE, and for Leukocyte transendothelial migration (LTM) and Regulation of actin cytoskeleton (RAC) gene sets.

Using the approach of Storey to control for the false discovery rate (FDR), Ribosome, SLE and OT remained significantly up-regulated, Ribosome, LTM and RAC remained significantly down-regulated, and the Ribosome gene set was significantly bi-directionally regulated (table 2.3). The heterogeneity of individual pair contrasts in the three gene sets significantly perturbed in both directions can be seen in Figure 2.3.

Table 2.3. Most up- regulated and most down-regulated KEGG gene sets in cases compared to controls.

Most up-regulated gene sets			**Most down-regulated gene sets**			**Most bi-directionally regulated gene sets**		
Gene sets	*P*- values	*q*-values$	Gene sets	*P*-values	*q*-values$	Gene sets	*P*-values	*q*-values$
Ribosome	4.4e-40	**9.1e-38**	Ribosome	4.1e-42	**8.3e-40**	Ribosome	1.9e-14	**3.9e-12**
SLE	1.7e-06	**1.8e-04**	LTM	4.2e-04	**3.5e-02**	BTF	3.3e-03	>0.1
OT	5.8e-05	**3.9e-03**	RAC	5.1e-04	**3.5e-02**	APM	4.9e-02	>0.1
CC	2.4e-03	>0.1	CC	1.8e-03	9.5e-02	PD	5.0e-02	>0.1
OP	5.6e-03	>0.1	SLE	9.2e-03	>0.1	MODY	5.1e-02	>0.1

$ *q*-values were calculated using *q*-values R package with FDR set at 0.05. SLE: Systemic lupus erythematosus, OT: Olfactory transduction, CC: Cell cycle, OP: Oxidative phosphorylation, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, BTF: Basal transcription factors, APM: Aminophosphonate metabolism, PD: Parkinson's disease, MODY: Maturity onset diabetes of the young.

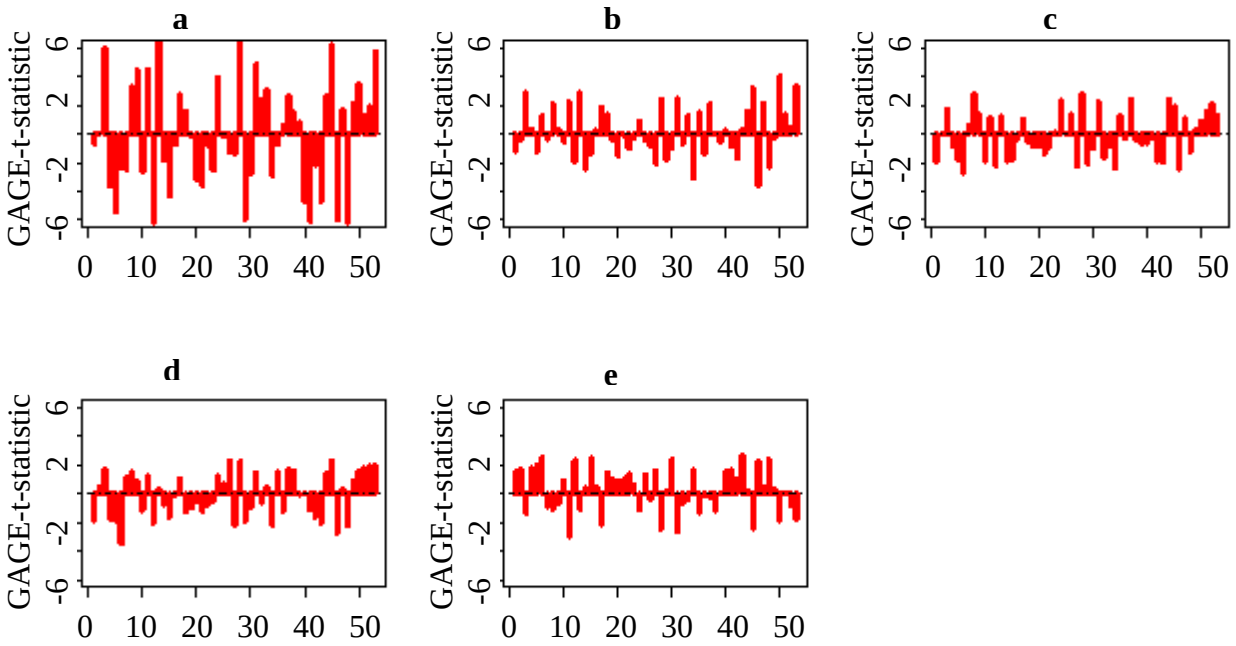

Figure 2.3. GAGE-t-statistics of five most significant KEGG gene sets.
a, b, c, d, e respectively: Ribosome, SLE: Systemic Lupus Erythematosus, LTM: Leukocyte Transendothelial Migration, RAC: Regulation of Actin Cytoskeleton, OT: Olfactory Traduction.

Assessing Significant Gene Set Findings in CP sub-sets

The heterogeneity of GAGE *t*-statistics across pairs suggests the need to stratify on covariates such as gestational age and motor type. Table 2.2 shows the findings for each gene set for children born at 37 weeks or later (N pairs = 33), before 37 weeks (N pairs= 20), with quadriplegia (N pairs = 24), diplegia (N pairs = 15) and hemiplegia (N pairs = 13).

FIRS shows an interaction with gestational age; among 20 premature pairs, FIRS was significantly up-regulated in CP cases, whereas among 33 term-born pairs, FIRS was significantly down-regulated among cases. The FIRS up-regulation seen in premature cases was paralleled by up-regulation in diplegic cases, who are dominantly premature. In parallel, the strongest contribution to down-regulation of inflammation came from hemiplegic cases (N = 13) who are nearly all born at term. Quadriplegia also showed down regulation of FIRS, but not as strongly as hemiplegia.

The empirical asphyxia gene set also showed significant up-regulation in premature cases and the opposite with term cases. As in the case of FIRS, the down-regulated signal was stronger than the up-regulated signal, reflecting the larger number of term-born cases in our sample. Hemiplegia and quadriplegia showed down-regulation of the asphyxial gene set in about equal measure. The thyroidal up-regulation signal was derived entirely from prematures and from children with quadriplegia.

After adjusting for multiple testing, the ribosome gene set is significantly up- and down-regulated in both term and preterm CP cases compared to controls (q-values $<<0.0001$) (Table 2.4). The SLE, OP and Pathogenic Escherichia coli infection (PECI) gene sets are significantly up-regulated in preterm born CP (q-values <0.05) only. OT, on the other hand, is significantly up-regulated in term CP (q-values <0.05). After adjusting for multiple testing, the ribosome gene

set is significantly up- and down- regulated in all 3 CP types (q-values <0.0001). SLE is significantly up regulated in diplegic CP (q-value <0.05) (table 2.5).

Analysis of individual genes

The analysis for all individual genes available in the arrays reveals that no individual gene was significantly differentially expressed between cases and controls, after adjusting for multiple testing. The lack of single gene expression differences confirms the value of gene set analysis for aggregating coordinated expression signals from related genes in gene sets in exploring pathophysiological pathways to disease.

Table 2.4. Results of KEGG gene sets stratified by GA.

	GA>=37wks				GA<37wks			
	Top up regulated gene sets		Top down regulated gene sets		Top up regulated gene sets		Top down regulated gene sets	
Gene sets	P-values	q-values	P-values	q-values	P-values	q-values	P-values	q-values
Ribosome	3.2e-18	6.6e-16	1.3e-27	2.7e-25	1.2e-24	2.5e-22	1.1e-16	2.2e-14
SLE	9.2e-03	>0.1	6.5e-03	>0.1	6.5e-06	6.6e-04	>0.05	>0.1
OT	2.2e-04	2.2e-02	>0.05	>0.1	3.6e-02	>0.1	9.2e-03	>0.1
CC	3.2e-02	>0.1	5.4e-03	>0.1	1.2e-02	>0.1	>0.05	>0.1
OP	>0.05	>0.1	>0.05	>0.1	1.2e-03	4.8e-02	>0.05	>0.1
LTM	>0.05	>0.1	8.3e-04	5.6e-02	4.1e-03	>0.1	>0.05	>0.1
RAC	>0.05	>0.1	6.4e-04	5.6e-02	1.6e-02	>0.1	>0.05	>0.1
Apoptosis	>0.05	>0.1	5.7e-03	>0.1	3.7e-02	>0.1	>0.05	>0.1
GHD	>0.05	>0.1	6.4e-03	>0.1	>0.05	>0.1	>0.05	>0.1
PECI	>0.05	>0.1	>0.05	>0.1	1.1e-03	4.8e-02	>0.05	>0.1

SLE: Systemic lupus erythematosus, OT: Olfactory transduction, CC: Cell cycle, OP: Oxidative phosphorylation, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, GHD: Graft-versus-host disease, PECI: Pathogenic Escherichia coli infection.

Table 2.5. Results of KEGG gene sets stratified by CP type.

Gene sets	Quadriplegia (n=24)				Diplegia (n=15)				Hemiplegia (n=13)			
	P-values up	q-value	P-values down	q-value	P-values up	q-value	P-values down	q-value	P-values up	q-value	P-values down	q-value
Ribosome	1.6e-22	3.4e-20	2.2e-28	4.6e-26	8.2e-10	1.6e-07	5.1e-08	1.1e-05	4.3e-12	8.9e-10	1.8e-10	3e-08
SLE	4.4e-03	>0.1	4.6e-02	>0.1	1.9e-04	1.9e-02	>0.05	>0.1	2.4e-02	>0.1	4.3e-03	>0.1
OT	1.3e-03	>0.1	>0.05	>0.1	1.2e-02	>0.1	>0.05	>0.1	>0.05	>0.1	>0.05	>0.1
LTM	1.1e-02	>0.1	4.1e-02	>0.1	>0.05	>0.1	>0.05	>0.1	>0.05	>0.1	4.5e-03	>0.1
RAC	>0.05	>0.1	2.4e-02	>0.1	>0.05	>0.1	2.0e-02	>0.1	>0.05	>0.1	3.2e-02	>0.1
CC	>0.05	>0.1	3.1e-02	>0.1	>0.05	>0.1	>0.05	>0.1	2.7e-02	>0.1	2.4e-02	>0.1
OP	>0.05	>0.1	>0.05	>0.1	3.7e-02	>0.1	>0.05	>0.1	6.9e-03	>0.1	>0.05	>0.1

Analysis of imputed data

Briefly, for the 7 preselected gene sets, results are similar to those of non-imputed data and the similarity is clearer when stratifying on GA (table 2.6). For the KEGG gene sets, the top up regulated gene sets are the same and in the same order as those of the dataset with missing values. The results of the top down regulated KEGG gene sets are also similar to those of the dataset with missing values, except for a slight difference in order, and one additional down regulated gene set (UMP) (table 2.7). Thus imputing values to missing data did not materially change our results on gene set analysis.

Table 2.6. Results of 7 pre-selected gene sets from dataset with imputation for missing values.

	All GA		GA>=37wks		GA<37wks	
Gene sets	P values UP	P-values DOWN	P values UP	P-values DOWN	P values UP	P-values DOWN
Inflammatory						
canonical	0.95413	0.75982	0.98644	0.45653	0.49302	0.90513
empirical	0.04142	**0.000014**	0.52683	**0.000009**	**0.00377**	0.11562
Thyroidal						
canonical	0.95934	0.83449	0.81971	0.88131	0.94878	0.51051
empirical	**0.00038**	0.12671	0.11309	0.88424	**0.00011**	0.43578
Asphyxial						
canonical	0.42742	0.88891	0.51491	0.86248	0.34002	0.69644
empirical	**0.00478**	0.017	0.09259	**0.02492**	**0.00614**	0.16318
Coagulative						
canonical	0.94373	0.91647	0.95229	0.89563	0.65688	0.71440

Table 2.7. Results of gene set analysis of KEGG gene sets from dataset with imputation for missing values.

Top up regulated gene sets		Top down regulated gene sets	
Gene sets	P-values	Gene sets	P-values
Ribosome	4.0e-47	Ribosome	2.5e-46
SLE	3.7e-08	RAC	5.6e-04
OT	2.3e-05	SLE	8.1e-04
OP	2.4e-03	UMP	6.0e-03
CC	5.8e-03	LTM	6.1e-03
		CC	8.7e-03

SLE: Systemic lupus erythematosus, OT: Olfactory transduction, CC: Cell cycle, OP: Oxidative phosphorylation, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, UMP: ubiquitin mediated proteolysis.

qPCR validation of mRNA data

To validate our microarray findings, we used qPCR techniques to examine the housekeeping genes, *ACTB* (beta actin) and *PPIA* (peptidylprolyl isomerase A), both commonly used in the literature to validate microarray findings. To validate genes differentially expressed, we selected *FCGR2A* (Fc fragment of IgG, low affinity IIa receptor), a representative gene of the Lupus pathway, which was among the most significantly differentially regulated gene sets in the KEGG database.

For *ACTB* and *PPIA* genes, the correlation coefficient between the $\log_2$ intensity of microarray data and mean CT (cycle threshold value) of qPCR was –0.52 ($P < 0.0001$) (figure 2.4) For *FCGR2A*, the correlation coefficient between the $\log_2$ intensity of microarray data and mean CT of qPCR is –0.43 ($P < 0.0001$) and the correlation coefficient between $\log_2$ fold change of microarray data and $\log_2$ fold change of qPCR data was 0.38 ($P = 0.0197$) (figure 2.5).

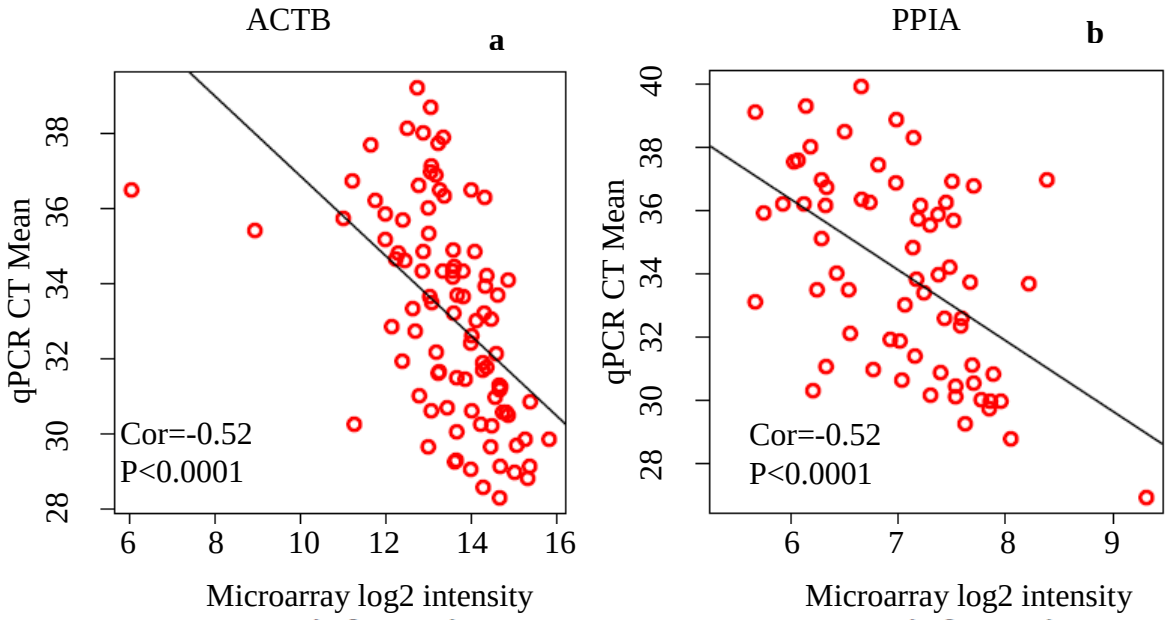

Figure 2.4. Correlation between qPCR and microarray data of housekeeping genes.
a: ACTB gene, b: PPIA gene. Cor: correlation coefficient.

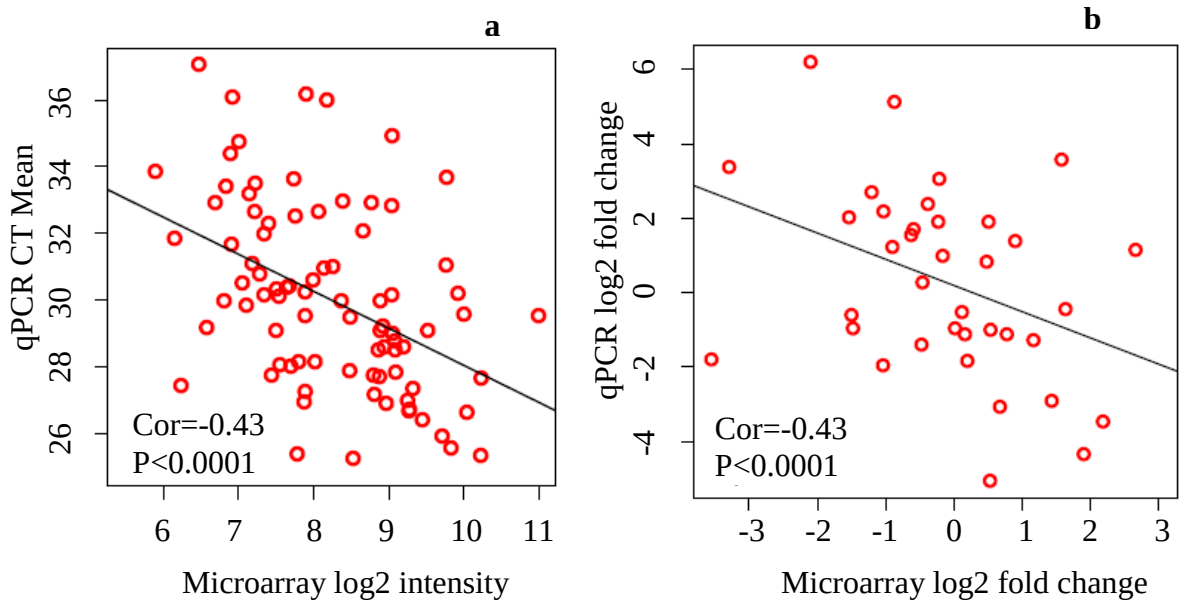

Figure 2.5. Correlation between qPCR and microarray data of representative gene (FCGR2A) of SLE gene set.
a: log2 expression, b: log2 fold change. Cor: correlation coefficient.

Discussion

uDNBS are available in many states and may be a rich resource for epidemiological research. [91] If gene expression profiling, especially via gene set analysis, can reveal potential risk factors or causal pathways to CP, we may set a scientific precedent for future studies using gene expression profiling from uDNBS to investigate the etiology of other disorders potentially of perinatal origin such as autism, etc.

Apart from a few genetic risk factors that we do not focus on, many environmental or behavioral exposures or risk factors may be altered, treated or avoidable. Most of the factors contributing to the inflammatory, thyroidal and hypoxic/asphyxial pathways may be averted. Thus, if these pathways are among important causal pathways of CP, there should be a hope for the reduction of CP prevalence. In other words, understanding causal pathways of cerebral palsy via gene expression and clinical data may help develop strategies to prevent CP in the future and thus may help reduce lifetime health care burden caused by CP.

The analysis of differential expression has provided some interesting results. The FIRS (empirical inflammatory gene set) contains the genes that are up-regulated in preterm newborns. These genes are also up-regulated in preterm born CP cases vs. controls. This suggests that CP may share some mechanism or pathological processes with FIRS or FIRS is a part of the path leading to CP in preterm. However, these same genes are more down regulated in term born CP. It is hard to speculate the pathological mechanism but this may suggest that there may be a difference in mechanisms leading to CP between term and preterm CP. For empirical hypoxia gene set, the pattern of differential expression is similar to that of FIRS gene set in that it is more down regulated in term and slightly more up regulated in born CP. Thus, the genes in this gene set may be regulated similarly to the genes in FIRS gene set. The empirical thyroidal gene set is

only up regulated in preterm born CP. This may be linked to the fact that preterm newborns often have thyroid hormone disorder.

The significant differential expression of the three above empirical gene sets represent inflammatory, thyroidal and hypoxic pathways suggests that there is a possible coordination of genes in FIRS, hypoxic and thyroidal pathways in developing CP. In other words, many pathways may coordinate to produce CP, not just a single pathway.

In addition, there may be an important effect of GA on differential expression patterns of the above three gene sets. Thus, pathological processes leading to CP may be different in term and preterm newborns. The effect of GA is biological and patho-phyiological meaningful and may match with current understanding of CP in preterm-born CP. The findings on term-born CP is quite new and may lead to new hypothesis about difference in mechanism or causal pathways to CP in term and preterm born CP.

Among KEGG gene sets, several gene sets (SLE, LTM) related to inflammatory processes are significantly either up or down regulated. Lupus gene set contains genes related to multiple inflammatory processes such as T cell and B cell receptor signaling pathways, LTM, Jak-STAT signaling pathway, cytokine-cytokine receptor interaction, complement and coagulation cascades, Apoptosis, tissue injury and damage, etc. RAC largely overlap with LTM gene set and thus, the significant result of RAC may be due to overlapping genes with LTM.

The pattern of differential expression of SLE and LTM are similar to FIRS in that they are more up-regulated in preterm born CP and more down regulated in term born CP. Several gene sets related to inflammation are more up-regulated in preterm born CP (SLE, LTM, Apoptosis, PECI) and more down-regulated in term born CP (SLE, LTM, Apoptosis, GHD). However, only SLE is significant (down regulation) after adjusting for multiple testing. The

results of KEGG supplement and strengthen the results of 7 selected gene sets and indicate that different processes related to inflammation in peri-partum period may play important role in developing CP.

Methodologically, a broad and deep analysis strategy has been employed for the analysis of differential expression. First, although analysis of individual genes reveals no significance gene, the analysis aggregating genes as gene sets produces important and interesting results. Thus, gene set or pathway analysis is essential in investigating pathways to CP. Second, examining pre-selected gene sets for the four pre-hypothesized pathways helps focus on the most potential pathways to CP and also reduces the need for adjusting for multiple testing. Extensively exploration of other gene sets from KEGG database help fully examine and discover other possible pathways contributing to the development of CP. Third, we performed the test for both uni-directional gene sets (genes in the set are regulated in the same direction) and bi-directional gene sets (genes in the set are regulated in different directions). Fourth, the global test of GAGE based on gamma distribution helps detect differential expression across pairs either homogeneously or heterogeneously. Fifth, for a given gene set, a significant global gamma test in a direction of differential expression (up or down) indicates that there are at least some pairs in the samples significantly differentially expressed in that direction. Thanks to this characteristic, the global gamma test of GAGE help detect heterogeneity in differential expression across matched pairs (some pairs are up regulated and some other pairs are down regulated). Other available gene set analysis methods are incapable of detecting differential expression when there is large heterogeneity. Our samples are clinically heterogeneous in gestational age, CP types. Thus, the GAGE method is appropriate to use to screen for differential expression either homogeneously or heterogeneously.

Furthermore, technically, the quality of our microarray data is validated via the acceptably good correlation with qPCR data, a so called “gold standard”. Statistically, repeating the analysis on the dataset with imputation for missing values can evaluate how missing values due to filtering of unqualified spots affect the results of the analysis of differential expression. The results of both qPCR data and analysis of imputed data indicate that the quality of microarray data is acceptably good and the results of gene set analysis using the GAGE method on our microarray data is relatively robust.

Analysis of differential expression in term of gene sets: methodological issues

The interpretation of the results should take into account the fact that the global gamma test of GAGE may be sensitive to extremely small p-values of a few pairs. It is unknown about type I error or false positive rate of the GAGE method. Thus, simulation and permutation on randomly selected gene sets of different sizes on actual microarray data is necessary to evaluate type I error of the GAGE method.

In addition, the GAGE method may also have some limitations. Although it is probably the only published gene set analysis method applicable to matched data, this method is based on fold change of gene expression values between case vs. matched control for each matched pair. While fold change may be appropriate for matched paired data, it is a relative measurement of difference which is sensitive to small values. Thus, inference from hypothesis testing must be cautious. In addition, except for stratified analysis, the GAGE method is not capable of evaluating the interaction or confounding effects of multiple covariates, especially continuous covariates. In clinical or epidemiological studies, many covariates may need to be taken into account when interpreting of gene set analysis results. Thus, there should be a need to develop

more rigorous approaches for gene set analysis to handle different aspects of microarray data in clinical epidemiological context.

Besides, several gene sets or pathways may have similar effects or may cooperate in disease development. There is also overlap in genes between gene sets that may affect the results and interpretation of the results when examining multiple gene sets. Therefore, a statistical method that can describe and test for the correlation or co-effect of gene sets as well as the influence of overlapping genes between gene sets is necessary to understand and quantify the relationship between gene sets or pathways.

Data processing and validation: methodological issues

First, some loss of mRNA, especially in older unfrozen blood spots, has been documented, and may affect the quality or quantity of mRNA used for gene expression profiling and thus effect the quality of microarray data generated from those mRNA samples. In older samples, with lower microarray data quality, the expression signal detected from microarray data may be too weak in all phenotypes for the genes of interest to be shown to be differentially expressed. Therefore, it is important to systematically explore the patterns of mRNA microarray quality which may indirectly reflect mRNA degradation or the amount of extractable mRNA with respect to age of NBS samples. It is also essential to evaluate the influence of mRNA microarray data quality on the possibility of detecting differential expression.

Second, a good normalization method should be able to account for systematic variations across arrays due to systematic variations in biological or clinical or technical characteristics across mRNA samples. The conventional quantile normalization method that we are using and other available published normalization methods can barely address this issue. Thus, an approach that can better address the trend of mRNA microarray quality over time as well as the effects of

other covariates on the distribution of detected expression signals across samples needs to be developed to process microarray data like ours.

Third, when using qPCR to validate mRNA microarray data, housekeeping genes are often needed to accurately quantify RNA. Housekeeping genes are often selected as genes with similar expression level in microarray data across samples. However, in old mRNA samples, the level of gene expression signal detected may decrease across samples over time. As a result, it may not be appropriate to apply the same criteria for selection of house-keeping genes for microarray data from degraded mRNA. We are using some commonly used housekeeping genes in literature for qPCR. However, it is unknown that whether these genes are the appropriate housekeeping genes for our actual microarray data. Thus, an approach for selecting housekeeping genes taking into account the pattern of variation of detected expression signal over age of samples is needed.

CHAPTER 3: ASSESSMENT OF MICROARRAY DATA QUALITY

Abstract

Background: Archived unfrozen newborn blood spots (NBS) collected on filter paper are widely available, and have been shown to retain mRNA sufficient for gene expression profiling. Quantifying the external factors that influence the number of mRNA species and/or quality of the mRNA retained by the filter paper can suggest the most efficient ways of using mRNA from NBS to explore the perinatal origins of diseases.

Methods: We evaluated the effect of storage time on the mRNA species detected using gene expression microarrays. NBS samples were stored for various times (3 – 16 years) and we investigated how mRNA storage time affected the expression patterns. We used the NBS of 53 cerebral palsy cases and 53 matched controls from an ongoing case-control study in whom differential expression of several gene pathways has been demonstrated.

Results: The RNA integrity number (RIN) (2.3+/-0.71) and the 28s/18s rRNA ratio (~0) are persistently low across NBS samples of all ages. We found that for the majority of genes, the signal intensity depended on the storage time of the NBS sample. This decrease in signal was detected using both microrrays and qPCR detection methods. Nonetheless, differential expression at the individual gene level (of the gender-specific genes XIST and KDM5D) and at the gene set level (the fetal inflammatory response syndrome (FIRS) gene set) is detectable, but the signal is more pronounced in NBS samples that were stored for six years old or less.

Conclusion: Differential expression of genes by experimental and control conditions can be ascertained in NBS stored between 3-16 years. Since mRNA microarray data quality decreases over time, potentially muting the extent of differential expression, we recommend prioritizing NBS of six years old or less for study.

Introduction

Newborn blood spotted on filter paper (NBS) is used in every state and most industrialized countries for genetic screening. Archives of leftover blood spot material are available in many states,[92] and mRNA can be extracted from archived NBS, even after many years of unfrozen storage.[93 94 95 96] We have shown that mRNA extracted from archived unfrozen NBS is of sufficient quality to permit mRNA microarray analysis for genome-wide expression profiling.[97] Thus NBS are a potential resource for research that uses gene expression patterns in the newborn period to study diseases of perinatal origin.

Newborn blood spot collection differs from most human blood collection modalities in that no collecting tube is used, but blood is spotted directly from a heel stick incision onto filter paper and dries in minutes. The absence of a liquid environment appears to reduce the activity of the ribonucleases and micro RNAs that degrade mRNA, as most mRNA species are within the cell are quickly degraded after being transcribed. However, some loss of mRNA, especially in older unfrozen blood spot samples is likely and this loss may affect the quality of mRNA used for gene expression profiling.

A commonly used method for assessing RNA quality is the RNA integrity number (RIN), which ranges from 1 (which usually indicted large amounts of mRNA degradation) to 10 (which usually indicates largely intact mRNA).[98] Other methods of RNA quality assessment are the 28s/18s rRNA ratio method,[99 100] the degradometer,[101] and the RNA quality scale (RQS).[102] While RIN, 28s/18s rRNA ratios are useful proxies for measuring RNA quality, they do not inform us if it is possible to still detect a subset of differential expression of genes that can exist between experimental and control groups.

In this chapter, we address two topics related to assessing and addressing the quality of mRNA microarray data obtained from archived unfrozen dried blood spots using Agilent microarray platform. We first address approaches for *assessment* of patterns of mRNA microarray quality which may indirectly reflect mRNA degradation or the amount of extractable mRNA with respect to age of NBS samples. We also address the influence of mRNA microarray data quality on the possibility of *detecting* differential expression. We anticipate that our approaches may help provide guidelines for using NBS in gene expression profiling, helping to make the best use of DNBS, a potentially rich resource for epidemiological and clinical studies.

Methods

The mRNA microarray dataset used for illustrative purposes in this paper is derived from archived unfrozen newborn blood spots of 53 singleton cerebral palsy cases and 53 matched controls (year of birth, sex, gestational age) who are part of on-going case-control study investigating the etiology of cerebral palsy. Among the 106 study subjects, 31 are female and 75 male, with ages ranging from 2.9 - 16 years. More details about this study are described in previous chapters.

Approaches for assessing patterns of mRNA microarray data quality over time.

We examine the following features of raw microarray data:

(1) The distribution of RIN and 28s/18s rRNA ratio across age of NBS samples.

(2) The brightness of raw images across samples and across age of blood spots;

(3) The overall distribution of raw unprocessed probe intensity data by age of blood spots as assessed using density plot and box plots;

(4) The slope of detected expression intensity by age of blood spots for raw aggregated gene data for all genes available in the arrays.

We employ the commonly used method for filtering unqualified spots of microarray data of Paterson et al, in which probe intensity is removed when the gProcessed signal is less than two times the gProcessed signal error.[103] We compare the number of aggregated genes being filtered out and number of genes remaining after filtering across samples and across age of blood spots. We applied quantile normalization for filtered microarray data with a simple modification by stratifying by age groups.

To have a further detailed look at the pattern of the detected expression signal of mRNA microarray data of individual genes, we examined the pattern of detected expression signals of some common housekeeping genes including PPIA, ACTB, GAPDH. We then examined whether this pattern can also be seen or can be validated in qPCR data of these genes.

Effect of mRNA microarray quality over time on detecting differential expression from microarray data

We assess the effect of mRNA data quality over time on two commonly used approaches in microarray work:

(1) The effect of mRNA microarray data quality by age of samples on detecting differential expression of *individual genes*; We illustrate this process using our work on detecting genes (XIST and KDM5D) differentially expressed by gender.

(2) The effect of mRNA microarray data quality by age on detecting differential expression of *gene sets.* We illustrate this process using our work showing that an inflammatory gene set is differentially expressed in children with and without CP.

All data exploration and analysis were done in R version 2.13.2. The R limma package was used for data processing and linear model for microarray data as implemented in the R limma package [104 105] is used for the analysis of differential expression of individual genes

between genders. The GAGE (Generally applicable gene set analysis) method is used for gene set analysis.[106]

Results

Pattern of mRNA microarray data quality over age of blood spots

The RIN values were determined for the mRNA samples isolated from 106 NBS cases. The average RIN for all samples was 2.3±0.71 with the exception of a small percentage of outliers. Next, we looked for an association between RIN value and storage time of the NBS. Somewhat unexpectedly, the RIN tends to be similar between NBS sample regardless of the storage time (figure 3.1a). These results are consistent with the 28s/18s ratio. This ratio is zero for nearly all samples and does not show a time-related trend (figure 3.1b).

Although, the RIN numbers did not show a trend with the storage time of the NBS, we noticed that there was a variation of the overall fluorensensce intensity that was hybridized to the microarrays images. In this visual inspection, the mRNA species isolated from NBS samples stored for a shorter amount of time tended to produce "brighter" array images and the mRNA isolated from NBS samples stored for a long amount of time tended to produce darker array images (figure 3.2). Based on this visual inspection, we tested whether flouresence intensity of the overall array varied by NBS storage time.

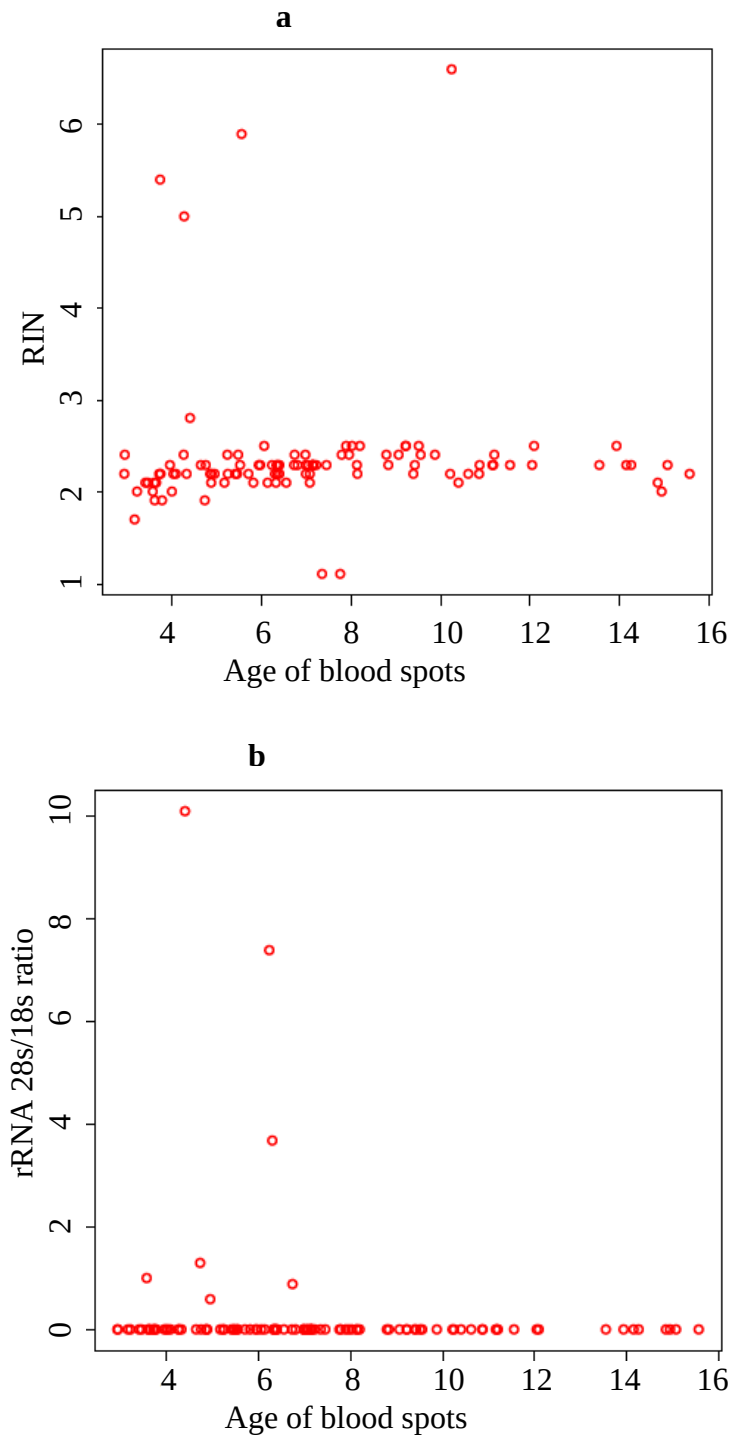

Figure 3.1. The distribution of RIN and 28s/18s ratio across samples with respect to age of blood spots.
a: RIN, b: rRNA 28s/18s ratio.

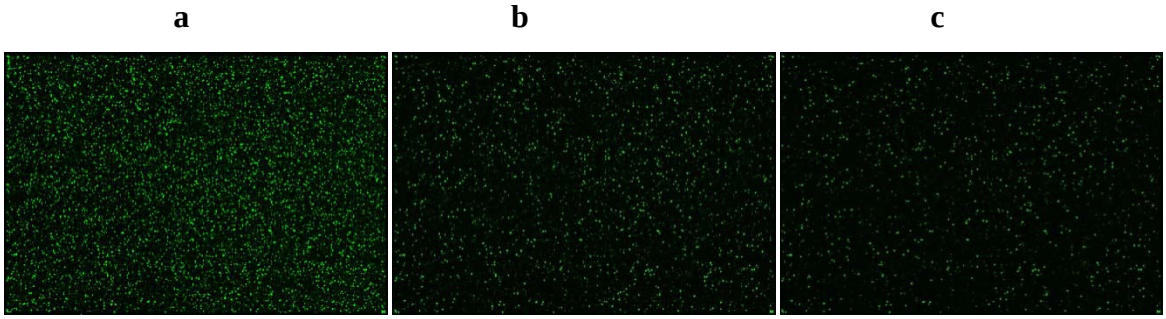

Figure 3.2. Examples of brightness of raw microarray images with respect to age of blood spots.
(a: 4 years; b: 8 years; c: 14 years).

Density plots and box plots of log2-transformed raw (non-normalized) intensity data for mRNA populations isolated from all 106 samples showed that the NBS with shorted storage times tended to produce more higher intensity signals and NBS with longer storage time tended to produce more lower intensity signals (figure 3.3a). The trend of the distribution over age of blood spots (samples are ordered by age of blood spots continuously) can be seen more visually with the box plots (figure 3.3b). The median values of log2 intensity of each of all arrays are significantly linearly decreased over age of blood spots (figure 3.4).

To determine if the decrease in fluoresence intensity was due a decrease in the entire population of mRNA or was limited to a small set of genes, a linear modeling approach was used. For this approach, for each gene, the expression values from each of the 106 samples were isolated and relationship between the log2-transformed intensity and NBS storage time was determined. For 89% of the 21500 genes tested, a significant decrease in expression value was associated with NBS storage time (figure 3.5). In other words, the raw log2 intensity of almost all of the genes linearly decreased as storage time increased. We call the linear slope indicating the decreasing of detected expression signal over age of blood spots for each gene the “decreasing slope”.

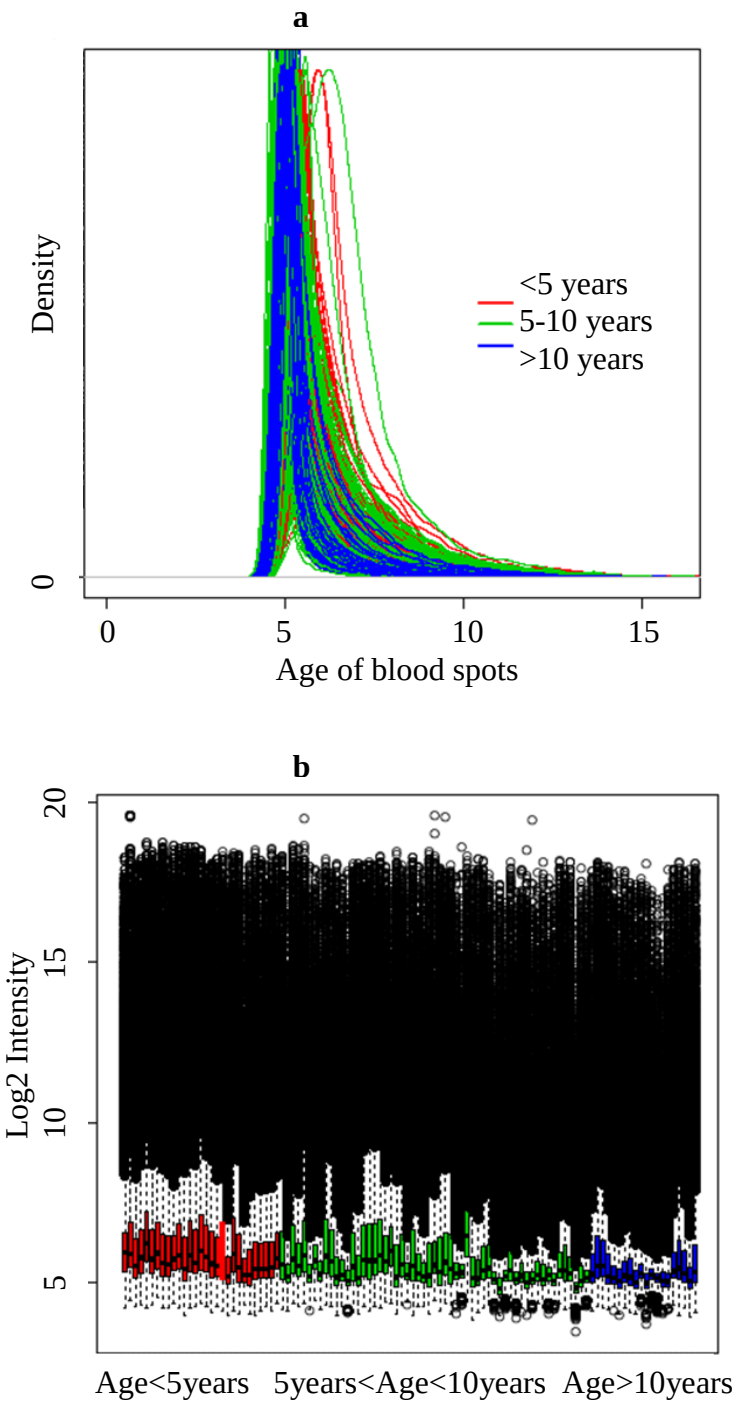

Figure 3.3. Distribution of expression intensity before quantile normalization by age of blood spots.
a: density plot; b: box plot.

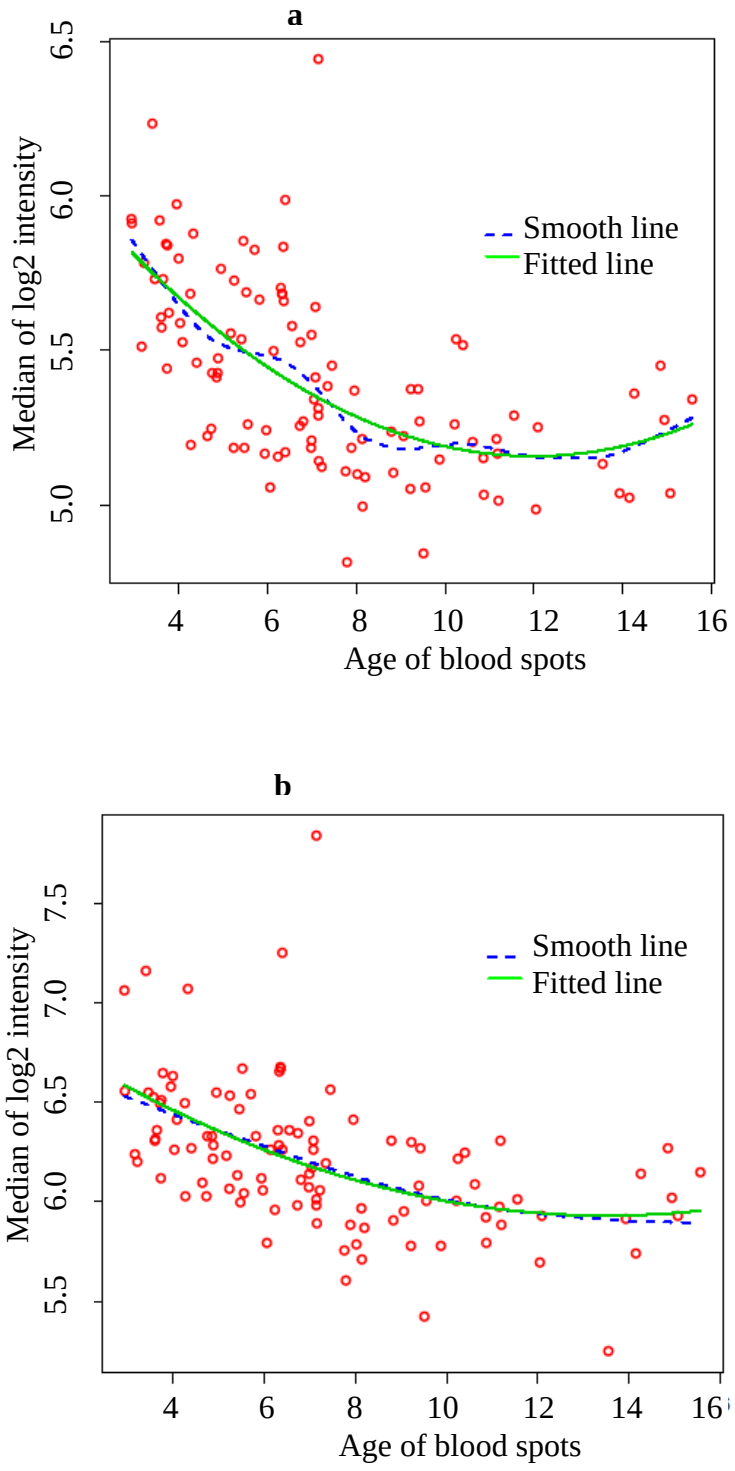

Figure 3.4. Median of log2 intensity of all genes of each of all arrays by age of blood spots before and after filtering.
a: before filtering; b: after filtering.

Another method to determine if detection of mRNA species varies with NBS storage time is to examine the number of features (probes) on the microarrays that show signal intensity above a presumed background level. After applying the filtering method of Paterson et al, the number of genes that were determined to be within the background range was found to linearly increase based on the storage time of the NBS (p-value <0.0001). Likewise, the number of gene expression features remaining for subsequent analysis decreases linearly with the NBS storage time (figure 3.6). Of a total of 21500 genes available in the arrays approximately 35%, 50% and 55% genes are determined to be within the background noise range when the NBS are stored for less than 5 years, between 5 and 10 years, and greater than 10 years old, respectively. The mean number of genes remaining is approximately 13551, 10730, 9925 from blood spots <5, 5-10, >10 years old respectively.

The detected expression signal of the three housekeeping genes ACTB, PPIA, GAPDH are significantly decreased over age of blood spots in microarray data (p.value<0.0001; slope=-0.23, -0.15, -0.14, respectively) (figure 3.10 upper panel). The mean of cycle threshold (CTmean) of qPCR data of these genes significantly increases over age of NBS (except some outliers for PPIA) (figure 3.10 lower panel).

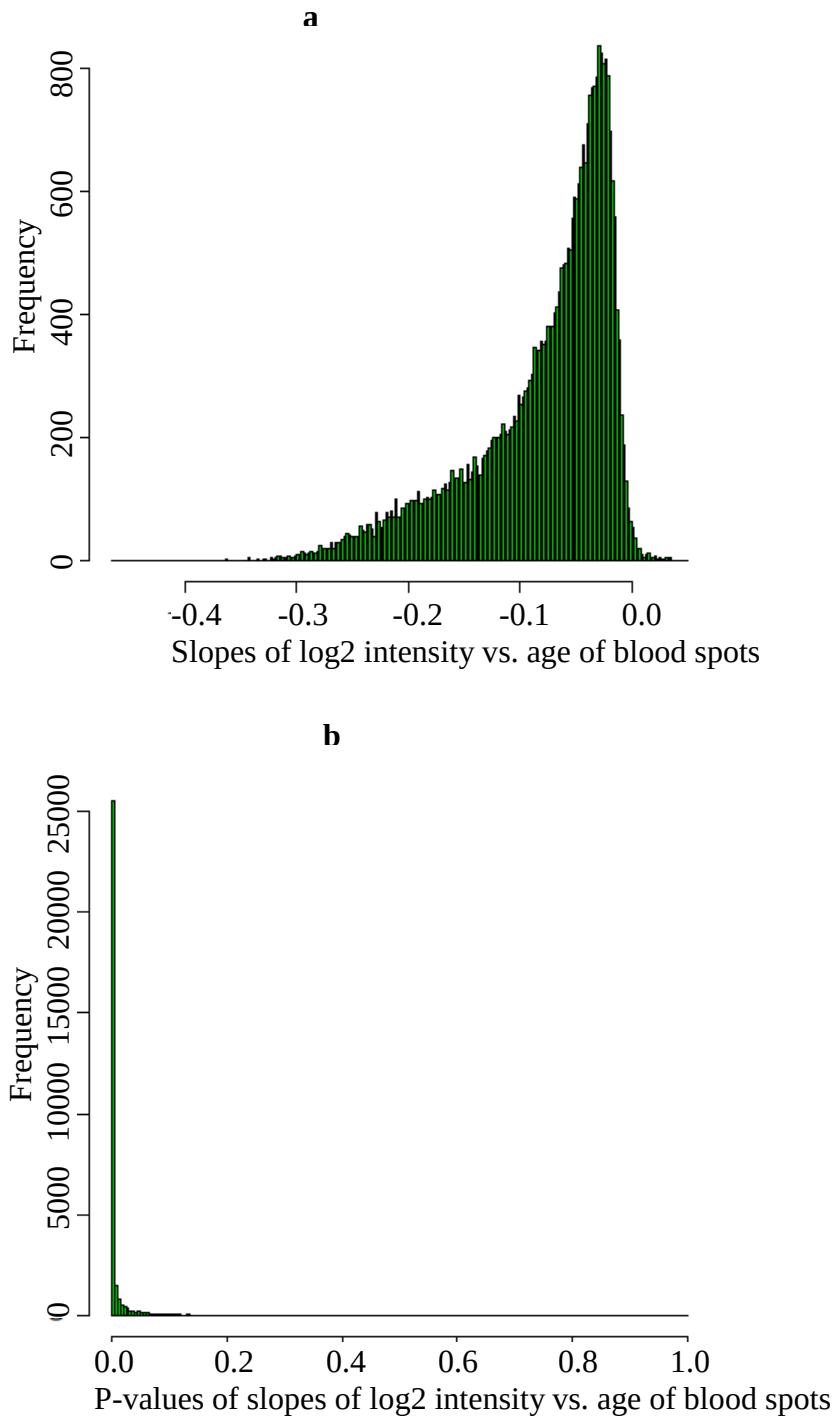

Figure 3.5. Slope of log2 expression signals of all genes vs. age of blood spots of raw microarray data after aggregated to gene level.
a: linear slopes; b: p-values of the slopes.

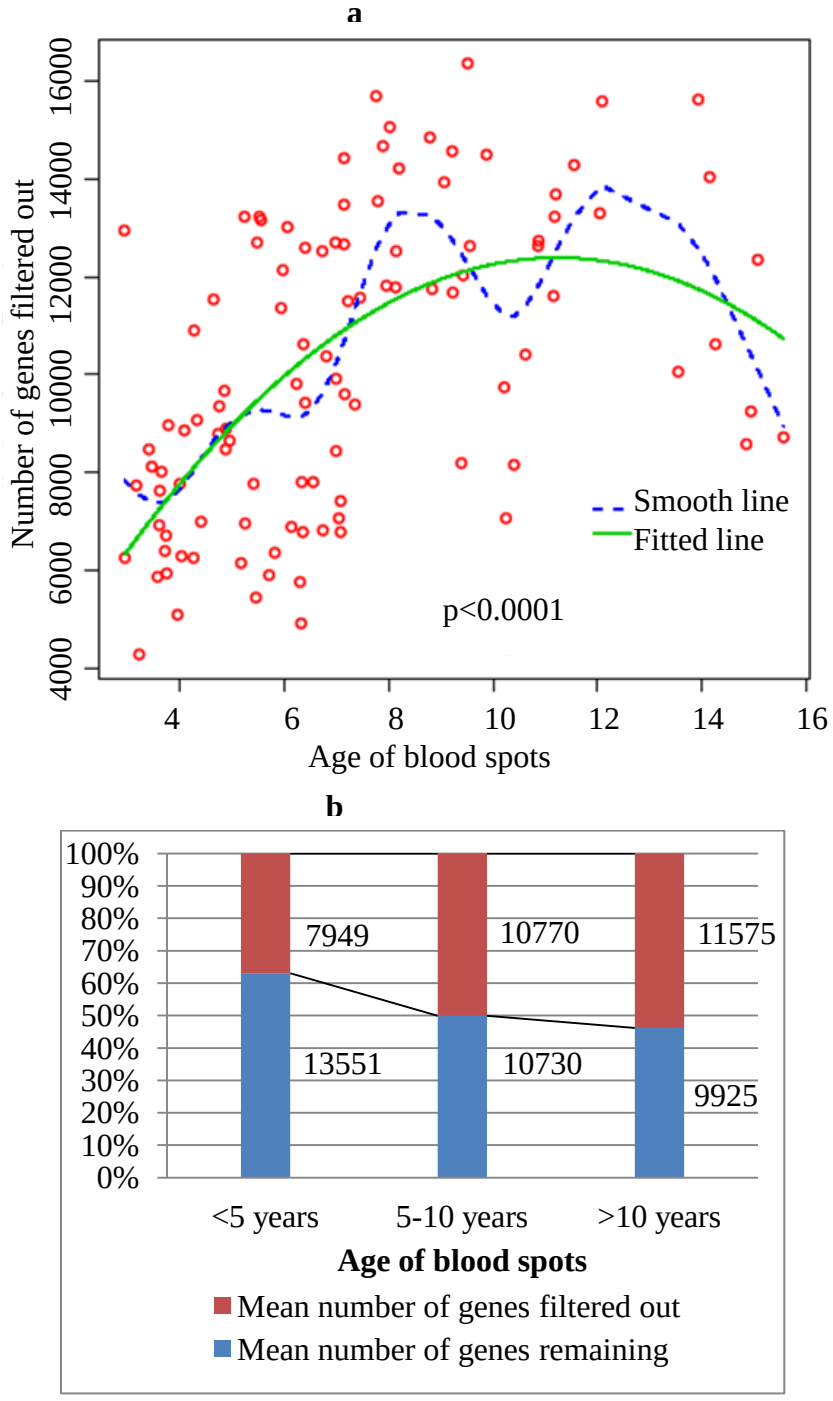

Figure 3.6. Number of genes filtered out and number of genes remaining in the arrays after filtering.

a: Number of genes filtered out; b: Mean number of genes filtered out and remaining by age group.

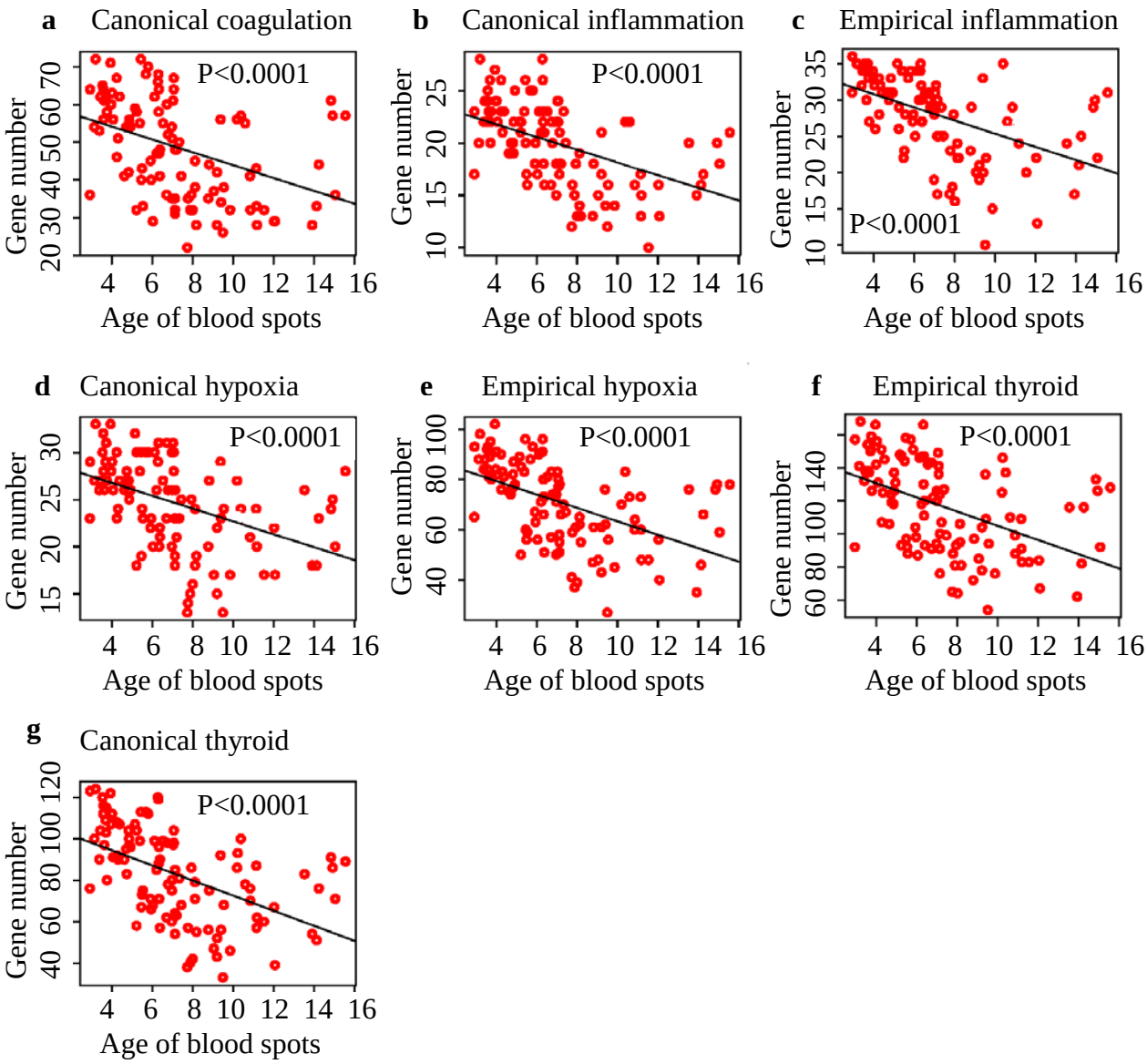

Figure 3.7. Number of genes of the 7 preselected gene sets remaining after filtering of unqualified spots by age of blood spots.
a, b, c, d, e, f, g respectively: Canonical coagulation, Canonical inflammation, Empirical inflammation, Canonical hypoxia, Empirical hypoxia, Empirical thyroid, Canonical thyroid.

Influence of decreasing pattern of detected expression signal on detecting differential expression

To determine if the remaining signal intensities detected by the microarray platform actually reflect the mRNA species examined, we examine two genes that are known to be differentially regulated between the male and female populations. The XIST gene is expressed exclusively in females as part of the system of X chromosome inactivation. Like the other genes, the XIST signal intensity significantly decreases with age of NBS storage in both the female (slope=-0.14, p.value=0.0012) and for male (slope=-0.02, p.value=0.0226) derived samples. However, log2-transformed signal intensity of the XIST gene is higher in females than in males for blood spots mainly before about age 6 years (figure 3.8a). T-test for the difference in mean of log2 intensity between males and females stratified by age groups are all significant but the difference is larger for age group ≤ 6 years than age group >6 years for both XIST (p.value <0.0001, Δmean=1.58 and 1.05 for age group ≤ 6 years and > 6 years respectively)

Analogous to XIST, the KDM5D gene is localized to the Y-chromosome and is expressed in several tissues in males. The log2 intensity of KDM5D significantly decreases over age of blood spots for males (slope=-0.08, p.value=0.0205) but not for females (slope=-0.09, p.value=0.1150) probably due to smaller sample size of females. Although less visually clear than XIST, for KDM5D, log2 intensity of males is higher than that of females for blood spots <6 years old and becomes less separate from that of females for blood spots >6 years old (p.value <0.01, Δmean=1.24 and 0.98 for age group ≤ 6 years and > 6 years respectively) (figure 3.8b).

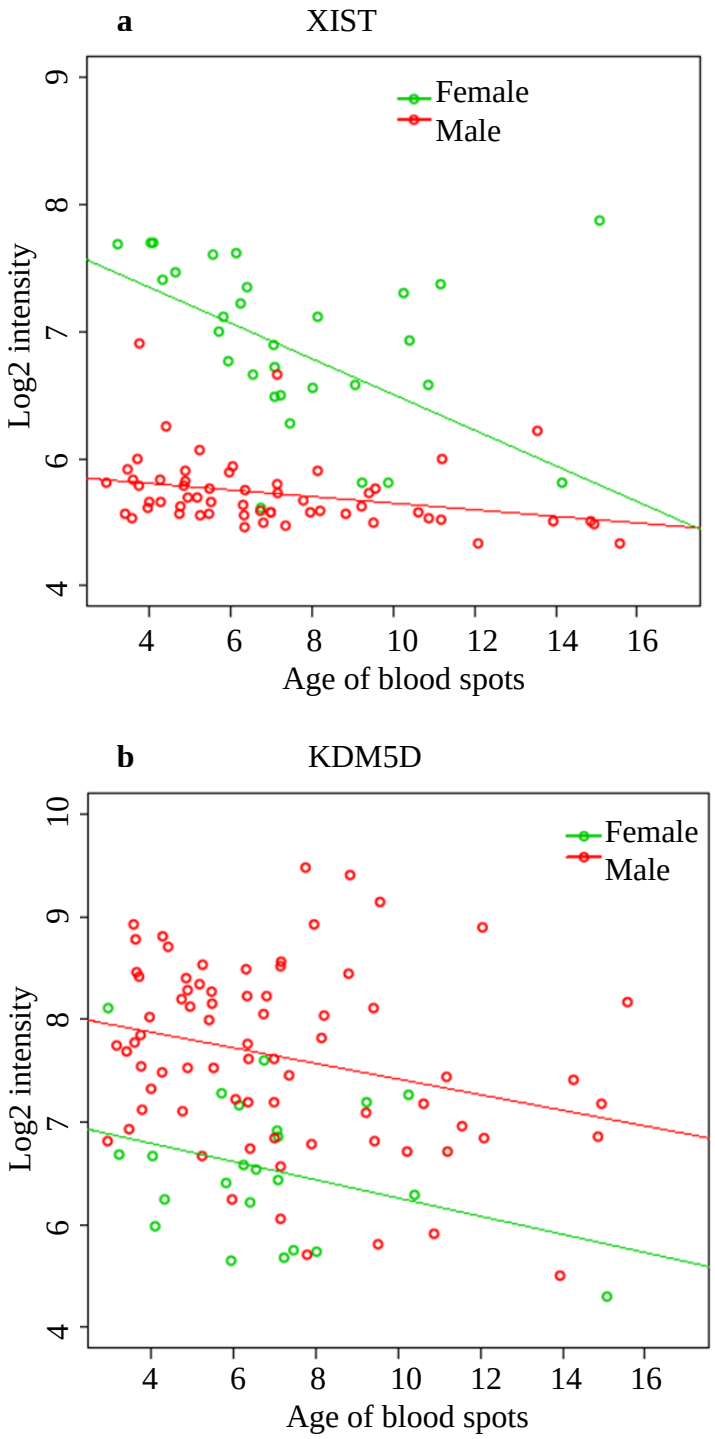

Figure 3.8. Expression signal of XIST and KDM5D genes between males and females.
a: XIST; b: KDM5D.

We made ratios of expression intensity between of CP case and age-matched control individuals. For almost all of the genes, linear modeling approach did not show significant linear relationship between relative expression values (log2 fold change) and NBS storage time. Then we examined whether gene sets of interest showed age-effects. The number of genes of the 7 preselected gene sets remaining after filtering of unqualified spots showed a significant relationship with age of blood spots (p-values <0.0001) (figure 3.7). For the three gene sets showing significant differential expression between cases vs. controls, the GAGE-t-statistics of each matched pairs are standardized by converting to equivalent z-statistics using Stouffer's method. The absolute values of these converted z-statistics are then plotted against age of NBS of the pairs. For the empirical inflammatory gene set which contain small number of genes (n=36), the absolute values of z-statistics are significantly linearly decreased over age of blood spots (slope x=-0.72, slope x^2 =0.04, p.value<0.001) (figure 3.9). Almost all of the z-statistics that are >1.96 are from the pairs with age <6. Chi-square test for the global significance of the z-statistics across pairs is strongly significant for the age group $\leq$ 6 years (p.value <0.0001) and not significant for age group > 6 years (p.value=0.43). For the other two gene sets (empirical hypoxia gene set and empirical thyroid gene set) which contain large number of genes (n=127 and n=140 respectively), the effect of age of blood spots on the magnitude of z-statistics is not linearly significant.

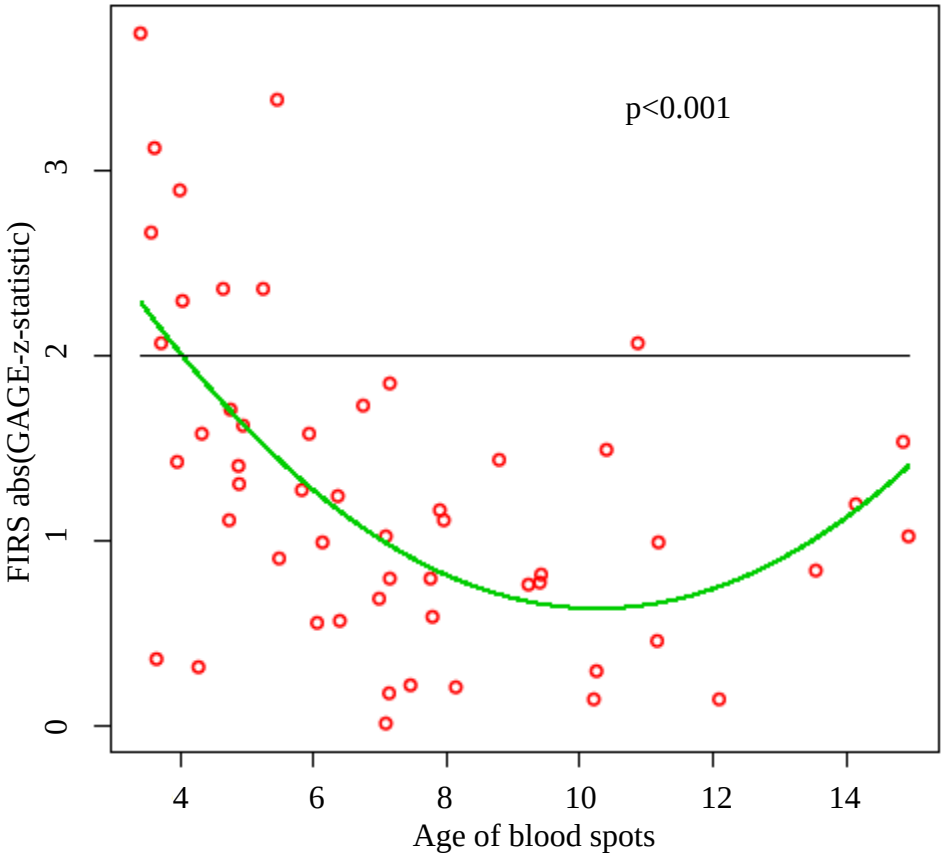

Figure 3.9. Absolute values of GAGE- z-statistics of FIRS gene set of matched pairs over age of blood spots.

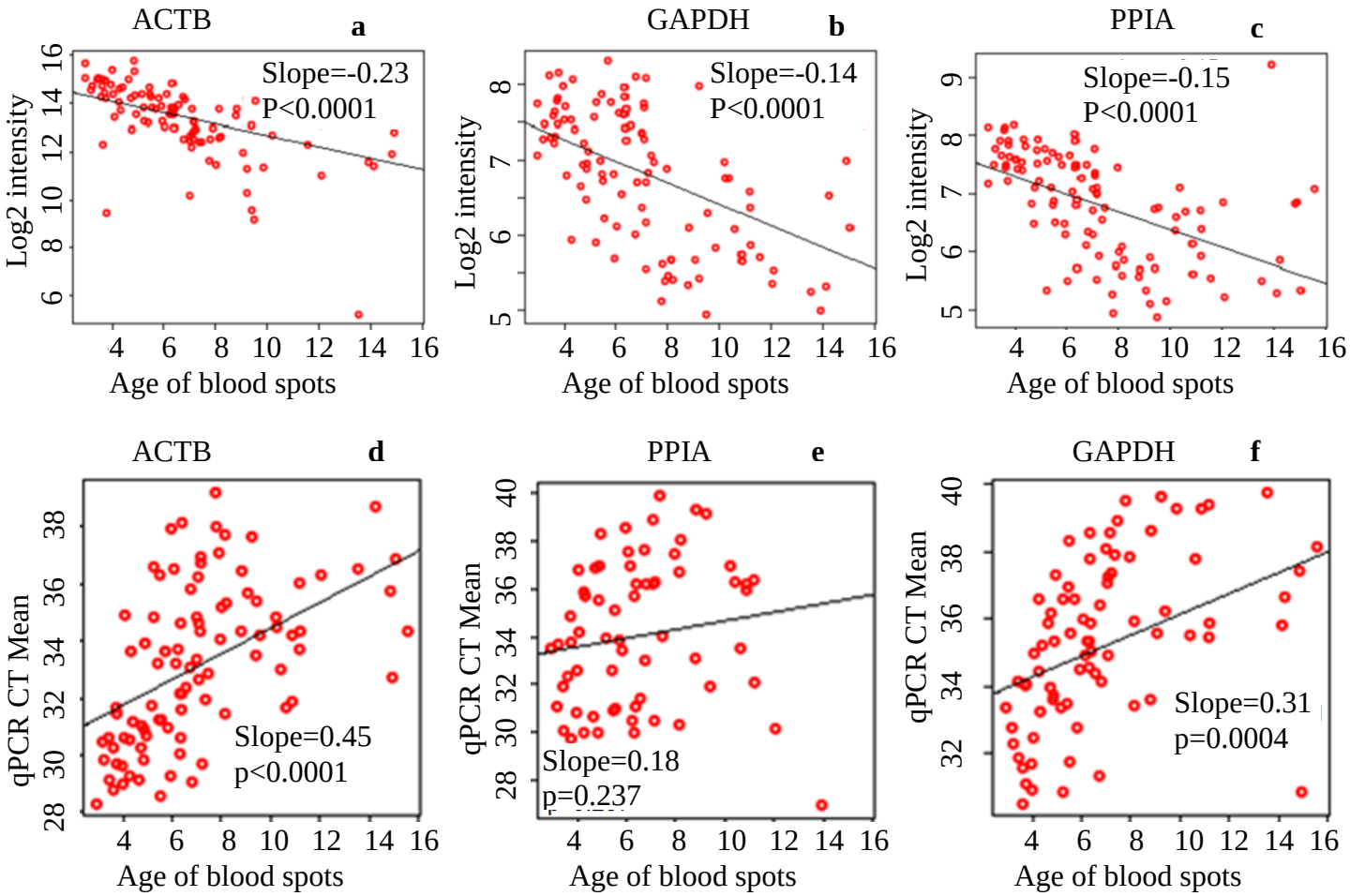

Figure 3.10. Detected expression signal of common housekeeping genes over age of blood spots.
a, b, c of upper panel: microarray data; d, e, f of lower panel: qPCR data.

Discussion

Our findings suggest that although RIN and 28s/18s ratio are low similarly across age of blood spot samples, younger spots do yield more detailed mRNA information than do older blood spots. The RIN results may suggest that mRNA retaining from NBS of age from 3 to 16 is degraded severely regardless of age of spot. Yet, the RIN may not accurately reflect the trend of mRNA microarray data quality. The RIN may describe the relative quality of mRNA but cannot depict how microarray data generated from that mRNA is qualified across samples or whether it is qualified enough for detecting differential expression.

The 28s/18s ratio of zero in almost all samples are due to the absence of the 28s peak. The 28s/18s ratio is an indirect measurement of mRNA degradation since it is based on the availability of rRNA. Some studies have suggested that the correlation between 28s/18s ratio and mRNA integrity is weak. [107] [108] Thus, theoretically, in samples severely degraded based on 28s profile, the 18s and other RNAs species may remain intact. Thus, 28s/18s may not be informative as to the chances of detecting differential gene expression in the experimental situation.

Our explorations of microarray data produced from NBS samples show a consistently decreasing pattern of detected expression signal with increasing age of spot. Since there is a systematic decrease in intensity between younger blood spots and older blood spots, a normalization method should be able to account for this systematic variation across age of blood spots. Our simple modification to the conventional quantile normalization method can address this decrease trend in part. However, we are developing a customized supervised normalization approach that can better address the trend of mRNA microarray quality over time as well as the effects of other covariates on the distribution of detected expression signals across samples.

For almost all genes available in the arrays, the expression signal is significantly diminished over age of blood spots. This decreasing trend is validated by the qPCR data of the three common housekeeping genes ACTB, PPIA and GAPDH. The low signal of all genes in all older blood spots may make the detection of differential expression of genes between different biological statuses more difficult in older blood spots.

The decreasing trend of detected expression signal either in microarray data or in qPCR data may indirectly reflect the increasing degradation trend of mRNA in blood spots over age of blood spots. However, there is no visible trend of RIN and 28s/18s over age of blood spots; thus, this decreasing trend may not reflect the quality (degradation) of mRNA from the samples but rather the decreasing quantity of mRNA that is extractable from the same amount of blood spot sample over age of NBS.

The differential expression of the two gender specific genes, XIST and KDM5D, were much more clearly detected in blood spots <6 years old. At the gene set level, although gene sets with large number of genes may not be affected, example of the empirical inflammatory gene set which contains a small number of genes also shows that significant z-statistics are almost from the blood spots of pairs younger than 6 years old. Thus, although mRNA may still be obtained for blood spots of >6 years old, the quantity or the quality of extractable mRNA may be low and thus lower the possibility of detecting differential expression from the microarray data generated from these mRNA samples.

In summary, considerable amount of mRNA can be obtained from NBS permitting the detection of differential expression between biological statuses, even after ten years of unfrozen storage. However, due to mRNA degradation or the decrease in the amount of extractable mRNA over time, the older blood spots may produce microarray data of lower quality and this may

lower the possibility of detecting differential expression of genes. The possibility of detecting differential expression either at individual gene level or gene set level is higher in blood spots < 6 years old and become less likely for blood spots > 6 years old. Thus, for studies examining gene expression from NBS to investigate the differential expression pattern between biological statuses, blood spots of < 6 years old should be prioritized for study.

CHAPTER 4. APPROACH FOR SELECTING HOUSEKEEPING GENES FROM MICROARRAY DATA OF HETEROGENEOUSLY DEGRADED MRNA SAMPLES

Abstract

When using qPCR to validate mRNA microarray data, housekeeping genes are often needed to accurately quantify RNA. Housekeeping genes are often selected as genes with similar expression level in microarray data across samples. However, in degraded mRNA samples, the levels of gene expression signal detected may decrease across samples over time. Thus, it may not be appropriate to apply the same criteria for selection of house-keeping genes for microarray data from degraded mRNA. In this chapter, we propose an approach for selecting housekeeping genes based on the slopes of detected expression signal over age of samples of all genes for microarray data from heterogeneously degraded mRNA samples. This approach can be generalized to other types of microarray data in which detected expression signal of genes may be influenced by other variables. mRNA microarray and qPCR data from archived unfrozen dried newborn blood spots (uDNBS) of different storage duration will be used to illustrate this approach.

Introduction

Real time quantitative reverse transcription polymerase chain reaction (qPCR) is a reliable method to quantify expression level of genes and thus is often used to validate mRNA microarray data. Artifactual variation or error in measuring gene expression level by qPCR can be due to variations across samples in initial sample amount, mRNA content per unit mass of total RNA, sample preparation, sample loading, sample or nucleic acid quality, RNA degradation, cDNA synthesis efficiency. Thus, housekeeping genes are often used as endogenous controls for determining the availability of relatively intact RNA (cDNA) and especially for

normalization of qPCR data to correct the above potential errors in quantifying the expression level of the target genes. [109] [110] Housekeeping genes are genes that are often believed to express ubiquitously at a stable level in different biological contexts. Thus, housekeeping genes are often selected from array-based expression profiles as genes with expression levels above background and with similar expression levels across samples. However, some commonly used housekeeping genes such as GAPDH, and beta-actin (ACTB) have been reported to express differently across tissue types, or respond differently to different stimuli or experimental conditions. [111] [112] [113] [114] [115] [116] [117] [118] Therefore, in practice, usually a set of housekeeping genes are tested to find the proper ones that are not regulated in the studied condition.

Selecting proper housekeeping genes is essential to accurately quantify RNA. Thus, carefully examining the behavior or pattern of expression signal of potential housekeeping genes together with other genes in studied samples is critical. In some types of tissues, or in some experimental conditions, commonly known housekeeping genes in the candidate list for testing may all not satisfy the criteria for being unregulated housekeeping genes to be used. Thus, it is often necessary to explore or screen a large list of genes of all genes available, or to find out the not-well-known but appropriate genes to serve as control genes. [119] [120] More extensive meta-analysis of multiple gene array samples can be helpful in identifying novel potential housekeeping genes with enhanced stability.[121] Sometimes, no single housekeeping gene is qualified enough. In such circumstances, the geometric means of multiple, carefully-selected housekeeping genes will improve accuracy in the normalization of qPCR data.[122] A model based variance estimation approach proposed by Andersen et al, which evaluates both variation of candidate housekeeping genes and variation between sample subgroups, can provide a

systematic and robust strategy to identify stably expressed genes appropriate for normalization.[123]

In degraded mRNA samples, the detected gene expression signal of non-regulated housekeeping genes may not be similar across samples but may be decreased or different across the age of samples. Thus, it may not be appropriate to apply the conventional criteria for selection of house-keeping genes to microarray data from degraded mRNA. In this paper, we will describe the slopes of detected expression signals over age of samples of all genes available in the arrays. We then propose an approach for selecting housekeeping genes for microarray data from degraded mRNA based on overall slopes of all genes available in the arrays over age of samples. This approach will be illustrated by mRNA microarray and qPCR data from archived unfrozen dried newborn blood spots (uDNBS) of different storage duration. Our approach can be generalized to other types of microarray data in which detected expression signal of genes may be influenced by other variables. In addition, our proposed approach can be easily implemented and can quickly provide potential housekeeping genes for specific microarray study.

Methods

For microarray data of ideal homogeneous tissue with no or little mRNA degradation, the slopes of detected expression signal of all genes over a given variable (e.g. age of samples) center around zero. In other words, the median of the slopes of the detected expression signal of all genes over a given variable is approximate zero. The qualified housekeeping genes are those with similar detected expression signals across samples and thus are those genes with slopes of detected expression signals of all genes over a given variable (e.g. age of samples) that are close to zero, or close to the median of the slopes. In other words, the qualified housekeeping genes are those with slopes that approximate the median of the slopes of all genes available in the arrays.

For microarray data of mRNA that is degraded heterogeneously across samples, the detected expression signal of all genes may be systematically different across samples or may be decreased across age of samples. Thus, the slopes of detected expression signal of all genes over a given variable (e.g. age of samples) may not center around zero. For qualified housekeeping genes, the detected expression signal may not be similar across samples but may decrease over age of samples. Alternatively, the slopes of detected expression signals of all genes over age of samples may not be close to zero but may be negative. Following the above logic, the qualified housekeeping genes are not those with similar detected expression signal across samples but are those with slopes of detected expression signal that approximate the median of the slopes of all genes available in the arrays. Following this logic, selection of housekeeping genes can be done with the following steps:

(1) compute the slopes of detected expression signal (log2 intensity) over age of samples for each of all genes available in the arrays;

(2) compute the robust estimate of the median of all of the slopes with 95% confidence interval (95CI) by using a re-sampling approach. This can be done by randomly sampling 1000 slopes out of the computed slopes of all genes and calculating the median of these 1000 slopes. This procedure can be repeated many times to calculate many median slopes (e.g. 1000 times to calculate 1000 median slopes). The 95CI of the median slope would fall within the 2.5% centile and 97.5% centile of the calculated median slopes.

(3) select the genes with slopes within 95CI of the median slope as housekeeping genes.

To ensure the selected housekeeping genes are not regulated by the studied condition (cases vs. controls), the above procedures are repeated for case subgroups and control subgroups of samples. The genes with slopes within 95CI of the median slopes of both the case and control

subgroups are qualified in both subgroups and thus are not regulated by their membership in the studied groups. In addition, to assure that the selected housekeeping genes are not influenced by experimental technical conditions, such as batch effects, the above procedures are repeated for two batches of the data. The genes with slopes within 95CI of the median slopes of both batches are qualified in both batches and thus are not affected by laboratory batches. Furthermore, to minimize the possibility of selecting false positive genes, and to enhance the robustness of the selection, the selected housekeeping genes are those with slopes within 95CI of the median slopes of both case and control sub-groups and 95CI of the median slopes of both batches. For even more reliable result, among these genes, the genes with slopes within 95CI of the median slope of all subjects are selected as housekeeping genes for qPCR.

Housekeeping genes = $A \cap (B1 \cap B2) \cap (C1 \cap C2)$

where A, B1, B2, C1, C2 are the genes with slopes within 95CI of the whole dataset, batch 1, batch 2, case group, control group, respectively ($\cap$ means joint probability). Ideally, the selected housekeeping genes are the genes that belong to all A, B1, B2, C1, C2.

ACTB, PPIA, GAPDH, RN28S1 are of the most commonly used housekeeping genes and have been evaluated in newborn blood. ACTB was reported to be the least variable while GAPDH was the most variable in neonate blood with hypoxic and acidotic condition. RN28S1 has shown to be the least variable gene in hypoxic condition. Examples of slopes of some commonly used housekeeping genes including ACTB, PPIA, GAPDH, and three rRNA genes RN28S1, RN18S1 and RN5-8S1 are examined to evaluate the pattern of RNA degradation of mRNA and rRNA. In addition, slopes of separate probes of these genes are inspected to assess how degradation is different for different probes of a gene thus allowing us to select the most optimal probes for designing primers for qPCR assays. To evaluate the quality of microarray

data as compared to qPCR data, we examined the trend of qPCR CT mean over age of blood spots for the four genes ACTB, PPIA, GAPDH, RN28S1.

All analyses are done using statistical software R (version 2.13.2). Qualified housekeeping genes should have expression signals above background noise. Thus, we employed the commonly used method for filtering unqualified spots of microarray data of Paterson et al, in which probe intensity is removed when the gProcessed signal is less than two times the gProcessed signal error.[124] The filtered probe data are normalized using quantile normalization method and then aggregated to the gene level using the mean value of the expression signal of all available probes of each gene. The processed data is used to compute degradation slopes over age of blood spots of each gene. R codes for implementing the proposed housekeeping gene selection procedures are available upon request.

The mRNA microarray dataset used for illustration in paper is from archived unfrozen dried blood spots of 53 cerebral palsy cases and 53 matched controls of an on-going case-control study investigating the etiology of cerebral palsy. CP cases and controls are matched by year of birth, and gestational ages. Microarray assays were done in two batches in which batch 1 contains 21 case - matched control pairs and batch 2 contains 32 case – matched control pairs. All 106 study subjects are singletons and aged from 2.9-16 years by the time mRNA is extracted from their newborn blood spots for microarray assays.

Results

Microarray data

The slopes of detected expression signal over age of blood spots of all genes (and their p-values) of raw data and of filtered and normalized data are shown in figure 4.1. For raw data, the degradation slopes of almost all of the genes are negative and almost all (89%) of the p-values of

the slopes are significant. This indicates that the detected expression signals of almost all genes significantly decrease over age of samples, that is, the RNAs of almost all genes are degraded increasingly over time. For filtered and normalized data, because low intensity signals were filtered out generating large percentage of missing values, most of the slopes are still negative, although less negative than those of raw data.

The 95%CI of the median of the slopes of all genes over age of blood spots of all 106 subjects, 53 cases, 53 controls, 42 subjects of batch 1 and 64 subjects of batch 2 are (-0.041; -0.035), (0.044; -0.037), (-0.049; -0.043), (-0.044; -0.037), (-0.055;-0.043) respectively. There are 6 genes with slopes that fall within the 95%CI of the median of the slopes of both case and control groups and both batch 1 and batch 2 including BAIAP2, CSTL1, ZNF544, FLJ45340, PRDX2, RCCD1. Description of these genes is shown in table 1. After applying filtering, normalization and aggregating of probe signals to the gene level, the slopes of these genes are all approximate -0.04 except ZNF544 with slope of -0.05. All slopes are statically significant except ZNF544 and PRDX2 (figure 4.2). Among these genes, RCCD1 has its slope within 95CI of the median slope of all subjects. It is also notable that the variation of the detected expression signal of these genes is quite small (the difference between the lowest signals to highest signals is mostly within 1-3 unit of log2 intensity). This is a favorable characteristic of housekeeping genes. The slopes of different raw probes of each of these genes are shown in figure 4.3.

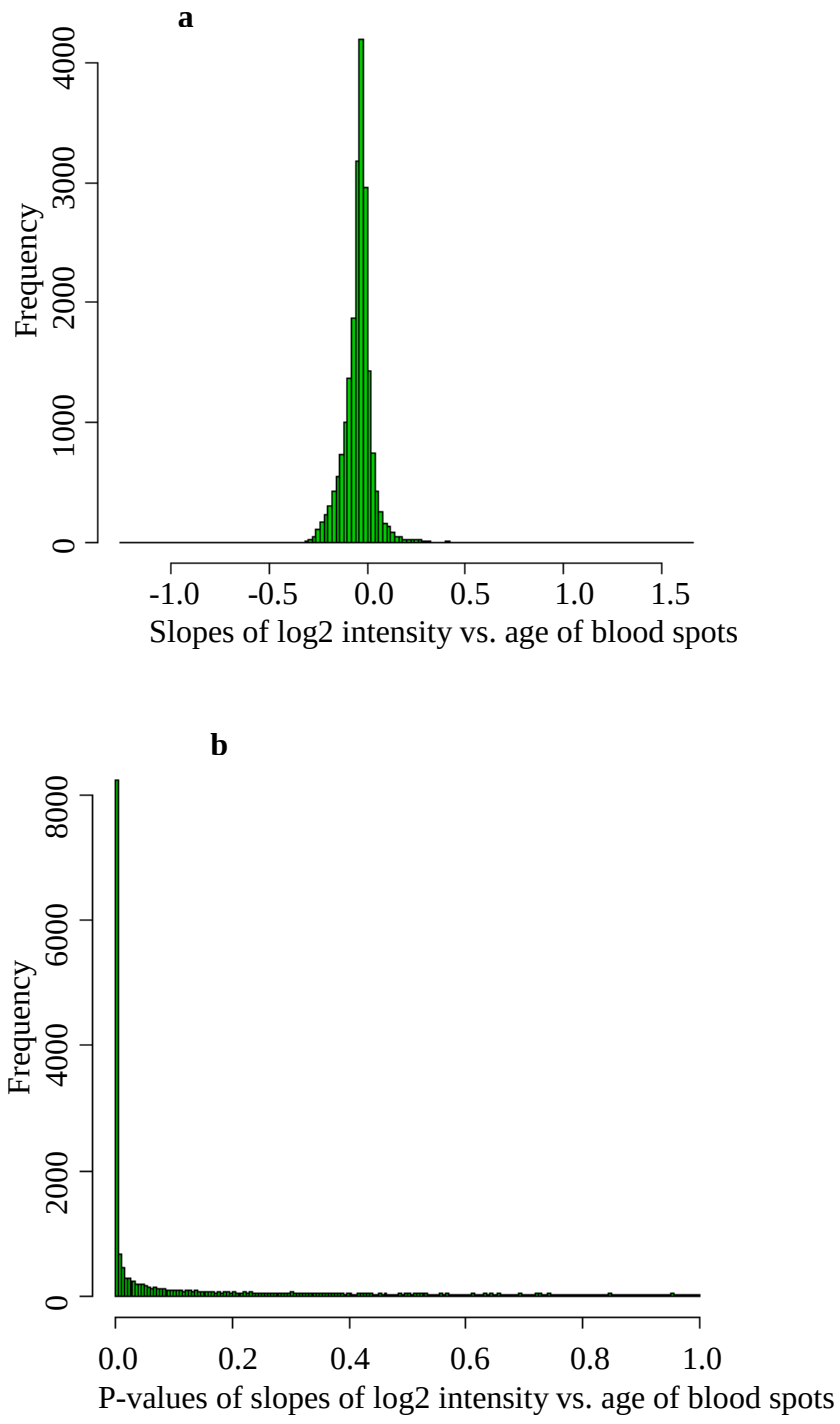

Figure 4.1. Slopes and p-values of quantile normalized log2 expression signal of all genes available in the arrays over age of blood spots.
a: linear slopes; b: p-values of the slopes.

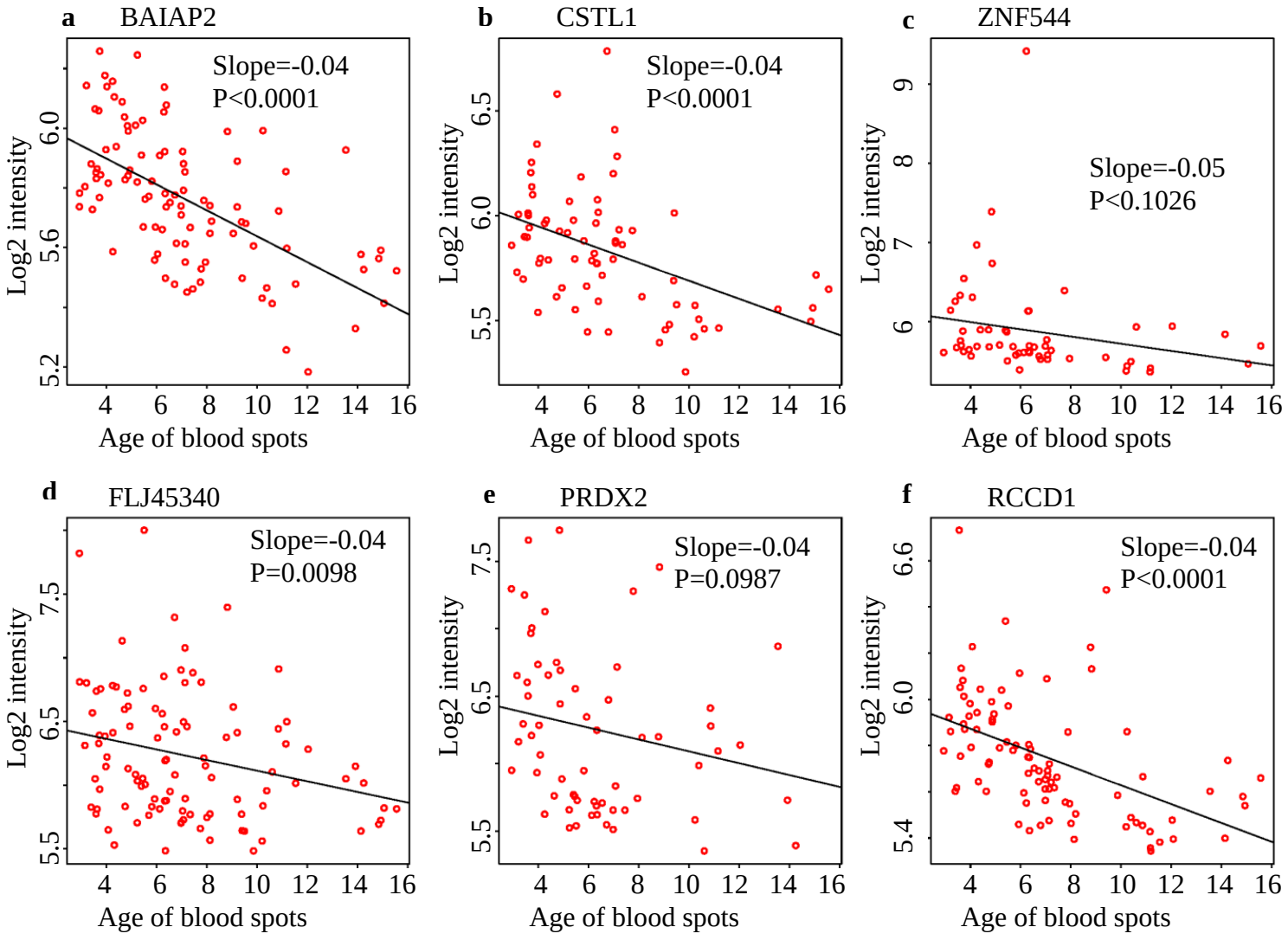

Figure 4.2. Slopes of log2 expression signal over age of blood spots of six selected housekeeping genes.
a, b, c, d, e, f: BAIAP2, CSTL1, ZNF544, FLJ45340, PRDX2, RCCD1.

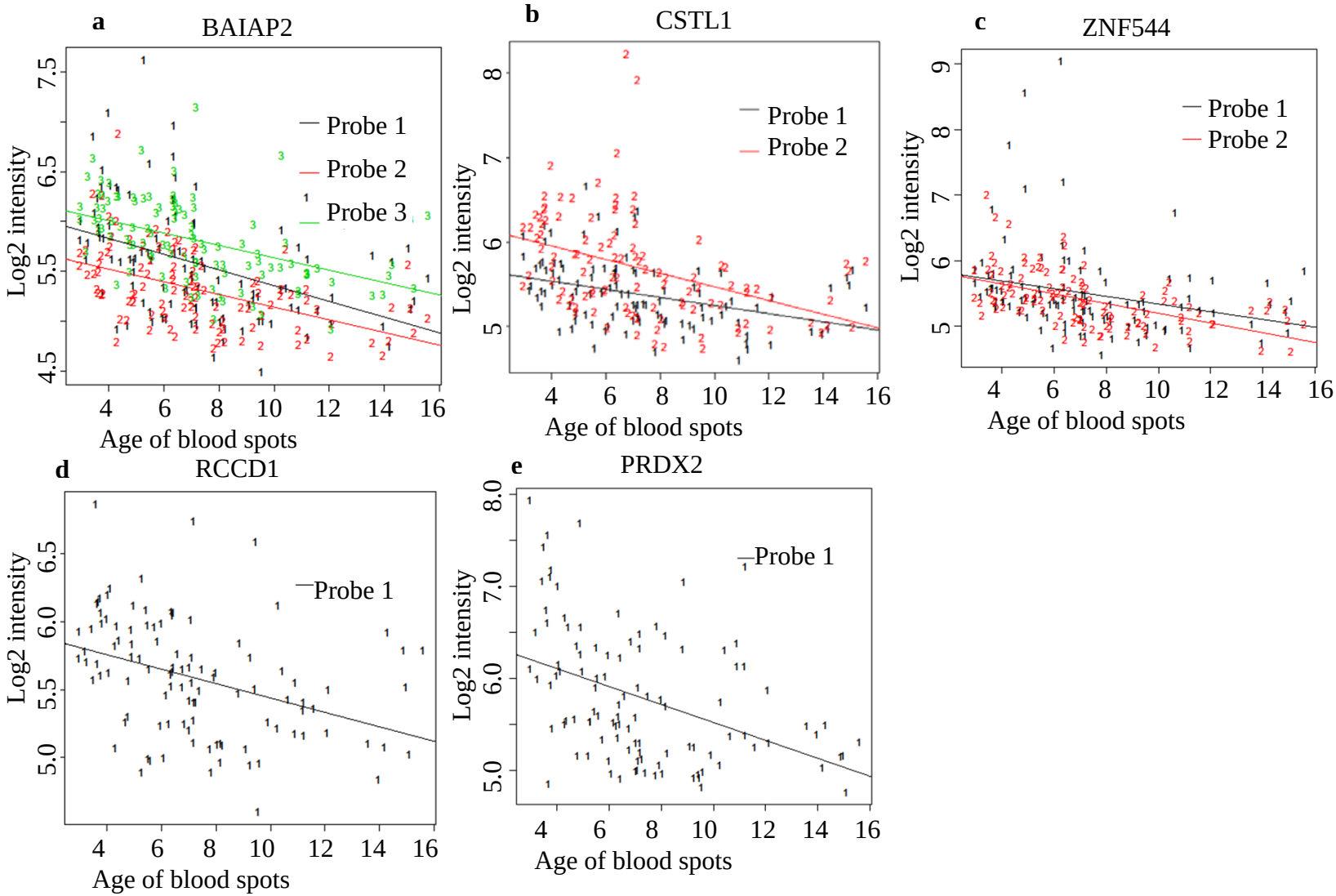

Figure 4.3. Slopes of log2 expression signal over age of blood spots of selected housekeeping genes.
a, b, c, d, e: BAIAP2, CSTL1, ZNF544, RCCD1, PRDX2.

Table 4.1. Description of housekeeping genes selected by the proposed approach.

Gene symbol	Gene full name	Systematic Name	Gene type	Gene function*	Selected Agilent probeUID	Suggested Agilent microarray probe sequence for qPCR
BAIAP2#	BAI1-associated protein 2	NM_006340	protein coding	Encode brain-specific angiogenesis inhibitor (BAI1)-binding protein	13151	TGGCACTACGGAGAGAGTGA GAAGACCAAGATGCGGGGCT GGTTTCCCTTCTCCTACACC
CSTL1#	cystatin-like 1	NM_138283	protein coding	encompasses proteins that contain multiple cystatin-like sequences	38832	AACAATGCCAGCAACGACAC CTACTTATATCGAGTCCAGAG GCTAATTCGAAGTCAGATG
ZNF544	zinc finger protein 544	NM_014480	protein coding	No description in Pubmed	11935	AGCTATCAGTGCGACGTGTAT TAAGCAGCGGTTGTGACTCAT TGAACATCAGAGGACATA
FLJ45340	uncharacterized LOC402483					
PRDX2	peroxiredoxin 2	NM_181738	protein coding	encodes a member of the peroxiredoxin family of antioxidant enzymes, which reduce hydrogen peroxide and alkyl hydroperoxides	31413	TGACTTCAAGGCCACAGCGGT GGTTGATGGCGCCTTCAAAGA GGTGAAGCTGTCGGACTA
RCCD1#	RCC1 domain containing 1	NM_033544	protein coding	No description in Pubmed	5541	TTGCTTTTGAGTGTTAGATAAA TGGAATCCTGTGTATGTGCTTT TGTGTCGTTTTTGTCA

*From Pubmed Gene database. #Better choice for qPCR assay.

For the commonly used housekeeping genes, the slopes of ACTB, PPIA, GAPDH are -0.23, -0.13, -0.15 respectively with p-values <0.0001 (figure 4.4). The slopes of these 3 genes are at 1.5, 10.5, and 8 centile of the slopes of all the genes available in the arrays, respectively. The slopes of the detected expression signals of different probes of ACTB, PPIA, GAPDH follow the general decreased trend of the genes (figure 4.6 upper panel). For five probes of GAPDH, the expression signals of different probes are different and the slope of the probe with highest expression signal is the most negative. For these genes, the sequence of the probe with highest detected expression signal of each gene is used to help design primer for qPCR assays.

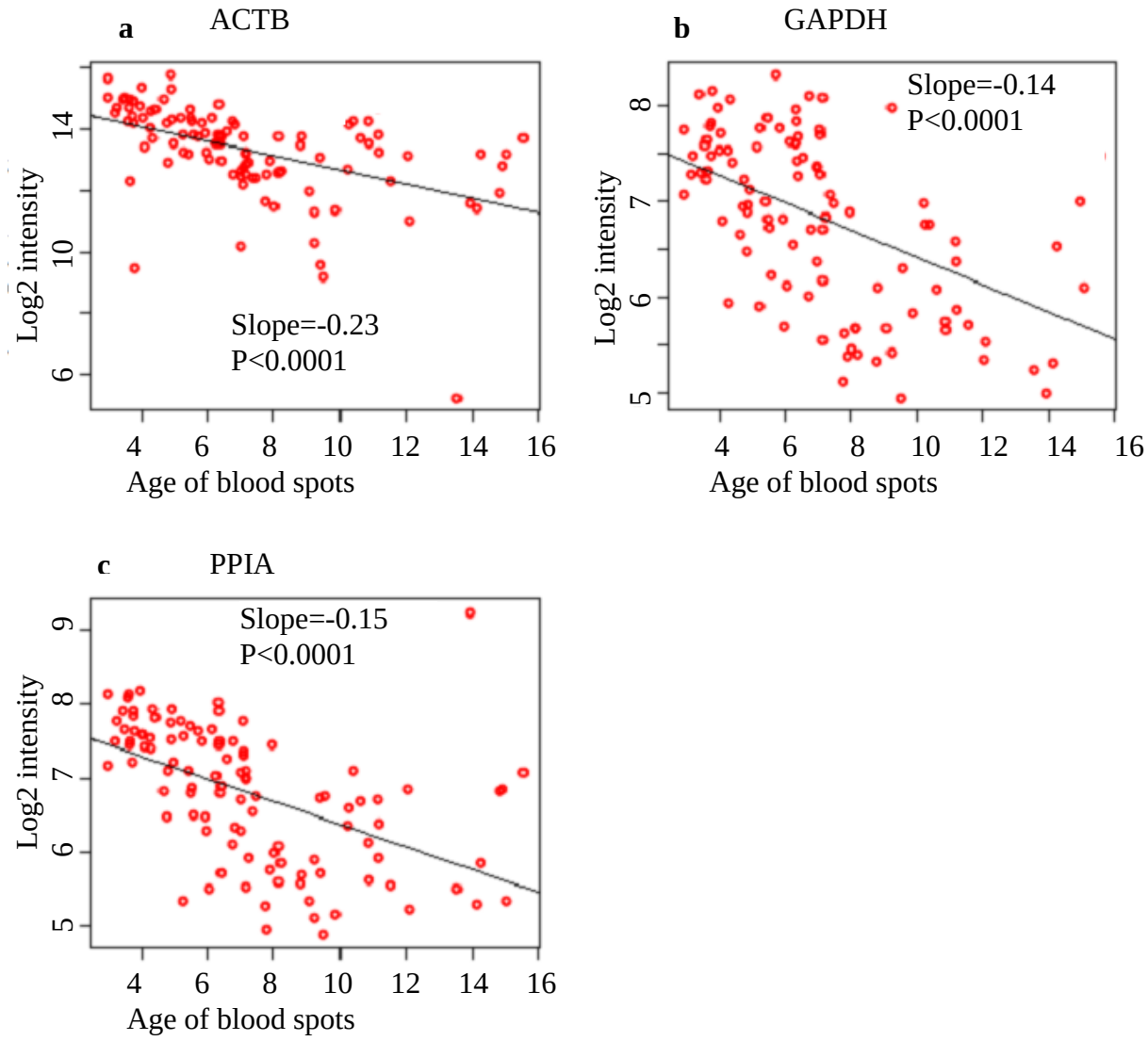

Figure 4.4. Slopes of detected expression signal (log2 intensity) over age of blood spots of commonly used housekeeping genes.
a: ACTB, b: GAPDH, c: PPIA.

There are three ribosomal RNAs (rRNAs) available in the arrays: RN28S1, RN18S1 and RN5-8S1. Their detected expression signals appears to vary or disperse a lot across subjects and seems not to clearly follow the general decreasing trend over time as of mRNAs. Although the slopes of RN28S1, RN18S1 and RN5-8S1 are negative (-0.03, -0.05, -0.07 respectively), none of these degradation slopes are statistically significant with P-value >0.1. (figure 4.5, figure 4.6 lower panel).

qPCR data

CT mean of the ACTB, GAPDH and PPIA (except some outliers) increases with age of blood spots. In other words, the amount of cDNA of these genes decreases over age of blood spots. CT mean or the amount of cDNA of RN28S1 is similar across age of blood spots. This indicates that the decrease trend of detected expression signals of those genes over age of blood spots seen in microarray data is also seen in qPCR data.

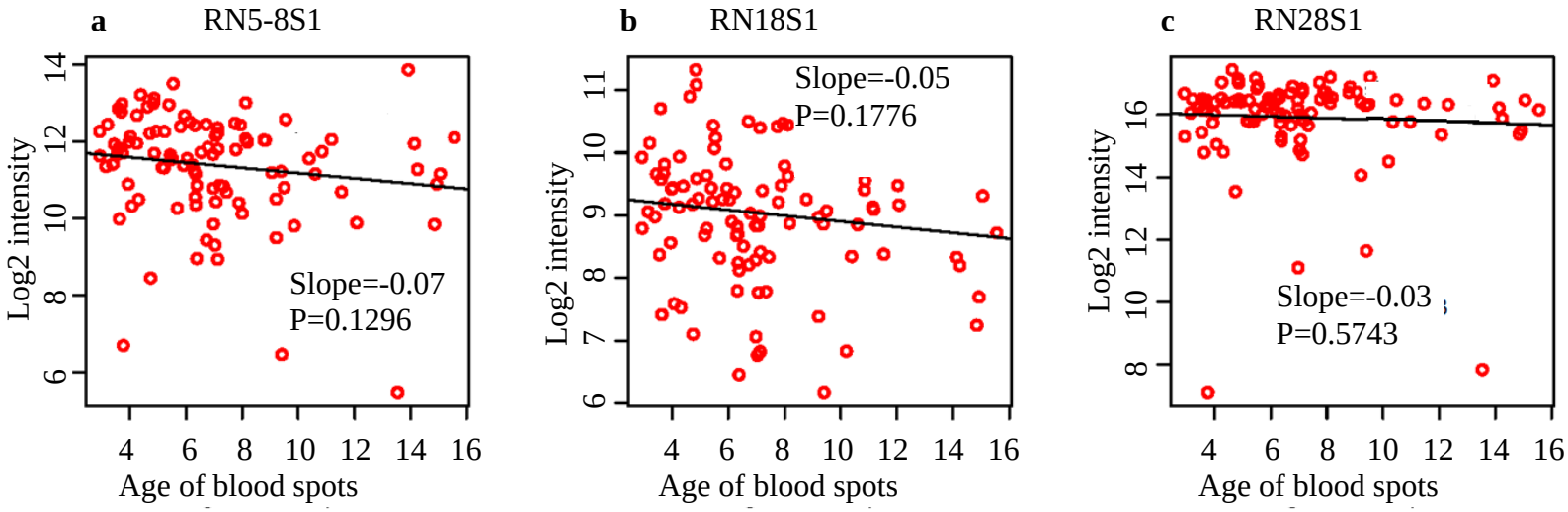

Figure 4.5. Slopes of detected expression signal (log2 intensity) over age of blood spots of rRNA genes.
a, b, c: RN5-8S1, RN18S1, RN28S1.

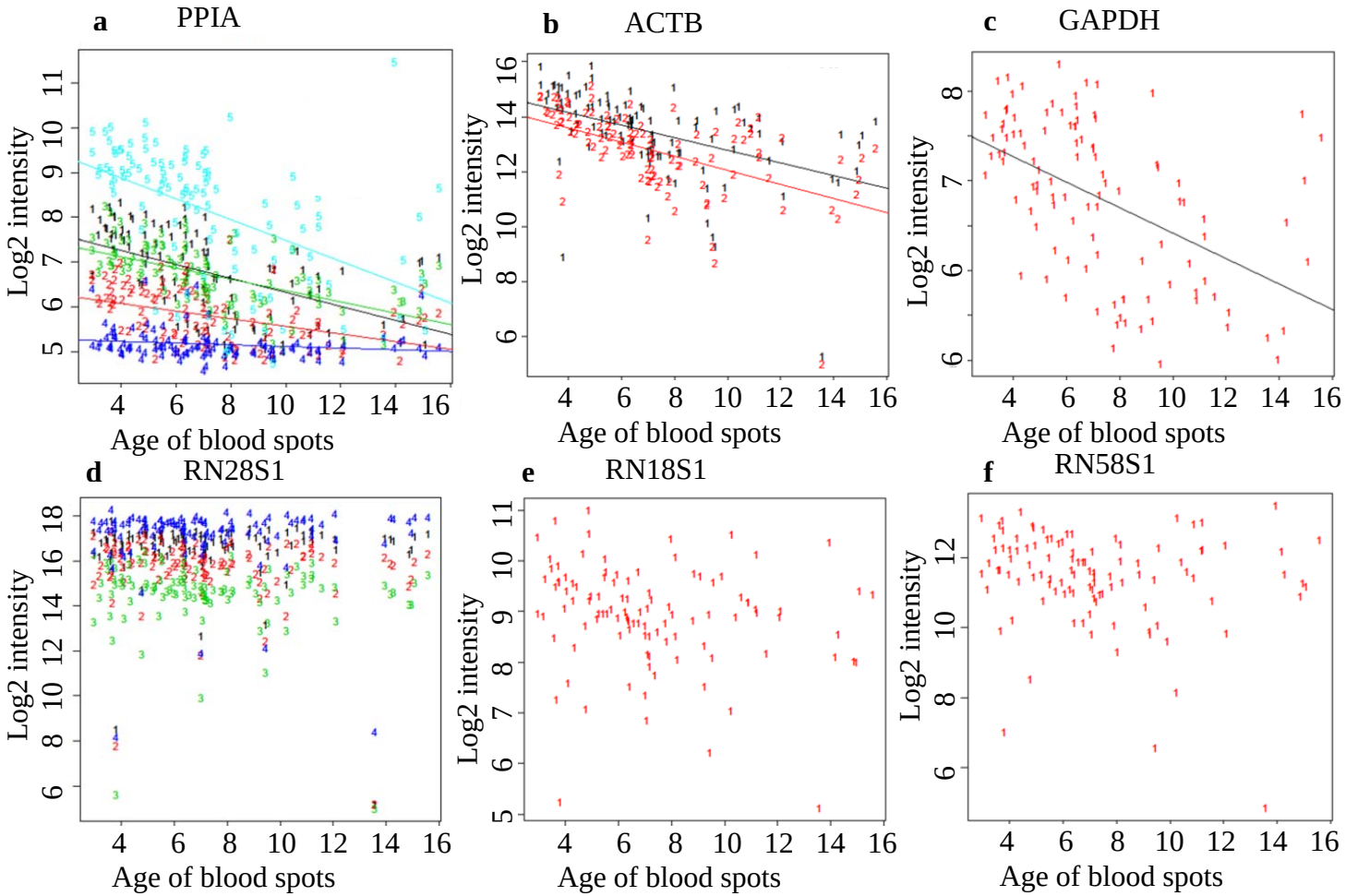

Figure 4.6. Different patterns of slopes of detected expression signal (log2 intensity) over age of blood spots for mRNA and rRNA probes.
a, b, c: PPIA, ACTB, GAPDH; d, e, f: RN28S1, RN18S1, RN58S1.

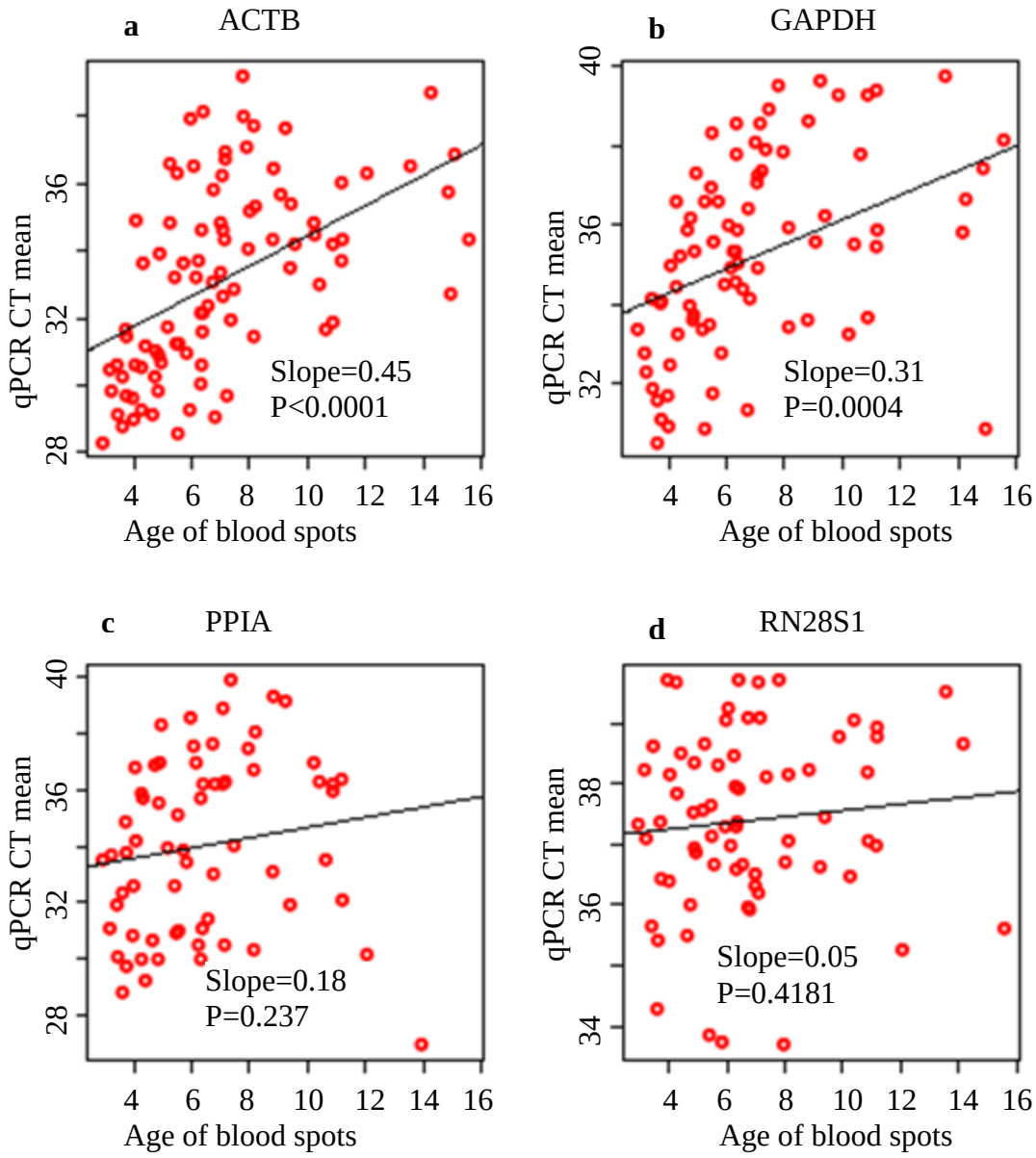

Figure 4.7. qPCR CT mean over age of blood spots of genes with qPCR data.
a, b, c, d: ACTB, GAPDH, PPIA, RN28S1.

Discussion

The degradation trend of mRNA over age of blood spots is reflected in the negative slopes of detected expression signal over age of blood spots of most of all genes in our microarray data. Since almost all of the slopes of the genes available in the arrays are negative, the median of the slopes is also negative. As a result, housekeeping genes are expected to have negative slopes of detected expression signal over age of blood spots.

Using our proposed approach for selecting housekeeping genes, six new potential housekeeping genes are discovered from our microarray data. These six genes all have similar degradation slopes and their degradation slopes are within 95CI of the median slopes of cases, controls and batches. The variation or dispersion of the detected expression signal of the three genes RCCD1, BALAP2, CSTL1 is smallest, and their negative degradation slopes are statistically significant. Although these genes still follow the decreasing trend due to mRNA degradation over age of blood spots, the small variation of their expression signal across subjects within a given age of blood spots make them appropriate to be used as housekeeping genes. Among these genes, RCCD1 has its degradation slope within 95CI of the median slope of all subjects, and is thus better candidate for a housekeeping gene. In other words, the order from most to least favorable housekeeping gene among these six selected genes is RCCD1 > BALAP2, CSTL1 > ZNF544, FLJ45340, PRDX2. The detected expression signals of different microarray probe types of each of the above genes are different. Thus, for the genes selected for qPCR, the sequences of the probes with higher intensity in microarray data should be used to help design the sequences for the qPCR primers of the corresponding genes.

The expression signal of rRNA genes does not clearly follow the decreasing trend over age of blood spots as other mRNA genes. The slopes of these genes are close to the slopes of the

genes selected by our proposed approach. However, these slopes are not statistically significant and more importantly, there is a huge variation of the detected expression signal of these genes across subjects within a given age of blood spots. Therefore, in our microarray data, rRNA genes may be less qualified to serve as housekeeping genes.

For the three commonly used housekeeping genes encoding mRNA ACTB, PPIA, GAPDH, there is a large variation of expression signal across samples within a given age of blood spots. The slopes of these genes are much more negative, or the expression signal of these genes are more prominentlydecreased over age of blood spots than those of the above six genes selected by our proposed approach. Their slopes (-0.23, -0.13, -0.15) are at 1.5, 10.5, and 8 centile of the degradation slopes of all the genes available in the arrays, respectively and thus, are at the extreme end as compared to those of all other genes available in the arrays. Therefore, for our microarray data, these genes may be less likely to qualifiedly serve as housekeeping genes.

The qPCR CT mean over age of blood spots of the three commonly used housekeeping genes ACTB, PPIA, GAPDH and of the gene RN28S1 confirm both the decreasing trend of expression signal over age of blood spots for the genes encoding mRNA, and the unclear trend of expression signal over age of blood spots for the genes encoding rRNA. This indicates that the quality of our microarray data is validated by qPCR data, and that what we observe from our microarray data such as the decreasing trend of detected expression signal over age of blood spots is reliable. Therefore, our proposed approach for selecting housekeeping genes based on the decreasing trend of detected expression signal over age of blood spots observed from our data is rooted in reliable evidence.

Our proposed approach can provide a robust selection of housekeeping genes. First, our approach is based on the robust estimates of the median of the degradation slopes of all genes

using re-sampling technique. Second, we employ several validation strategies for the selection including validation of selection between the experimental vs. control groups, between laboratory technical batches as well as all subjects as a whole. This approach can also be used to test the eligibility of a list of candidate housekeeping genes by comparing their degradation slopes with the estimated median of the slopes of all the genes in the arrays.

Our proposed approach of selecting housekeeping genes is based on the logic that the slopes of detected expression signal of housekeeping genes over a given variable (age of blood spots in our data) approximate the median of the slopes of all genes over that variable. Thus, our approach can be generalized to other types of microarray data in which detected expression signal of genes may be influenced by other variables. In addition, our proposed approach can be easily implemented and can quickly provide potential housekeeping genes for a specific microarray study.

There may be some more work to be done to further evaluate the performance of our proposed approach such as qPCR for the selected housekeeping genes (using, for example, the three best genes RCCD1, BALAP2, CSTL1) and for some important target genes from significant pathways. The qPCR data of these selected housekeeping genes are then used to normalize qPCR data of target genes using the geometric mean method of Vandesompele et al. Then the qPCR data of the target genes are compared with microarray data of those genes. In addition, the degradation slopes of genes in microarray data may also be influenced by the normalization method applied for microarray data. Thus, proper normalization method for microarray data should be developed and used. Then our proposed approach should be repeated to select housekeeping genes from the new normalized data.

CHAPTER 5. GENE SET ANALYSIS OF MATCHED GENE EXPRESSION DATA: AN EVALUATION OF EXISTING AND PROPOSED METHODS ON POWER, TYPE I ERROR AND INFLUENCE OF MISSING VALUES

Abstract

Motivation: Methods for gene set analysis of matched gene expression data have not been well established and evaluated. We address essential issues where the published literature is meager: (1) We test and apply a two stage z-test approach for gene set analysis (ZZ-GSA) of matched genome-wide expression data based on modifications of existing methods using log fold change to assess both homogeneity and heterogeneity in differential expression across matched pairs; (2) We evaluate power of existing methods and our two stage z-test approach for gene set analysis of matched microarray data corresponding to different sample sizes, effect sizes, gene set sizes; (3) Evaluate type I error of existing methods and two stage z-test approach; (4) We evaluate the performance of existing methods and the two stage z-test approach to data with and without missing values, and accordingly propose panels of adjustments for statistical significance for microarray data with and without missing values.

Results: Our simulation study, permutation study and analysis results of actual data have shown that for matched microarray data: (1) the ZZ-GSA approach can assess gene set differential expression homogeneously and heterogeneously across matched pairs; (2) this approach has high power and reasonable type I error in detecting differential expression of gene sets when applied to existing log fold change methods of gene set analysis; (3) with proper implementation, both ZZ-GSA and existing methods perform well on microarray data with and without missing values; (4) our permutation approaches can be used to create reference panels

for type I errors, and for adjustment of statistical significance for different methods of gene set analysis on different microarray datasets with different levels of missing values.

Introduction

Gene set analysis methods for mRNA microarray data can be classified into several categories. One category of methods is based on t-tests of individual genes such as the Kolmogorov-Smirnov running sum summarization; [125] "maxmean" summarization; [126] assessment of gene set overlap using hypergeometric tests, [127] the two-step Q1-Q2 test, [128] the "GSEA-made-simple" method of Irizarry et al.[129] Another category is based on regression models of individual genes; examples are the empirical Bayesian generalized linear models of Goeman et al (2004), [130] the linear model of Jiang and Gentleman (2006).[131] Another category is based on log fold change of individual genes between experimental and control groups including the parametric t-profiler, [132] Parametric analysis of gene set enrichment (PAGE)[133] and generally applicable gene set enrichment (GAGE) (Lou 2009). [134]

Of the methods employing the log fold change, the PAGE method was originally developed for non-matched data. The central limit theorem is applied to log2 fold change between the mean expression of experimental and control groups to test the difference of mean log fold change of genes within the gene set of interest (m) from that of all genes on the array (M):

$$z = \frac{\mathrm{m} - \mathrm{M}}{\mathrm{S}/\sqrt{\mathrm{g}}}$$

where S is the standard deviation of the log fold change over all genes, and g is the number of genes in the gene set of interest.

The t-profiler method was also originally developed for non-matched data. A t-test is used to compare mean of log fold change of the genes in the gene set vs. the remaining genes in the array:

$$t_G = \frac{\mu_G - \mu_{G'}}{S\sqrt{1/N_G + 1/N_{G'}}} \sim t_{N_G - 2}$$

Where

$$S = \sqrt{\frac{(N_G - 1) \times S_G^2 + (N_{G'} - 1) \times S_{G'}^2}{N_G + N_{G'} - 2}}$$

μG is the mean expression log-fold change of the NG genes in gene set G, μG' is the mean expression log fold change of the remaining NG' genes and s is the pooled standard deviation obtained from the estimated variances for gene set G and remaining genes G'.

The GAGE method appears to be the only published gene set analysis method specifically applicable for matched case-control studies. For each matched pair, the GAGE method uses a two-sample-like t-test to compare the expression of genes in the gene set of interest to the expression of all of the genes measured on the array:

$$t_{kl} = (m - M) / \sqrt{\frac{s^2}{n} + \frac{S^2}{n}}$$

where m is the mean of log fold change of genes in the set, M is the mean of all genes in the array, s is the standard deviation of the log fold change of genes in the set, S is the standard deviation of the log fold change of all genes in the array, n is the number of genes in the set. This procedure is followed by a meta-test for global significance, derived from the sum of the

negative logarithms of the p-values of the individual within-pair t-tests based on a Gamma distribution.

$$x = -\frac{1}{L}\sum_{kl} \log P_{kl},$$

where (PX>x)~Gamma(K,1).

However, the two-sample t-test assumes a normal distribution and independence of two samples. While the log2 fold changes of all genes for each pair are not necessarily normal distributed, and the gene set is a tiny part of all genes in the array. In addition, its global meta-test may be oversensitive to extreme values in just a few pairs and can produce a significant p-value for a data set with differential expression in only one or two pairs. In addition, for data set with large heterogeneity among pairs (e.g. some pairs are up-regulated and some pairs are down-regulated) the meta-test can be confusingly significant for both up and down regulation for a gene set. As a result, the meta-test for global significance of the GAGE method may be not a robust test. Furthermore, there has been no paper we know of systematically addressing the issue of heterogeneity in differential expression among matched pairs.

Recently, Brooke L. Fridley et al evaluated power of various self-contained gene set analysis methods at different sample sizes based on simulation.[135] However, there is no published method we know of with formulas for estimating sample size and power for the analysis of gene sets especially for matched data.

Missing values after filtering background noise or unqualified spots in microarray data can be troublesome for any statistical method. Statistical assumptions may be satisfied in data without missing values but no longer satisfied in data with missing values, especially when the percentage of missing values is large. However, not much attention has been paid to the influence of missing values on the performance of gene set analysis methods.

In this chapter, we propose:

(1) A two stage z-test approach that modifies the existing methods of gene set analysis based on log fold change for non-matched data to make these methods applicable for matched data to assess both homogeneity and heterogeneity in differential expression across pairs for both uni-directional gene set and bi-diectional gene set.

Furthermore, we aim to:

(2) Evaluate power corresponding to different sample sizes, effect sizes, gene set sizes of different existing methods after applying our modification for gene set analysis of matched data.

(3) Evaluate type I error of different existing methods after applying our modification proposed method for gene set analysis of matched data.

(4) Evaluate the performance of different existing methods after applying our modification approach on matched microarray data with and without missing values and accordingly propose panels for type I error and panel of adjustments for statistical significance for data with and without missing values.

We will demonstrate our method and approaches by simulation study, permutation and analysis of actual mRNA microarray data from Michigan archived dried newborn blood spots (DNBS) of our on-going matched paired case-control study on cerebral palsy.

Methods

The two-stage z-test approach for gene set analysis (ZZ-GSA)

The general idea of the two-stage z-test approach is that to test the differential expression of a given gene set, a z-statistic is calculated for each matched pair first and then a second z-test is calculated to test for global significance across pairs. This approach may be applied to modify the methods of gene set analysis based on the log fold change of non-matched data to make it

applicable for matched data. The procedures are as below. In stage one, instead of calculating the test statistic based on the log fold change of the mean expression of the experimental group vs. that of the control group, the test statistic is calculated based on the log fold change of expression of each individual experimental subject vs. its matched control. If these test statistics of individual matched pairs are z-statistics, a stage two z-test is performed on these z-statistics to test for global significance across pairs. If the initial test statistics of individual matched pairs in stage one are not z-statistics (for example t-statistics), these test statistics are converted into z-statistics based on their corresponding one sided p-values, and then a stage two z-test is performed on these converted z-statistics to test for global significance across pairs. The stage two global test based on converted z-statistics is similar to the approach of Stouffer et al first proposed for meta-analysis.[136] We applied the ZZ-GSA approach to the PAGE and the t-profiler methods.

When ZZ-GSA is applied to the PAGE method, instead of calculating one z-score based on the log2 fold change of mean expression of the case group vs. the control group, for each matched case-control pair *i*, a z-score z_i is calculated in testing the difference of mean log fold change of genes within the gene set of interest (mi) from that of all genes on the array (Mi):

$$zi = \frac{\mathrm{mi} - \mathrm{Mi}}{\mathrm{Si}/\sqrt{\mathrm{gi}}}$$

where Si is the standard deviation of the log fold change over all genes of pair *i*, and gi is the number of genes in the gene set of interest. The z-scores (z_1, …, z_n) of n pairs are iid ~ $N(0, 1)$. A second stage global z-score is then calculated for the *n* pairs in the study with:

$$Z = (\sum_{i=1}^{n} zi)/\sqrt{n}$$

.

Under the null hypothesis of no significance, Z ~ $N(0,1)$.

When ZZ-GSA is applied to the t-profiler method, in stage one, a tiG is calculated for each matched pair. One sided p-value pi is calculated for each tiG and then converted to zi corresponding to pi. A stage two z-test for global significance based on the converted zi is calculated similarly to above.

While we were preparing this manuscript, the authors of the GAGE method, Luo et al, modified their meta-test for global significance using Stouffer's method to adjust for the drawbacks of summarizing using gamma distribution. In this update, the p-value calculated for the GAGE t-statistic of each pair is converted to a z-score corresponding to that p-value. The global Z is calculated from these converted zi similarly to the global z-test described above. The modified version of the GAGE method may be considered another version of the ZZ-GSA approach when applying to the GAGE method.

Our simulation study and permutation and analysis of actual mRNA microarray data evaluates power and type I error of the ZZ-GSA approach when applied to the PAGE and t-profiler methods, in comparison with the modified GAGE method (Stouffer GAGE) and the early version of the GAGE method (gamma GAGE).

When case-control pairs are heterogeneous (i.e. some pairs are up-regulated while some pairs are down-regulated for a given gene set), the global z-test based on the mean of zi as above become insensitive. Thus, a simple global chi-square statistic:

$X^2 = (z_1^2 + \ldots + z_n^2)$

with n degrees of freedom should be calculated, instead. This chi-square statistic can detect perturbation of gene sets in any direction and thus, can be sensitive to either homogeneous pairs or heterogeneous pairs. This chi-square test can be over-sensitive, however, when the number of pairs is large. So, in a study with the number of case-control pairs n > 20, we propose using a standardized chi-square test to detect the heterogeneity among pairs:

$$\mathrm{Ez} = [\sum_{i=1}^{n} (\mathrm{Zi} - \overline{Zi})^2 - (n-1)]/2(\mathrm{n}-1)$$

Ez follows standard normal distribution if n>20.[137] This test is sensitive if the pairs are heterogeneous, but may be not sensitive if pairs are homogeneous.

Bidirectional gene set. For gene sets in which some genes are up regulated and some genes are down regulated in cases vs. controls, the stage 1 z-test described above is insensitive and thus, should be modified as below:

$z_{i2d} = (m_{i2d} - M_{i2d})/(S_{i2d}/\sqrt{g})$,

where m_{i2d}, M_{i2d} is the mean of the absolute value of log fold change of genes in the set and of all genes, respectively, S_{i2d} is standard deviation of the absolute values of log fold change of all genes for pair *i*. The stage two global z-test is calculated similarly as above. However, since the testing hypothesis is that the genes in the gene set are more perturbed than overall genes in the array, for both zi of stage 1 and Z of stage 2 global test, only upper 1 sided test should be considered.

Estimating required sample size and power

The sample size (n= number of matched pairs) and power calculation for the global z-test follows the formulas of an one sample z-test:[138]

$n = \{ Z_{\alpha/2} + Z_{1-\beta}\}^2 \ (\sigma/\Delta)^2$ and;

power = $\Phi\{ Z_{\alpha/2} + \sqrt{n}(\Delta/\sigma)\}$,

where Δ/σ is the effect size to be detected:

$\Delta/\sigma = (|\mu 1 - \mu o|)/\sigma = |\overline{Zi}| = \overline{[(mi - Mi)/Si]}\sqrt{g}$

in this case, Φ is the cumulative density function of the standard normal distribution $N(0,1)$, n is the number of pairs, g is the number of genes in the gene set, $\mu 1$ is the mean of Zi, μo is the reference mean, which is zero in this case, σ is the reference standard deviation of Zi which is 1 in this case, mi is the mean of log fold change of gene set for pair i, Mi is the mean of log fold change of all genes for pair i, Si is the standard deviation of the log fold change of all genes of pair *i*, and g is the number of genes in the gene set. Thus, power and required sample size depend on the size of the change between case and control and also the size of gene set. $(|\mu 1 - \mu o|)/\sigma$ is the conventional standardized effect size which is dependent on gene set size. $\overline{(mi - Mi)/Si}$ is the gene set standardized effect size, which is independent of gene set size.

For bidirectional gene sets, the formula is similar, but only an upper 1 sided test can be used for global significance. Thus, $Z\alpha$ should replace $Z\alpha/2$.

Simulation study

We simulated microarray data imitating the overall distribution shape of our real quantile normalized log2 intensity microarray data by generating a random number of chi-square distributions of 3 degree of freedom plus 5 and then taking values of the range from <5 and <18. Each simulated dataset contains either 20, 50, 100, 200 or 500 pairs and each simulated array contains 20000 genes. We aimed to test the performance of the ZZ-GSA approaches when applied to the PAGE (PAGE-ZZ-GSA), the t-profiler (t-profiler-ZZ-GSA) and the GAGE method (Stouffer GAGE or GAGE-ZZ-GSA) in comparison with the original GAGE (gamma

GAGE). We evaluated the methods in term of power and type I error of the tests corresponding to different sample sizes, effect sizes and gene set sizes in testing for gene set significance. For each scenario, 1000 simulated expression datasets are created and thus, 1000 tests using ZZ-GSA approaches and gamma GAGE were done.

Power estimation for ZZ-GSA vs. gamma GAGE. We manipulated the expression values of simulated datasets to create differential expression (up-regulation in this simulation study) between cases and controls at different effect sizes for different gene set sizes. A total of 5x3x5x3 (=225) non-null scenarios were generated. The proportion of significant tests (power) out of a total of 1000 tests using ZZ-GSA approaches and gamma GAGE on 1000 simulated datasets were calculated for each scenario corresponding to sample sizes of 20, 50, 100, 200, 500 pairs, gene set sizes of 20 (small size), 100 (average size), 400 (large size) genes and five different crude effect sizes (mean log2 fold change of genes in the set across pairs of 0.05, 0.1, 0.2, 0.5, 1). Gene set standardized effect sizes calculated from corresponding mean log2 fold changes of genes in the gene set across pairs are 0.015, 0.03, 0.06, 0.15, 0.3. Conventional standardized effect sizes for gene set of 20, 100, 400 genes respectively are 0.06, 0.12, 0.24, 0.63, 1.26 and 0.14, 0.28, 0.56, 1.43, 2.86 and 0.29, 0.57, 1.14, 2.85, 5.7. The crude effect size (mean log2 fold change of genes in the set across pair) was generated by random number of normal distribution with mean equal to crude effect size and standard deviation equal to 0.5, 1 and 1.5. For each sample size, each gene set size, and each effect size, the power of a test was calculated by averaging proportion of significant test out of 1000 tests from 1000 simulated datasets of those 3 different standard deviations.

Type I error of ZZ-GSA vs. gamma GAGE. A total of 15 null scenarios were generated. For each of the 1000 simulated microarray datasets of different sample sizes (20, 50, 100, 200,

500 matched case-control pairs), we performed ZZ-GSA and gamma GAGE tests for randomly selected gene sets of 3 different sizes: small (20 genes), average (100 genes) and large (400 genes). The proportion of significant tests out of 1000 tests (type I error) for each of these null scenarios was calculated.

Application to actual matched microarray data: the pilot data of a case-control study on cerebral palsy

To illustrate the method, we performed permutation and applied ZZ-GSA approach and gamma GAGE to our pilot batch of microarray data which contain 21 matched case-control pairs (about 10% of the total planned sample size of our on-going study). Degradation of mRNA is expected in these samples and thus, a large percentage of missing values after filtering unqualified spots is anticipated. Microarray data, after being normalized using quantile normalization method and aggregated to the gene level, were used for gene set analysis. Three versions of the dataset were used. In one version, we did not apply filtering on raw data and thus, this dataset contains no missing values and was used to test the performance of ZZ-GSA and GAGE on microarray data without missing values. In another version, we used the method of Patterson et al [139] to filter out unqualified spots, and thus, this dataset was used to test the performance of ZZ-GSA and GAGE on microarray data with a considerable percentage of missing values of from 20% to 70%. In the third version of the dataset, we replaced the missing values by the smallest expression value of the remaining expression data after filtering (smallest log2 intensity was 5.7 in this situation). The distribution of expression values of this dataset would be “strange” with a high peak due to the large number of genes with same log2 intensity values of 5.7. The distribution of log fold change of all genes of each pair would contain high peak due to a large number of genes with same log2 fold change values of 0. This dataset was

used to test the performance of ZZ-GSA and GAGE on microarray data with unusual distributions, or the distribution of log fold change across genes is not identical.

We carried out two sets of experiments on each of the three versions of our data: (1) Permutation to test the ZZ-GSA approach and gamma GAGE on randomly selected gene sets of different sizes from 10 to 500 genes to create a panel of reference of type I error, normal distribution evaluation, and proportion of outliers corresponding to each gene set size; (2) Testing the methods by exploring 205 KEGG (Kyoto Encyclopedia of Genes and Genomes) gene sets.

Permutation on randomly selected gene sets

Test for normal distribution of calculated z-statistics. We checked the assumptions on normal distribution of z-statistics of individual pairs by examining the z-statistics of gene sets of size from 10 to 500 genes. For each gene set size, for the first (no missing values) and the third (imputed missing values) version of our dataset, for each of the 21 pairs, we randomly sample 1000 times from the pool of all genes in the arrays and calculated the z-statistic for each of the 1000 samples. For the second version of the dataset (containing 20%-70% of missing values), since many gene set containing all missing values were expected to be randomly sampled and thus, no (or missing) z-statistics would be calculated, we sampled 2000 times instead of 1000 times. We then examined the distribution of z-statistics calculated from those 1000 or 2000 samples for each gene set size for each pairs by calculating mean, standard deviation and performing a one sample Kolmogorof-Smirnov test for normal distribution. For each gene set size, we also examined the observed probability of large z-statistics to evaluate the tails of the distribution of z-statistics by calculating the proportion of z-statistics <-3 or >3 of all z-statistics calculated.

Test for type I error. For each of the three versions of the datasets, we performed ZZ-GSA tests and gamma GAGE tests on each of the 1000 random samples for each gene set of different size from 10 to 500 genes. The proportion of significant tests out of 1000 tests (type I error) for each gene set size corresponding to p-values set at 0.01, 0.05, 0.1 were calculated.

Test for gene set significance using KEGG gene sets

KEGG gene sets of different sizes representing different pathways were obtained from the KEGG database. We chose to use KEGG gene sets because they are patho-physiologically relevant to clinical diseases. For each of the three versions of the datasets, we performed ZZ-GSA and GAGE tests for each of the 205 selected KEGG gene sets.

Results

Simulation study

The power of the global z-test of PAGE-ZZ-GSA and the Stouffer modified GAGE test are shown in figure 5.1 (the power of T-profiler-ZZ-GSA is similar to that of PAGE-ZZ-GSA and thus is not shown here). For each gene set size, each sample size, and each simulated up-regulation effect size, the power of ZZ-GSA in detecting up-regulation is higher than the power of Stouffer GAGE. For smaller simulated up-regulation effect sizes, the gamma GAGE test, confusingly, produced a large proportion of significant tests for both up and down regulation (figure 5.2). For example, for the gene set size of 100 genes with simulated mean log2 fold change for up-regulation of genes in the set across pairs at 0.05 and 0.1, all gamma GAGE tests of 1000 tests for up-regulation and all gamma GAGE tests of 1000 test for down-regulation are significant.

The type I error of different tests of ZZ-GSA approach and the gamma GAGE are shown in table 5.1. The type I error of ZZ-GSA is almost equal to random chance when applied for

PAGE, T-profiler, and their global simple chi-square test, and the test for bi-directional gene sets. The standardized chi-square test is more conservative and picks up no false positive. The type I error of ZZ-GSA is smaller than random chance when applied to GAGE (Stouffer GAGE). The gamma GAGE method produces less false positive than random chance.

The gamma GAGE method is more sensitive to large perturbations in only a few pairs and thus the results may not be robust and can be confusing. In another set of simulations where 1000 simulated datasets with only 5 out of 50 pairs (10%) with mean log2FC of 1 for a given gene set were generated, 100% of 1000 tests using gamma GAGE are significant while only 9% of 1000 tests using ZZ-GSA are significant. In another 1000 simulated datasets in which around 40% of pairs are up regulated and 40% are down regulated, 99% of 1000 gamma GAGE tests are significant for both up-regulation and down regulation. The global z-test of ZZ-GSA does not have these drawbacks. The combination of use of the global second stage z-test (test for perturbation of gene set in similar direction across pairs or homogeneous pairs) and a simple chi-square test (screen for any important differential expression of gene set in any direction across pairs) and a standardized chi-square test (test for heterogeneity of differential expression among pairs) can help detect any perturbation, and thus increase power without increasing type I error. The use of the three tests thus helps us to interpret the results properly.

Table 5.1. Type I error from simulated data.

Number of matched pairs		**20**			**100**			**200**			**500**		
Gene set size	Reference#	20	100	400	20	100	400	20	100	400	20	100	400
ZZ-GSA													
Z-test Up	0.010; 0.050; 0.100	0.001; 0.052; 0.099	0.006; 0.063; 0.122;	0.002; 0.037; 0.077;	0.005; 0.041; 0.082	0.006; 0.056; 0.099	0.001; 0.041; 0.092	0.009; 0.040; 0.087	0.006; 0.054; 0.102;	0.007; 0.037; 0.077;	0.007; 0.051; 0.083	0.010; 0.046; 0.091	0.008; 0.051; 0.072
Z-test Down	0.010; 0.050; 0.100	0.008; 0.023; 0.071;	0.001; 0.041; 0.111;	0.015; 0.069; 0.122;	0.015; 0.081; 0.112	0.001; 0.026; 0.081	0.011; 0.061; 0.102	0.007; 0.043; 0.072;	0.005; 0.041; 0.081;	0.008; 0.059; 0.091;	0.011; 0.041; 0.100	0.004; 0.036; 0.091	0.009; 0.051; 0.101
T-profiler Up	0.010; 0.050; 0.100	0.002; 0.060; 0.111	0.009; 0.069; 0.122	0.003; 0.050; 0.070	0.011; 0.022; 0.073	0.009; 0.068; 0.091	0.001; 0.041; 0.082	0.006; 0.041; 0.122	0.006; 0.060; 0.112	0.007; 0.037; 0.082	0.008; 0.032; 0.065	0.006; 0.048; 0.095	0.007; 0.051; 0.072
T-profiler Down	0.010; 0.050; 0.100	0.009; 0.030; 0.071	0.002; 0.061; 0.121	0.012; 0.060; 0.123	0.021; 0.061; 0.132	0.001; 0.033; 0.092	0.008; 0.043; 0.126	0.004; 0.026; 0.071	0.007; 0.054; 0.110	0.009; 0.061; 0.079	0.009; 0.051; 0.102	0.006; 0.043; 0.094	0.011; 0.053; 0.106
Chi-square	0.010; 0.050; 0.100	0.010; 0.068; 0.110	0.004; 0.043; 0.112	0.005; 0.021; 0.060;	0.011; 0.038; 0.081	0.004; 0.043; 0.092	0.005; 0.031; 0.061;	0.009; 0.030; 0.089	0.004; 0.033; 0.090	0.015; 0.051; 0.086;	0.009; 0.040; 0.009	0.004; 0.033; 0.100	0.011; 0.041; 0.104;
Standardized chi-square	0.010; 0.050; 0.100	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0
Bi-direction gene set	0.010; 0.050; 0.100	0.001; 0.041; 0.098	0.005; 0.049; 0.112	0.009; 0.067; 0.131	0.004; 0.032; 0.092	0.001; 0.049; 0.093;	0.001; 0.051; 0.083	0.002; 0.031; 0.110	0.008; 0.049; 0.073	0.004; 0.038; 0.079	0.001; 0.021; 0.120	0.009; 0.039; 0.082	0.001; 0.037; 0.069

Table 5.1. (cont'd)

Number of matched pairs		**20**			**100**			**200**			**500**		
Gene set size	Reference#	20	100	400	20	100	400	20	100	400	20	100	400
GAGE													
Gamma up	0.010; 0.050; 0.100	0; 0.001; 0.010	0; 0.002; 0.004	0; 0.001; 0.003	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0
Gamma down	0.010; 0.050; 0.100	0; 0.001; 0.002	0; 0.001; 0.003	0; 0.002; 0.002	0; 0; 0	0 0; 0	0; 0; 0	0; 0; 0	0 0; 0	0; 0; 0	0; 0; 0	0 0; 0	0; 0; 0
Bi-direction gene set	0.010; 0.050; 0.100	0; 0.003; 0.006	0; 0.001; 0.002	0; 0.002; 0.003	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0
Stouffer up$	0.010; 0.050; 0.100	0.001; 0.010; 0.120	0.001; 0.001; 0.004	0.001; 0.002; 0.005	0; 0.023; 0.094	0; 0; 0	0; 0; 0	0; 0.020; 0.111	0; 0.001; 0.003	0; 0; 0	0; 0.021; 0.114	0; 0.002; 0.004	0; 0; 0
Stouffer down$	0.010; 0.050; 0.100	0.001; 0.050; 0.110	0.001; 0.002; 0.003	0.001; 0.003; 0.004	0; 0.032; 0.110	0; 0; 0	0; 0; 0	0.002; 0.033; 0.102	0; 0; 0	0; 0; 0	0.001; 0.042; 0.112	0; 0; 0	0; 0; 0
Bi-direction gene set$	0.010; 0.050; 0.100	0.001; 0.010; 0.040	0.000; 0.001; 0.003	0.001; 0.001; 0.003	0.002; 0.016; 0.067	0; 0.002; 0.009	0; 0; 0	0.001; 0.007; 0.018	0; 0.003; 0.007	0; 0; 0	0.001; 0.006; 0.017	0; 0.001; 0.006	0; 0; 0

Type Ierror (proportion of significant tests out of 1000 tests on random sets of genes from 1000 simulated datasets) were calculated for significant level of p-values set at 0.1, 0.05, 0.01 respectively. #Reference: expected false positives due to random chance for significant p-value set at <0.01, <0.05, <0.1. $Modified GAGE using Stouffer's method.

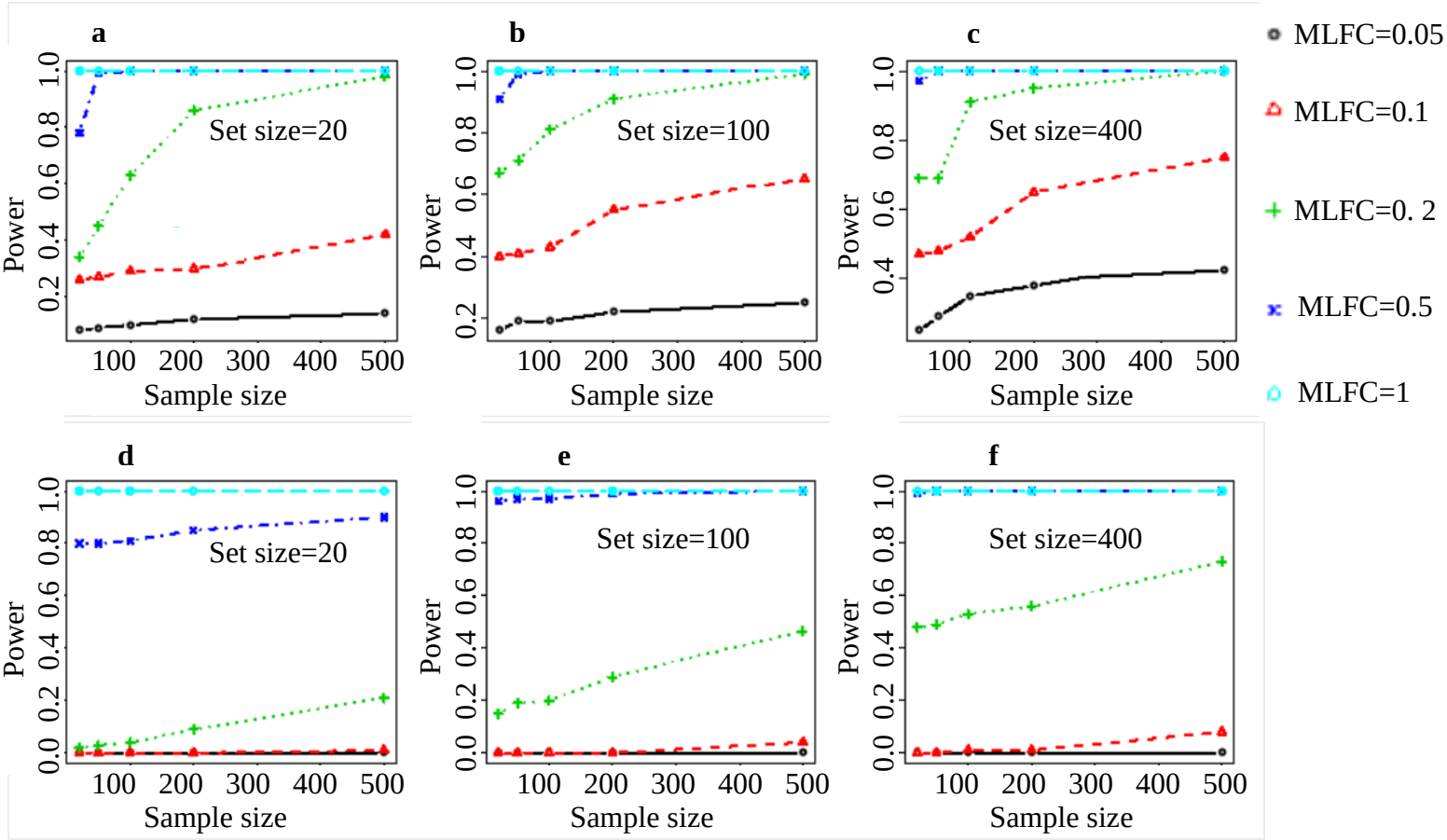

Figure 5.1. Power of PAGE-ZZ-GSA vs. (Stouffer) GAGE-ZZ-GSA.
The simulated mean log2 fold change is for up-regulation (MLFC: mean log2 fold change of genes in the gene sets across pair >0). Results of the tests for up-regulation are shown. Upper panels: ZZ-GSA, lower panels: Stouffer-GAGE. For each gene set size of 20, 100, 400 genes, power (proportion of significant tests out of 1000 simulations) was calculated corresponding to different crude effect sizes (mean log2 fold change of all genes in the gene set across pairs) 0.05, 0.1, 0.2, 0.5, 1. Gene set standardized effect sizes calculated from corresponding mean log2 fold change of genes in the gene set across pairs are 0.015, 0.03, 0.06, 0.15, 0.3. Conventional standardized effect sizes for gene set of 20, 100, 400 genes respectively are 0.06, 0.12, 0.24, 0.63, 1.26 and 0.14, 0.28, 0.56, 1.43, 2.86 and 0.29, 0.57, 1.14, 2.85, 5.7. a, b, c: PAGE-ZZ-GSA; d, e, f: GAGE-ZZ-GSA.

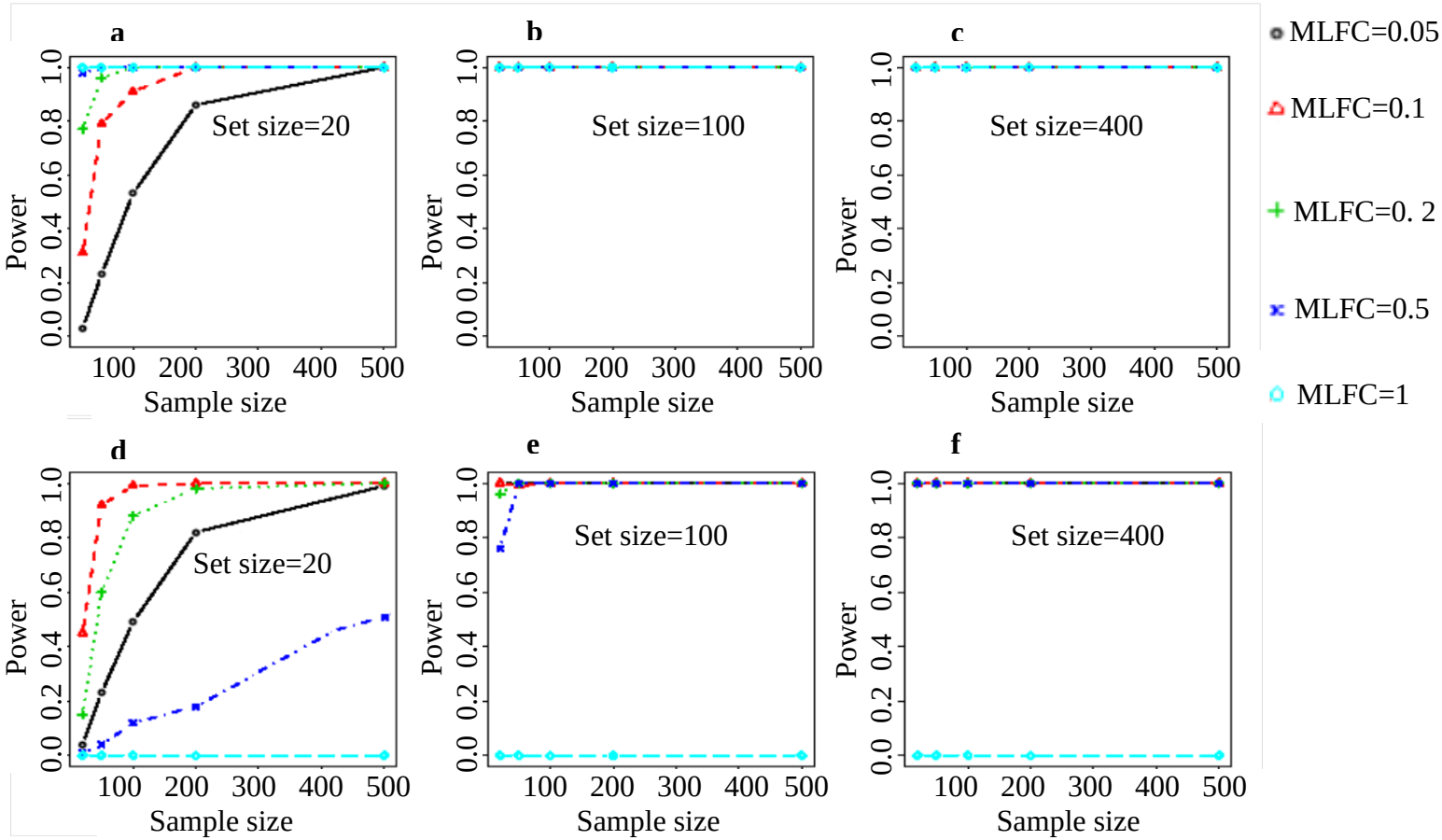

Figure 5.2. Power of gamma GAGE approach.
The simulated mean log2 fold change is for up-regulation (mean log2 fold change of genes in the gene sets across pair >0). Upper panels: gamma GAGE test for up-regulation, lower panels: gamma GAGE test for down-regulation. For each gene set size of 20, 100, 400 genes, power (proportion of significant tests out of 1000 simulations) was calculated corresponding to different crude effect sizes (mean log2 fold change of all genes in the gene set across pairs) 0.05, 0.1, 0.2, 0.5, 1. Gene set standardized effect sizes calculated from corresponding mean log2 fold change of genes in the gene set across pairs are 0.015, 0.03, 0.06, 0.15, 0.3. Conventional standardized effect sizes for gene set of 20, 100, 400 genes respectively are 0.06, 0.12, 0.24, 0.63, 1.26 and 0.14, 0.28, 0.56, 1.43, 2.86 and 0.29, 0.57, 1.14, 2.85, 5.7. a, b, c: test for up regulation; d, e, f: test for down regulation.

Pilot data of the cerebral palsy study

Permutation on randomly selected gene sets

Distribution of z-statistics. Table 5.2 and 5.3 shows the results for the tests for the normal distribution of calculated z-statistics for gene set sizes from 10 to 500 genes for three versions of the datasets when applying for PAGE, and GAGE respectively. The results for T-profiler are similar to those for PAGE. For all gene set sizes and for all pairs, the mean of the z-statistics is approximate zero and the standard deviation of the z-statistics is approximately 1. The results are similar for PAGE z-statistics and for T-profiler-z-statistics. For the first version of the dataset without missing values, one sample Kolmogorof-Smirnov (K-S) tests for the normal distribution of z-statistics for each of the 21 pairs are all non significant for gene set sizes from 20 or more, and are significant in some pairs for gene set sizes <20. For the third version of the dataset with imputed missing values, one sample K-S tests for the normal distribution of z-statistics for each of the 21 pairs are all not significant for gene set sizes from 30 or more and are significant in some pairs for gene set sizes <30. For the second version of the dataset with missing values, one sample K-S tests for the normal distribution of z-statistics for each of the 21 pairs are all not significant for gene set sizes from 60 or more and are significant in some pairs for gene set sizes <60. In other words, the z-statistics approximate the standard normal distribution for gene set sizes from 20 or more for data without missing values, for gene set sizes from 30 or more for data with imputed missing values, and for gene set sizes from 60 or more for data with large percentage of missing values. For the datasets without missing values and imputed missing values, the proportion of outliers (z-statistics <-3 or > 3) are approximately the tail probability of the standard normal distribution for gene set sizes from 50 or more, and 70 or more, respectively. For the second dataset with a large percentage of missing values of up to 70%, the proportion of

outliers (z-statistics <-3 or > 3) are approximately the tail probability of the standard normal distribution for gene set sizes from 120 or more (table 5.4 and 5.5). In other words, the shape of the distribution of the z-statistics approximate the standard normal distribution for average and large gene sets, and the distribution of z-statistics contains slightly heavy tails for gene sets of small sizes. GAGE-z-statistics follow the normal distribution better than PAGE and T-profiler z-statistics for data with missing values, especially for smaller gene set sizes.

Table 5.2. Panel of test for normal distribution of PAGE-z-statistics of random sets of genes from actual microarray data.

Gene set size	10	15	20	30	40	60	80	100	120	160	200	250	300	400	500
First version of microarray data without missing values															
K-S P-value (min;mean;max) Reference>0.05	0.004; 0.469; 0.997	0.023; 0.529; 0.978	0.123; 0.652; 0.985	0.113; 0.616; 0.949	0.094; 0.671; 0.999	0.287; 0.765; 0.998	0.359; 0.792; 0.987	0.391; 0.780; 0.997	0.196; 0.728; 0.998	0.195; 0.658; 0.964	0.296; 0.726; 0.998	0.162; 0.723; 0.981	0.231; 0.706; 0.985	0.182; 0.669; 0.996	0.255; 0.809; 0.987
Mean z-stastistics (min;mean;max) Reference=0	-0.077; -0.009; 0.048	-0.053; -0.010; 0.039	-0.075; -0.009; 0.045	-0.072; -0.015; 0.051	-0.053; -0.007; 0.046	-0.072; 0.005; 0.073	-0.069; -0.002; 0.063	-0.051; 0.007; 0.083	-0.043; -0.001; 0.036	-0.075; -0.006; 0.089	-0.078; 0.002; 0.074	-0.039; 0.003; 0.052	-0.083; -0.008; 0.032	-0.066; 0.004; 0.059	-0.042; 0.018; 0.082
SD z-stastistics (min;mean;max) Reference=1	0.954; 1.001; 1.053	0.959; 0.991; 1.068	0.958; 1.006; 1.053	0.975; 0.997; 1.032	0.959; 1.006; 1.039	0.963; 0.998; 1.028	0.970; 1.001; 1.037	0.958; 1.006; 1.053	0.962; 0.999; 1.038	0.962; 0.996; 1.041	0.956; 0.995; 1.040	0.949; 0.987; 1.036	0.967; 1.006; 1.050	0.981; 1.008; 1.038	0.947; 0.987; 1.036
Second version of microarray data with missing values															
K-S P-value (min;mean;max) Reference>0.05	<0.001; 0.089; 0.797	<0.001; 0.073; 0.462	<0.001; 0.087; 0.377	0.001; 0.231; 0.730	0.033; 0.325; 0.863	0.160; 0.521; 0.985	0.116; 0.603; 0.992	0.059; 0.529; 0.958	0.258; 0.687; 0.929	0.092; 0.746; 0.992	0.232; 0.651; 0.993	0.105; 0.651; 0.999	0.389; 0.791; 0.977	0.150; 0.663; 0.995	0.420; 0.766; 0.998
Mean z-stastistics (min;mean;max) Reference=0	-0.049; -0.011; 0.033	-0.036; -0.002; 0.039	-0.049; 0.002; 0.044	-0.044; -0.004; 0.038	-0.028; 0.001; 0.036	-0.021; 0.015; 0.051	-0.035; -0.005; 0.039	-0.031; 0.004; 0.042	-0.032; -0.001; 0.037	-0.033; 0.001; 0.053	-0.054; -0.002; 0.054	-0.052; -0.001; 0.064	-0.043; 0.003; 0.070	-0.066; -0.012; 0.057	-0.035; 0.006; 0.039
SD z-stastistics (min;mean;max) Reference=1	0.959; 1.002; 1.046	0.959; 0.999; 1.043	0.962; 1.000; 1.043	0.974; 1.007; 1.052	0.971; 0.994; 1.002	0.966; 0.998; 1.024	0.979; 1.000; 1.022	0.981; 1.003; 1.013	0.970; 0.996; 1.024	0.970; 0.990; 1.028	0.958; 0.992; 1.039	0.970; 0.992; 1.015	0.953; 0.991; 1.030	0.966; 0.991; 1.032	0.970; 0.988; 1.010

Table 5.2. (cont'd)

Gene set size	10	15	20	30	40	60	80	100	120	160	200	250	300	400	500
Third version of microarray data with imputed missing values															
K-S P-value (min;mean;max) Reference>0.05	0.004; 0.433; 0.985	0.048; 0.359; 0.896	0.046; 0.529; 0.973	0.057; 0.656; 0.998	0.098; 0.643; 0.950	0.111; 0.648; 0.999	0.089; 0.742; 0.986	0.351; 0.789; 0.992	0.266; 0.676; 0.969	0.411; 0.764; 0.998	0.305; 0.735; 0.995	0.433; 0.758; 0.998	0.254; 0.776; 0.999	0.235; 0.753; 0.994	0.145; 0.829; 0.997
Mean z-stastistics (min;mean;max) Reference=0	-0.081; -0.015; 0.044	-0.051; 0.004; 0.063	-0.057; -0.005; 0.068	-0.052; 0.001; 0.076	-0.038; -0.001; 0.044	-0.059; 0.007; 0.056	-0.064; -0.015; 0.024	-0.053; 0.001; 0.050	-0.053; -0.001; 0.038	-0.044; 0.004; 0.092	-0.067; 0.014; 0.063	-0.059; 0.001; 0.041	-0.067; -0.005; 0.065	-0.062; -0.008; 0.047	-0.058; 0.013; 0.060
SD z-stastistics (min;mean;max) Reference=1	0.940; 0.984; 1.055	0.968; 0.999; 1.045	0.963; 1.005; 1.055	0.971; 1.009; 1.030	0.964; 0.993; 1.045	0.971; 0.998; 1.032	0.956; 1.001; 1.046	0.940; 0.994; 1.050	0.951; 0.991; 1.048	0.963; 0.996; 1.035	0.925; 0.984; 1.032	0.949; 0.999; 1.062	0.930; 0.993; 1.052	0.927; 0.980; 1.008	0.958; 0.986; 1.038

Table 5.3. Panel of test for normal distribution of GAGE-z-statistics of random sets of genes from actual microarray data.

Gene set size	10	15	20	30	40	60	80	100	120	160	200	250	300	400	500
First version of microarray data without missing values															
K-S P-value (min;mean;max) Reference>0.05	0.004; 0.469; 0.997	0.023; 0.529; 0.978	0.123; 0.652; 0.985	0.113; 0.616; 0.949	0.094; 0.671; 0.999	0.287; 0.765; 0.998	0.359; 0.792; 0.987	0.391; 0.780; 0.997	0.196; 0.728; 0.998	0.195; 0.658; 0.964	0.296; 0.726; 0.998	0.162; 0.723; 0.981	0.231; 0.706; 0.985	0.182; 0.669; 0.996	0.255; 0.809; 0.987
Mean z-stastistics (min;mean;max) Reference=0	-0.077; -0.009; 0.048	-0.053; -0.010; 0.039	-0.075; -0.009; 0.045	-0.072; -0.015; 0.051	-0.053; -0.007; 0.046	-0.072; 0.005; 0.073	-0.069; -0.002; 0.063	-0.051; 0.007; 0.083	-0.043; -0.001; 0.036	-0.075; -0.006; 0.089	-0.078; 0.002; 0.074	-0.039; 0.003; 0.052	-0.083; -0.008; 0.032	-0.066; 0.004; 0.059	-0.042; 0.018; 0.082
SD z-stastistics (min;mean;max) Reference=1	0.954; 1.001; 1.053	0.959; 0.991; 1.068	0.958; 1.006; 1.053	0.975; 0.997; 1.032	0.959; 1.006; 1.039	0.963; 0.998; 1.028	0.970; 1.001; 1.037	0.958; 1.006; 1.053	0.962; 0.999; 1.038	0.962; 0.996; 1.041	0.956; 0.995; 1.040	0.949; 0.987; 1.036	0.967; 1.006; 1.050	0.981; 1.008; 1.038	0.947; 0.987; 1.036
Second version of microarray data with missing values															
K-S P-value (min;mean;max) Reference>0.05	0.075; 0.552; 0.897	0.018; 0.573; 0.998	0.124; 0.687; 0.997	0.171; 0.731; 0.730	0.193; 0.825; 0.863	0.126; 0.767; 0.985	0.196; 0.793; 0.992	0.112; 0.705; 0.998	0.209; 0.752; 0.996	0.392; 0.746; 0.992	0.232; 0.651; 0.993	0.346; 0.821; 0.999	0.389; 0.791; 0.977	0.150; 0.663; 0.995	0.420; 0.766; 0.998
Mean z-stastistics (min;mean;max) Reference=0	-0.049; -0.011; 0.033	-0.036; -0.002; 0.039	-0.049; 0.002; 0.044	-0.044; -0.004; 0.038	-0.028; 0.001; 0.036	-0.021; 0.015; 0.051	-0.035; -0.005; 0.039	-0.031; 0.004; 0.042	-0.052; 0.003; 0.047	-0.033; 0.001; 0.053	-0.054; -0.002; 0.054	-0.052; -0.001; 0.064	-0.043; 0.003; 0.070	-0.066; -0.012; 0.057	-0.035; 0.006; 0.039
SD z-stastistics (min;mean;max) Reference=1	0.959; 1.002; 1.046	0.959; 0.999; 1.043	0.962; 1.000; 1.043	0.974; 1.007; 1.052	0.971; 0.994; 1.002	0.966; 0.998; 1.024	0.979; 1.000; 1.022	0.981; 1.003; 1.013	0.973; 0.999; 1.046	0.970; 0.990; 1.028	0.958; 0.992; 1.039	0.970; 0.992; 1.015	0.953; 0.991; 1.030	0.966; 0.991; 1.032	0.970; 0.988; 1.010

Table 5.3. (cont'd)

Gene set size	10	15	20	30	40	60	80	100	120	160	200	250	300	400	500
Third version of microarray data with imputed missing values															
K-S P-value (min;mean;max) Reference>0.05	0.004; 0.433; 0.985	0.048; 0.359; 0.896	0.056; 0.529; 0.973	0.057; 0.656; 0.998	0.098; 0.643; 0.950	0.111; 0.648; 0.999	0.089; 0.742; 0.986	0.351; 0.789; 0.992	0.333; 0.779; 0.999	0.411; 0.764; 0.998	0.305; 0.735; 0.995	0.433; 0.758; 0.998	0.254; 0.776; 0.999	0.235; 0.753; 0.994	0.145; 0.829; 0.997
Mean z-stastistics (min;mean;max) Reference=0	-0.081; -0.015; 0.044	-0.051; 0.004; 0.063	-0.057; -0.005; 0.068	-0.052; 0.001; 0.076	-0.038; -0.001; 0.044	-0.059; 0.007; 0.056	-0.064; -0.015; 0.024	-0.053; 0.001; 0.050	-0.059; 0.004; 0.079	-0.044; 0.004; 0.092	-0.067; 0.014; 0.063	-0.059; 0.001; 0.041	-0.067; -0.005; 0.065	-0.062; -0.008; 0.047	-0.058; 0.013; 0.060
SD z-stastistics (min;mean;max) Reference=1	0.940; 0.984; 1.055	0.968; 0.999; 1.045	0.963; 1.005; 1.055	0.971; 1.009; 1.030	0.964; 0.993; 1.045	0.971; 0.998; 1.032	0.956; 1.001; 1.046	0.940; 0.994; 1.050	0.960; 1.000; 1.061	0.963; 0.996; 1.035	0.925; 0.984; 1.032	0.949; 0.999; 1.062	0.930; 0.993; 1.052	0.927; 0.980; 1.008	0.958; 0.986; 1.038

Mean, SD of z-statistics and Kolmogorof-Smirnov p-value were calculated from 1000 samples for each gene sets for each pairs. Mean(min;max) of these values of 21 pairs were then calculated.

Table 5.4. Panel of proportion of outliers of individual pair PAGE-z-statistics of random sets of genes.

Gene set size	10	15	20	30	40	60	80	100	120	160	200	300	500
First version of microarray data without missing values													
Reference# =0.00269	0.00576	0.00366	0.00452	0.00414	0.00376	0.00285	0.00276	0.00257	0.00352	0.00281	0.00300	0.00285	0.00204
Second version of microarray data with missing values													
Reference# =0.00269	0.14459	0.06961	0.03476	0.01352	0.00690	0.00466	0.00407	0.00435	0.00364	0.00319	0.00311	0.00288	0.00276
Third version of microarray data with imputed missing values													
Reference# =0.00269	0.00557	0.00576	0.00547	0.00414	0.00366	0.00380	0.00323	0.00342	0.00376	0.00257	0.00261	0.00242	0.00209

Outliers in this context are defined as z-statistics >3 or <-3. #Reference: proportion of outliers of standard normal distribution.

Table 5.5. Panel of proportion of outliers of individual pair GAGE-z-statistics of random sets of genes.

Gene set size	10	15	20	30	40	60	80	100	120	160	200	300	500
First version of microarray data without missing values													
Reference# =0.00269	0.00554	0.00376	0.00462	0.00421	0.00357	0.00292	0.00259	0.00261	0.00349	0.00279	0.00313	0.00279	0.00214
Second version of microarray data with missing values													
Reference# =0.00269	0.09941	0.03021	0.00905	0.00084	0.00007	0.00000	0.00000	0.00000	0.00000	0.00000	0.00004	0.00006	0.00004
Third version of microarray data with imputed missing values													
Reference# =0.00269	0.00549	0.00536	0.00537	0.00422	0.00359	0.00382	0.00318	0.00341	0.00368	0.00256	0.00251	0.00245	0.00206

Outliers in this context are defined as z-statistics >3 or <-3. #Reference: proportion of outliers of standard normal distribution.

Type I error. The results of type I error for the first and the third version of the dataset are shown in table 5.6 and 5.7 respectively. For the second version of the dataset with missing values, the results are similar to those of the first version of the dataset without missing values. These results show that type I error of PAGE-ZZ-GSA and t-profiler- ZZ-GSA is less than or equal to random chance. The false positive rate produced by the simple chi-square test of the ZZ-GSA is a bit larger than random chance, but <0.05 when the p-value for significance is set at <0.01. Thus for simple chi- square tests, the p-value should be <0.01 to claim significance. The proportion of false positive of standardized chi-square test is equal to zero for gene set sizes larger than 10. The proportion of false positives of the gamma GAGE and Stouffer- GAGE tests are less than chance for the first version of dataset without missing values, and for the third version of the dataset with imputed missing values. Thus these tests are more conservative in picking up false positives.

For tests for bidirectional gene sets, the type I error of ZZ-GSA when applied to PAGE, T-profiler, and GAGE is considerably higher than random chance. The type I error is <0.05 when the p-value for significance is set at <0.005 for the first dataset without missing values and for the second dataset with missing values. The type I error is <0.05 when the p-value for significance is set at <0.001 for the third dataset with imputed missing values. Since there is a large percentage of genes with log2 fold change values of zero in the dataset with imputed missing values, this produces more false positives for the test for bidirectional gene sets, where the mean of absolute values of log2 fold change of expression values of genes in the set is compared with the mean of the absolute values of log2 fold change of expression values of all genes. The type I error for the test of bidirectional gene set of gamma GAGE is smaller than random chance.

Table 5.6. Panel of type I error of random sets of genes for microarray data without missing values.

Gene set size	Reference#	10	20	30	40	60	80	100	120	160	200	250	300	400	500
ZZ-GSA															
Z-test Up	0.010; 0.050; 0.100	0.004; 0.027; 0.056	0.006; 0.027; 0.065	0.005; 0.028; 0.069	0.004; 0.023; 0.063	0.005; 0.031; 0.062	0.006; 0.033; 0.070	0.004; 0.030; 0.068	0.003; 0.034; 0.078	0.003; 0.026; 0.068	0.002; 0.024; 0.072	0.004; 0.032; 0.075	0.000; 0.023; 0.062	0.005; 0.031; 0.075	0.003; 0.032; 0.076
Z-test Down	0.010; 0.050; 0.100	0.006; 0.027; 0.078	0.005; 0.034; 0.072	0.003; 0.033; 0.079	0.003; 0.027; 0.075	0.002; 0.022; 0.059	0.004; 0.026; 0.070	0.007; 0.046; 0.079	0.007; 0.029; 0.072	0.005; 0.032; 0.082	0.003; 0.022; 0.061	0.003; 0.029; 0.075	0.005; 0.030; 0.073	0.006; 0.033; 0.058	0.001; 0.018; 0.059
T-profiler Up	0.010; 0.050; 0.100	0.002; 0.025; 0.063	0.001; 0.028; 0.072	0.002; 0.031; 0.07	0.005; 0.026; 0.063	0.002; 0.027; 0.066	0.006; 0.035; 0.072;	0.005; 0.030; 0.072	0.005; 0.033; 0.083	0.004; 0.038; 0.082	0.001; 0.029; 0.068	0.004; 0.041; 0.084	0.003; 0.025; 0.06	0.007; 0.033; 0.084	0.003; 0.034; 0.083
T-profiler Down	0.010; 0.050; 0.100	0.002; 0.023; 0.074	0.005; 0.030; 0.076	0.006; 0.032; 0.080	0.001; 0.037; 0.078	0.001; 0.026; 0.058	0.003; 0.032; 0.070	0.008; 0.041; 0.084	0.004; 0.029; 0.070	0.007; 0.033; 0.095	0.002; 0.028; 0.061	0.005; 0.035; 0.074	0.005; 0.039; 0.078	0.007; 0.038; 0.076	0.001; 0.022; 0.059
Chi-square	0.010; 0.050; 0.100	0.051; 0.120; 0.171	0.048; 0.101; 0.141	0.048; 0.098; 0.140	0.041; 0.112; 0.158	0.042; 0.090; 0.138	0.030; 0.082; 0.142	0.033; 0.099; 0.155	0.039; 0.097; 0.142	0.036; 0.095; 0.143	0.038; 0.084; 0.131	0.030; 0.080; 0.124	0.038; 0.101; 0.148	0.028; 0.086; 0.128	0.032; 0.080; 0.126
Standardized chi-square	0.010; 0.050; 0.10	0.000; 0.001; 0.001	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0
Bi-direction*	0.010; 0.050; 0.10	0.088; 0.148; 0.197	0.091; 0.172; 0.225	0.084; 0.152; 0.219	0.093; 0.174; 0.226	0.086; 0.169; 0.213	0.095; 0.178; 0.245	0.078; 0.146; 0.206	0.089; 0.161; 0.212	0.087; 0.153; 0.214	0.068; 0.149; 0.200	0.081; 0.170; 0.223	0.087; 0.180; 0.249	0.078; 0.160; 0.222	0.075; 0.152; 0.207

Table 5.6. (cont'd)

Gene set size	Reference#	10	20	30	40	60	80	100	120	160	200	250	300	400	500
GAGE															
Gamma up	0.010; 0.050; 0.10	0.001; 0.001; 0.004	0; 0.001; 0.004	0; 0; 0	0; 0; 0	0; 0; 0	0.001; 0.001; 0.003	0; 0.001; 0.002	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0.001; 0.004; 0.004	0; 0.001; 0.002
Gamma down	0.010; 0.050; 0.100	0; 0; 0.001	0; 0; 0	0.001; 0.001; 0.002	0; 0.001; 0.001	0.001; 0.002; 0.004	0; 0.002; 0.002	0; 0.001; 0.001	0; 0; 0	0; 0; 0	0.001; 0.002; 0.003	0; 0; 0	0.001; 0.004; 0.004	0; 0.001; 0.002	0; 0.001; 0.002
Bi-direction	0.010; 0.050; 0.100	0; 0.010; 0.21	0; 0.021; 0.043	0; 0.028; 0.045	0; 0.031; 0.041	0; 0.039; 0.056	0; 0.049; 0.063	0; 0.046; 0.068	0; 0.051; 0.074	0; 0.046; 0.065	0; 0.045; 0.069	0; 0.046; 0.071	0; 0.047; 0.061	0; 0.040; 0.064	0; 0.052; 0.080
Stouffer up$	0.010; 0.050; 0.100	0.001; 0.003; 0.028	0; 0.003; 0.018	0.001; 0.006; 0.022	0; 0.004; 0.017	0; 0.003; 0.025	0; 0.013; 0.016	0.001; 0.002; 0.021	0; 0.005; 0.025	0; 0.003; 0.022	0; 0.009; 0.032	0; 0.007; 0.031	0; 0.006; 0.021	0; 0.004; 0.027	0; 0.005; 0.020
Stouffer down$	0.010; 0.050; 0.100	0; 0.007; 0.018	0; 0.007; 0.022	0; 0.004; 0.019	0; 0.005; 0.022	0; 0.005; 0.029	0; 0.005; 0.022	0.001; 0.005; 0.027	0.001; 0.010; 0.022	0; 0.005; 0.029	0; 0.005; 0.033	0; 0.009; 0.022	0; 0.005; 0.040	0; 0.003; 0.023	0; 0.009; 0.029
Bi-direction$	0.010; 0.050; 0.100	0.034; 0.083; 0.115	0.062; 0.107; 0.135	0.068; 0.121; 0.136	0.061; 0.124; 0.139	0.073; 0.129; 0.171	0.079; 0.148; 0.181	0.099; 0.173; 0.181	0.098; 0.163; 0.186	0.090; 0.159; 0.186	0.094; 0.168; 0.218	0.093; 0.174; 0.189	0.097; 0.171; 0.225	0.096; 0.171; 0.218	0.112; 0.170; 0.217
Bi-direction	0.010; 0.050; 0.100	0; 0.010; 0.21	0; 0.021; 0.043	0; 0.028; 0.045	0; 0.031; 0.041	0; 0.039; 0.056	0; 0.049; 0.063	0; 0.046; 0.068	0; 0.051; 0.074	0; 0.046; 0.065	0; 0.045; 0.069	0; 0.046; 0.071	0; 0.047; 0.061	0; 0.040; 0.064	0; 0.052; 0.080
Stouffer up$	0.010; 0.050; 0.100	0.001; 0.003; 0.028	0; 0.003; 0.018	0.001; 0.006; 0.022	0; 0.004; 0.017	0; 0.003; 0.025	0; 0.013; 0.016	0.001; 0.002; 0.021	0; 0.005; 0.025	0; 0.003; 0.022	0; 0.009; 0.032	0; 0.007; 0.031	0; 0.006; 0.021	0; 0.004; 0.027	0; 0.005; 0.020

#Reference: expected proportion of false positive (type I error) corresponding to significant level of p-value set at 0.01, 0.05, 0.1.
*proportion of significant test for all gene set sizes are <0.05 for p-value set at 0.001.

Table 5.7. Panel of type I error of random sets of genes for microarray data with imputed missing values.

Gene set size	Reference#	10	20	30	40	60	80	100	120	160	200	250	300	400	500
ZZ-GSA															
Z-test Up	0.010; 0.050; 0.100	0.004; 0.025; 0.063	0.006; 0.033; 0.077	0.005; 0.045; 0.087	0.009; 0.035; 0.071	0.007; 0.039; 0.079	0.009; 0.036; 0.083	0.011; 0.042; 0.081	0.007; 0.039; 0.097	0.008; 0.039; 0.091	0.007; 0.047; 0.095	0.006; 0.037; 0.092	0.011; 0.040; 0.087	0.004; 0.029; 0.064	0.007; 0.044; 0.104
Z-test Down	0.010; 0.050; 0.100	0.011; 0.049; 0.100	0.007; 0.049; 0.098	0.007; 0.039; 0.092	0.002; 0.036; 0.083	0.004; 0.030; 0.058	0.006; 0.050; 0.093	0.005; 0.046; 0.078	0.006; 0.029; 0.098	0.008; 0.032; 0.082	0.007; 0.022; 0.061	0.010; 0.029; 0.084	0.006; 0.030; 0.081	0.006; 0.033; 0.086	0.010; 0.018; 0.074
T-profiler Up	0.010; 0.050; 0.100	0.004; 0.035; 0.075	0.004; 0.043; 0.085	0.008; 0.045; 0.086	0.008; 0.037; 0.088	0.006; 0.037; 0.083	0.004; 0.038; 0.081;	0.008; 0.044; 0.086	0.004; 0.042; 0.094	0.006; 0.044; 0.087	0.012; 0.045; 0.100	0.007; 0.039; 0.096	0.007; 0.042; 0.086	0.005; 0.036; 0.069	0.009; 0.053; 0.107
T-profiler Down	0.010; 0.050; 0.100	0.010; 0.043; 0.108	0.004; 0.044; 0.105	0.003; 0.043; 0.090	0.005; 0.041; 0.075	0.006; 0.033; 0.065	0.006; 0.052; 0.096	0.006; 0.038; 0.078	0.007; 0.040; 0.093	0.007; 0.040; 0.079	0.009; 0.040; 0.089	0.007; 0.044; 0.088	0.007; 0.033; 0.083	0.008; 0.056; 0.095	0.010; 0.036; 0.085
Chi-square	0.010; 0.050; 0.100	0.038; 0.094; 0.131	0.044; 0.106; 0.155	0.035; 0.088; 0.146	0.032; 0.074; 0.125	0.032; 0.076; 0.127	0.023; 0.078; 0.136	0.026; 0.070; 0.114	0.031; 0.079; 0.109	0.018; 0.068; 0.094	0.028; 0.057; 0.127	0.018; 0.071; 0.124	0.022; 0.074; 0.148	0.014; 0.053; 0.093	0.012; 0.063; 0.099
Standardized chi-square	0.010; 0.050; 0.10	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0
Bi-direction*	0.010; 0.050; 0.10	0.114; 0.21; 0.243	0.115; 0.200; 0.249	0.114; 0.196; 0.246	0.115; 0.179; 0.237	0.112; 0.194; 0.230	0.113; 0.187; 0.230	0.120; 0.199; 0.245	0.110; 0.189; 0.243	0.121; 0.194; 0.238	0.093; 0.217; 0.273	0.081; 0.186; 0.247	0.087; 0.185; 0.257	0.107; 0.194; 0.251	0.105; 0.188; 0.241

Table 5.7. (cont'd)

Gene set size	Reference#	10	20	30	40	60	80	100	120	160	200	250	300	400	500
GAGE															
Gamma up	0.010; 0.050; 0.10	0.001; 0.001; 0.004	0; 0.001; 0.004	0; 0; 0	0; 0; 0	0; 0; 0	0.001; 0.001; 0.003	0; 0.001; 0.002	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0; 0; 0	0.001; 0.004; 0.004	0; 0.001; 0.002
Gamma down	0.010; 0.050; 0.100	0; 0; 0.001	0; 0; 0	0.001; 0.001; 0.002	0; 0.001; 0.001	0.001; 0.002; 0.004	0; 0.002; 0.002	0; 0.001; 0.001	0; 0; 0	0; 0; 0	0.001; 0.002; 0.003	0; 0; 0	0.001; 0.004; 0.004	0; 0.001; 0.002	0; 0.001; 0.002
Bi-direction	0.010; 0.050; 0.100	0; 0.010; 0.21	0; 0.021; 0.043	0; 0.028; 0.045	0; 0.031; 0.041	0; 0.039; 0.056	0; 0.049; 0.063	0; 0.046; 0.068	0; 0.051; 0.074	0; 0.046; 0.065	0; 0.045; 0.069	0; 0.046; 0.071	0; 0.047; 0.061	0; 0.040; 0.064	0; 0.052; 0.080
Stouffer up$	0.010; 0.050; 0.100	0.001; 0.003; 0.028	0; 0.003; 0.018	0.001; 0.006; 0.022	0; 0.004; 0.017	0; 0.003; 0.025	0; 0.013; 0.016	0.001; 0.002; 0.021	0; 0.005; 0.025	0; 0.003; 0.022	0; 0.009; 0.032	0; 0.007; 0.031	0; 0.006; 0.021	0; 0.004; 0.027	0; 0.005; 0.020
Stouffer down$	0.010; 0.050; 0.100	0; 0.007; 0.018	0; 0.007; 0.022	0; 0.004; 0.019	0; 0.005; 0.022	0; 0.005; 0.029	0; 0.005; 0.022	0.001; 0.005; 0.027	0.001; 0.010; 0.022	0; 0.005; 0.029	0; 0.005; 0.033	0; 0.009; 0.022	0; 0.005; 0.040	0; 0.003; 0.023	0; 0.009; 0.029
Bi-direction$	0.010; 0.050; 0.100	0.034; 0.083; 0.115	0.062; 0.107; 0.135	0.068; 0.121; 0.136	0.061; 0.124; 0.139	0.073; 0.129; 0.171	0.079; 0.148; 0.181	0.099; 0.173; 0.181	0.098; 0.163; 0.186	0.090; 0.159; 0.186	0.094; 0.168; 0.218	0.093; 0.174; 0.189	0.097; 0.171; 0.225	0.096; 0.171; 0.218	0.112; 0.170; 0.217
Bi-direction	0.010; 0.050; 0.100	0; 0.010; 0.21	0; 0.021; 0.043	0; 0.028; 0.045	0; 0.031; 0.041	0; 0.039; 0.056	0; 0.049; 0.063	0; 0.046; 0.068	0; 0.051; 0.074	0; 0.046; 0.065	0; 0.045; 0.069	0; 0.046; 0.071	0; 0.047; 0.061	0; 0.040; 0.064	0; 0.052; 0.080
Stouffer up$	0.010; 0.050; 0.100	0.001; 0.003; 0.028	0; 0.003; 0.018	0.001; 0.006; 0.022	0; 0.004; 0.017	0; 0.003; 0.025	0; 0.013; 0.016	0.001; 0.002; 0.021	0; 0.005; 0.025	0; 0.003; 0.022	0; 0.009; 0.032	0; 0.007; 0.031	0; 0.006; 0.021	0; 0.004; 0.027	0; 0.005; 0.020

#Reference: expected proportion of false positive (type I error) corresponding to significant level of p-value set at 0.01, 0.05, 0.1.
*proportion of significant test for all gene set sizes are <0.05 for p-value set at 0.001.

KEGG gene sets

The analysis results of 205 KEGG gene sets for 3 versions of the data are shown in table 5.8, 5.9, 5.10 respectively. For the first version of the dataset without missing values and the third version of the dataset where missing values were imputed, the Lupus gene set is picked up as significantly up-regulated by both ZZ-GSA and GAGE. Ribosome is picked up as heterogeneously differentially expressed across pairs by ZZ-GSA, picked up as both significantly up-regulated and down-regulated by gamma GAGE, and not picked up by Stouffer GAGE. Although the Ribosome gene set remains the most perturbed gene sets by the chi-square test of ZZ-GSA and by gamma GAGE for both up and down regulation, fewer gene sets are picked up as significantly differentially expressed (before adjusting for multiple testing) in the second version of the dataset with missing values than in the first and the third version of the dataset. This suggests that missing values may cause loss of important expression data especially for the analysis method using ratio where missing only in either cases or controls can cause missing values for the ratio between cases and controls. For the third dataset where missing values were imputed as the smallest expression values, the log2 fold change data contain large percentages of the same value of zero. As a result, larger false positives may be produced for the test for bidirectional gene sets. The results presented are before adjusting for multiple testing. After adjusting for multiple testing using the q-value R package (cite Storey FDR 2002), the Ribosome gene set remains significant consistently in the three versions of the dataset and the Lupus gene set remains significant in the first and third version of the dataset.

Table 5.8. Test for significance of KEGG gene sets for microarray data without missing values.

Gene sets	PAGE ZZ-GSA	T profiler ZZ-GSA	Chi square	S chi square	Bi direction	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
Up									
SLE	0.00016		8.4e-18	1.4e-02	0.01150	8.7e-05	7.6e-02	0.00814	>0.1
PCM	0.03142		>0.1	>0.1	0.00849	>0.1	>0.1	>0.1	>0.1
MODY	0.04918		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
GM	0.08776		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Down									
TRSP	0.00525		1.5e-04	>0.1	>0.1	2.3e-02	>0.1	0.02919	>0.1
LTM	0.00848		4.7e-13	4.9e-02	>0.1	1.7e-03	>0.1	0.05248	>0.1
RAC	0.00993		2.1e-10	>0.1	>0.1	2.3e-03	>0.1	0.04232	>0.1
AtRB	0.01122		8.6e-02	>0.1	>0.1	>0.05	>0.1	0.05470	>0.1
PPP	0.01195		4.3e-06	>0.1	0.03901	5.5e-02	>0.1	0.02107	>0.1
Apoptosis	0.01289		2.3e-09	>0.1	>0.1	4.6e-03	>0.1	0.05639	>0.1
CCRI	0.01449		5.0e-02	>0.1	>0.1	>0.1	>0.1	0.05594	>0.1
PCB	0.02795		>0.1	>0.1	>0.1	>0.1	>0.1	0.09409	>0.1
FAM	0.04611		>0.1	>0.1	>0.1	>0.1	>0.1	0.09934	>0.1
UCMAG	0.07030		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Heterogeneity in differential expression across pairs									
Ribosome	>0.1	>0.1	5.1e-108	1.5e-43	0.00000	***	2.9e-17	>0.1	2.6e-10
NKCMC	>0.1	>0.1	2.6e-14	2.4e-02	>0.1	6.3e-03**	>0.1	>0.1	>0.1
OT	>0.1	>0.1	1.5e-10	8.5e-02	0.01228	0.01012*	2.4e-02	>0.1	8.4e-02
PD	>0.1	>0.1	1.9e-09	>0.1	0.02252	2.2e-02 **	>0.1	>0.1	>0.1
OP	>0.1	>0.1	3.0e-09	>0.1	0.04498	2.4e-02**	>0.1	>0.1	>0.1
BCRSP	>0.1	>0.1	4.05e-09	>0.1	>0.1	5.7e-02**	>0.1	>0.1	>0.1

Table 5.8. (cont'd)

Gene sets	PAGE ZZ-GSA	T profiler ZZ-GSA	Chi square	S chi square	Bi direction	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
EHEC	>0.1	>0.1	1.4e-08	>0.1	0.00572	>0.1	9.4e-02	>0.1	>0.1
EPEC	>0.1	>0.1	1.4e-08	>0.1	0.00572	>0.1	9.4e-02	>0.1	>0.1
Bi-directional gene set									
NM	>0.1	>0.1	>0.1	>0.1	0.00849	>0.1	>0.1	>0.1	8.9e-02
FAEM	>0.1	>0.1	>0.1	>0.1	0.01036	>0.1	>0.1	>0.1	>0.1

SLE: Systemic lupus erythematosus, PCM: Porphyrin and chlorophyll metabolism, MODY: Maturity onset diabetes of the young, GM: Glutathione metabolism, TRSP: Toll-like receptor signaling pathway, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, AtRB: Aminoacyl-tRNA biosynthesis, PPP: Pentose phosphate pathway, CCRI: Cytokine-cytokine receptor interaction, PCB: Pantothenate and CoA biosynthesis, FAM: Fatty acid metabolism, UCMAG: Urea cycle and metabolism of amino groups, NKCMC: Natural killer cell mediated cytotoxicity, OT: Olfactory transduction, PD: Parkinson's disease, OP: Oxidative phosphorylation, BCRSP: B cell receptor signaling pathway, EHEC: Pathogenic Escherichia coli infection – EHEC, EPEC: Pathogenic Escherichia coli infection – EPEC, NM: Nitrogen metabolism, FAEM: Fatty acid elongation in mitochondria. *P-value for up regulation, **P-value for down regulation, ***Both P-value for up and down regulation are <<0.0001. #Test for bi-directional gene sets using GAGE Gamma test. $ Test for bi-directional gene sets using GAGE Stouffer test.

Table 5.9. Test for significance of KEGG gene sets for microarray data with missing values.

Gene sets	PAGE ZZGSA	T profiler ZZ-GSA	Chi square	S chi square	Bi direction	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
Up									
SLE	>0.1		6.1e-06	>0.1	>0.1	0.0933	>0.1	>0.1	>0.1
PCM	0.02106		>0.1	>0.1	0.02589	>0.1	>0.1	>0.1	0.07048
MODY	>0.1		2.2e-02	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
GM	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Down									
TRSP	0.02491		>0.1	>0.1	>0.1	>0.1	>0.1	0.0616	>0.1
LTM	0.00368		4.2e-06	>0.1	>0.1	1.05e-2	>0.1	0.0289	>0.1
RAC	0.01132		5.6e-04	>0.1	>0.1	>0.1	>0.1	0.0563	>0.1
AtRB	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
PPP	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Apoptosis	>0.1		>0.1	>0.1	>0.1	3.7e-2	>0.1	>0.1	>0.1
CCRI	0.00183		1.2e-03	>0.1	>0.1	>0.1	>0.1	0.0201	>0.1
PCB	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
FAM	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
UCMAG	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Heterogeneity in differential expression across pairs									
Ribosome	0.00190**	>0.1	7.6e-40	7.3e-08	0.00035	***	8.2e-10	0.0176	0.00224
NKCMC	0.01478**	>0.1	2.3e-02	>0.1	>0.1	>0.1	>0.1	0.0694	>0.1
OT	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	5.8e-05	>0.1	>0.1
PD	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
OP	>0.1	>0.1	5.3e-03	>0.1	>0.1	>0.1	3.7e-02	>0.1	>0.1

Table 5.9. (cont'd)

Gene sets	PAGE ZZGSA	T profiler ZZ-GSA	Chi-square	S chi square	Bi direction	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
BCRSP	0.04286**	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
EHEC	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
EPEC	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Bi-directional gene set									
NM	>0.1	>0.1	>0.1	>0.1	0.00514	>0.1	>0.1	>0.1	>0.1
FAEM	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
ASP@	>0.1		>0.1	>0.1	0.00011	1.3e-02*	>0.1	>0.1	0.00165

SLE: Systemic lupus erythematosus, PCM: Porphyrin and chlorophyll metabolism, MODY: Maturity onset diabetes of the young, GM: Glutathione metabolism, TRSP: Toll-like receptor signaling pathway, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, AtRB: Aminoacyl-tRNA biosynthesis, PPP: Pentose phosphate pathway, CCRI: Cytokine-cytokine receptor interaction, PCB: Pantothenate and CoA biosynthesis, FAM: Fatty acid metabolism, UCMAG: Urea cycle and metabolism of amino groups, NKCMC: Natural killer cell mediated cytotoxicity, OT: Olfactory transduction, PD: Parkinson's disease, OP: Oxidative phosphorylation, BCRSP: B cell receptor signaling pathway, EHEC: Pathogenic Escherichia coli infection – EHEC, EPEC: Pathogenic Escherichia coli infection – EPEC, NM: Nitrogen metabolism, FAEM: Fatty acid elongation in mitochondria, ASP: Adipocytokine signaling pathway. *P-value for up regulation, **P-value for down regulation, ***Both P-value for up and down regulation are <<0.0001. #Test for bi-directional gene sets using GAGE Gamma test. $ Test for bi-directional gene sets using GAGE Stouffer test. @Gene set are not significant in the dataset without missing values.

Table 5.10. Test for significance of KEGG gene sets for microarray data with imputed missing values.

Gene sets	PAGE ZZ-GSA	T profilerZZ-GSA	Chi-square	S chi square	Bi direction @@	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
Up									
SLE	2.8e-05		1.7e-15	4.1e-02	2.0e-10	1.1e-04	2.1e-05	0.00355	>0.1
PCM	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
MODY	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
GM	5.6e-02		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Down									
TRSP	0.00184		>0.1	>0.1	7.7e-11	>0.1	>0.1	0.02340	>0.1
LTM	0.00307		1.9e-12	6.6e-02	0.00000	1.8e-03	2.2e-07	0.03764	8.3e-09
RAC	0.00773		2.8e-09	>0.1	4.6e-14	9.8e-03	2.9e-06	0.04906	1.6e-07
AtRB	0.00344		>0.1	>0.1	>0.1	>0.1	>0.1	0.03221	>0.1
PPP	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Apoptosis	0.00799		1.6e-07	>0.1	0.00000	2.2e-02	4.0e-07	0.05797	7.4e-09
CCRI	0.00010		2.9e-08	>0.1	1.6e-08	>0.1	>0.1	0.00464	>0.1
PCB	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
FAM	>0.1		>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
UCMAG	0.00707		>0.1	>0.1	>0.1	>0.1	>0.1	0.05388	>0.1
CSP@	0.00164		>0.1	>0.1	>0.1	>0.1	1.5e-05	0.02710	3.5e-07
Heterogeneity in differential expression across pairs									
Ribosome	>0.1	>0.1	1.8e-119	9.7e-52	0.00000	***	8.7e-34	>0.1	4.5e-28
NKCMC	0.00191**	>0.1	6.9e-12	8.2e-02	0.00000	2.7e-03 **	1.2e-08	0.03027**	5.7e-10
OT	2.0e-02*	>0.1	1.2e-12	5.1e-02	>0.1	6.5e-05 *		0.04591*	
PD	>0.1	>0.1	8.2e-08	>0.1	1.9e-13	2.2e-02 **	5.8e-08	>0.1	>0.1

Table 5.10. (cont'd)

Gene sets	PAGE ZZ-GSA	T profiler ZZ-GSA	Chi square	S chi square	Bi direction @@	GAGE Gamma	Bi direction#	GAGE Stouffer	Bi direction$
OP	>0.1	>0.1	2.8e-08	>0.1	3.1e-06	2.4e-02**		>0.1	>0.1
BCRSP	>0.1	>0.1	1.7e-06	>0.1	0.00000	5.0e-02**	2.5e-08	>0.1	1.7e-10
EHEC	>0.1	>0.1	6.6e-06	>0.1	0.00000	>0.1	4.0e-09	>0.1	2.0e-11
EPEC	>0.1	>0.1	6.6e-06	>0.1	0.00000	>0.1	4.0e-09	>0.1	2.0e-11
Bi-directional gene set									
NM	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	
FAEM	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
MAPKSP@	>0.1	>0.1	>0.1	>0.1	0.00000	>0.1	1.8e-11	>0.1	1.4e-14
CC@	>0.1	>0.1	>0.1	>0.1	0.00000	>0.1	9.2e-09	>0.1	6.2e-11

SLE: Systemic lupus erythematosus, PCM: Porphyrin and chlorophyll metabolism, MODY: Maturity onset diabetes of the young, GM: Glutathione metabolism, TRSP: Toll-like receptor signaling pathway, LTM: Leukocyte transendothelial migration, RAC: Regulation of actin cytoskeleton, AtRB: Aminoacyl-tRNA biosynthesis, PPP: Pentose phosphate pathway, CCRI: Cytokine-cytokine receptor interaction, PCB: Pantothenate and CoA biosynthesis, FAM: Fatty acid metabolism, UCMAG: Urea cycle and metabolism of amino groups, CSP: Calcium signaling pathway, NKCMC: Natural killer cell mediated cytotoxicity, OT: Olfactory transduction, PD: Parkinson's disease, OP: Oxidative phosphorylation, BCRSP: B cell receptor signaling pathway, EHEC: Pathogenic Escherichia coli infection – EHEC, EPEC: Pathogenic Escherichia coli infection – EPEC, NM: Nitrogen metabolism, FAEM: Fatty acid elongation in mitochondria, MAPKSP: MAPK signaling pathway, CC: Cell cycle. *P-value for up regulation, **P-value for down regulation, ***Both P-value for up and down regulation are <<0.0001. #Test for bi-directional gene sets using GAGE Gamma test. $ Test for bi-directional gene sets using GAGE Stouffer test. @Gene sets that are not significant in the datasets without missing values. @@More than 20 pathways are extremely significant. This is more likely to be false positive due to many imputed values of zero..

Discussion

This chapter focuses heavily on simulation and permutation to test and evaluate the performance of the two-stage z-test approaches when applied to three existing methods (PAGE, GAGE and T-profiler) in comparison with the gamma GAGE method for gene set analysis of matched microarray data. The gene set analysis methods and approaches were examined for power, type I error, satisfaction of statistical assumptions, influence of missing values and influence of unusual data distributions or unusual data structure.

Both the results from purely simulated data and experiments on actual microarray datasets supplement and support each other. Expression values in our simulated datasets are independent and identical distributed. Our simulated datasets do not account for gene specific expression patterns or correlations in expression among genes. Thus, simulated datasets were mainly used for the comparison of power of different approaches, and for demonstration of some special situations. Permutation on actual microarray data can evaluate the performance of different approaches on real usual and unusual microarray data where the expression pattern can be different across genes or the correlation in expression among genes exist, or large ties are produced by a large percentage of genes having the same ratio of expression values between cases and matched controls.

Statistically, the two stage z-test approach for gene set analysis assumes that for each matched pair the log fold changes are independently and identically distributed (IID) across genes. However, on actual microarray data where the expression of genes may be correlated, and the distribution of expression of genes may not be identical, the permutation results show that the z-statistics satisfy standard normal distribution for gene set sizes of 20 genes or more. For the third version of the dataset, missing values after filtering are imputed as the smallest values of

log2 expression values and thus the log2 fold changes of each pair contain from 20% to 70% of genes with log2 fold changes at the same value of zero. The distribution of log2 fold changes in this imputed data set is very unusual, contains large ties among genes, and may not be identical across genes. However, the z-statistics still satisfy the standard normal distribution for gene set sizes from 30 or more.

In theory, the z-statistics may have some drawbacks. Firstly, since they are based on the mean, they may be sensitive to some extreme values of the log2 fold change of some genes in the gene sets, especially for gene sets of small size. This is reflected by the fact that the proportion of large z-statistics ($z<-3$ or $z>3$) is slightly higher than the expected proportion from the standard normal distribution for gene set sizes ≤ 50 and approximately equal to the expected proportion for gene set sizes > 50 for the dataset without missing values and for the dataset with imputed missing values. For the second version of the dataset, since there is a large percentage of missing values, the number of genes in the set that actually have expression values is much smaller than the gene set size itself. As a result, a considerably higher proportion of large z-statistics for small and average gene set sizes is predicted. However, the second stage z-test average one more time the z-statistics of individual pairs, thus the type I error of the global z-test is just equal to, or smaller than random chance for both simulated data and for actual microarray data with and without missing values. To avoid a higher risk of false positives due to large z-statistics of individual pairs, especially for datasets with large percentages of missing values, the panel of the proportion of large z-statistics for different gene set sizes from 10 to 500 genes (table 5.4 and 5.5) can be used to adjust the p-value of the global z-test. One possible solution is that if the observed proportion is k times larger than the expected proportion, the adjusted global p-value of global z-test of a gene set of corresponding size may be k times the observed global p-value.

Secondly, the stage 1 z-test does not account for variance within gene sets. However, the combination of the tests for one-directional gene set, and the test for bi-directional gene sets addresses the variability of expression values of genes in the set. In addition, for matched microarray data, variation of differential expression across matched pairs may be of more interest than variation of expression ratio across genes in the set. The variation of differential expression across matched pairs can be partially addressed by the combined use of a global z-test, a simple chi-square test and a standardized chi-square test.

The combination of the three stage 2 global tests: z-test, simple chi-square test and standardized chi-square test can detect both homogeneity and heterogeneity in differential expression across pairs. The global z-test can only detect differential expression in the direction that is more predominant across pairs (homogeneous). The global standardized chi-square test can detect differential expression in 2 two opposite directions across pairs (heterogeneous). This test is not very sensitive, and may not be a good test for sample sizes with < 20 pairs. The global simple chi-square test can detect differential expression homogeneously or heterogeneously across pairs for any sample size. Thus, a simple chi-square test can be used as a "screening" test for perturbation of gene sets. However, this test can be very sensitive and can produce false positive rates slightly higher than random chance. Thus for simple chi- square tests, the p-values should be <0.01 to claim significance, and this test should be confirmed by a global z-test or a standardized chi-square test.

In term of power, the gamma GAGE approach is over sensitive for perturbation in only a few pairs, and thus, can confusingly produce high power for both up and down regulation for up-regulation simulation. The global z-test of ZZ-GSA when applied to PAGE, GAGE, and T-profiler does not have this drawback. The global z-test of PAGE-ZZ-GSA has considerably

higher power than the (Stouffer) GAGE-ZZ-GSA in detecting differential expression in one direction in the gene set.

In terms of type I errors of the tests for one directional gene sets, false positive rates of both gamma and Stouffer GAGE approach are lower than random chance and lower than the PAGE-ZZ-GSA and T-profiler. However, except for the global simple chi-square test whose type I error is predictably slightly higher than random chance, type I error of the global z-test of ZZ-GSA is lower or approximates random chance and type I error of the standardized chi-square test of ZZ-GSA is almost zero.

The false positive rate of the ZZ-GSA approach for bidirectional gene sets is considerably larger than random chance. For gamma GAGE, the type I error of the test for bidirectional gene sets is lower than random chance. Thus, the tests for bidirectional gene sets of the ZZ-GSA approach should be used with caution. For ZZ-GSA, the test can only be claimed for significant if p-value <0.001.

The PAGE-ZZ-GSA and T-profiler-ZZGSA have better power than the (Stouffer) GAGE-ZZ-GSA, while the type I error of ZZ-GSA is smaller or equal to random chance. While the advantage of (Stouffer) GAGE-ZZ-GSA is small type 1 error, its low power can be a concern. The gamma GAGE is sensitive to the perturbation in only a few pairs. Thus, although its type I error is much smaller than random chance, the use of this test and the interpretation of its results should be cautious.

The current GAGE software available in Bioconductor can only be used for microarray data without missing values. For microarray data with many missing values, this software package can produce unacceptably large false positive rate of up to 90-100%. Thus, it cannot be used. The current GAGE software needs to be re-programmed or modified in order to handle

missing values properly. Our R program for implementing the ZZ-GSA approach and modifying the GAGE software to handle microarray data with and without missing values is available upon request.

The permutation results on actual microarray data in terms of type I error and satisfaction of statistical assumptions in this paper can be used as a reference for future use of the ZZ-GSA and GAGE on different microarray datasets. However, different datasets may contain different percentages of missing values, so the reference panel for datasets with missing values may be somewhat different for different datasets. Thus, although it is not a must, rerunning the permutation could produce customized reference panels for datasets with missing values (our R program for simulation and permutation study will be freely available upon request).

The KEGG gene sets represent different patho-physiological and biological pathways and are of different sizes. Thus KEGG gene sets are appropriate for demonstrating the PAGE-ZZ-GSA and GAGE approaches for gene set analysis of our microarray data of CP cases and their matched controls. The combination of use of the global z-test, simple chi-square test and standardized chi-square test of PAGE-ZZ-GSA can pick up more significant gene sets than the Stouffer GAGE without increasing type I error. The combination use of these 3 tests can help describe how the gene set is differentially expressed across pairs, homogeneously or heterogeneously and whether up or down regulation is more predominant. For example, for the Lupus and the Leukocyte transendothelial migration (LTM) gene sets, without taking multiple test testing into account, the p-value of the Lupus gene set is <0.05 by the global z-test for up-regulation and by the global standardized chi-square test in the first and the third version of the dataset. For the LTM gene set, the p-value is <0.05 by the global z-test for down-regulation and by the global standardized chi-square test. Thus, differential expression of Lupus and LTM gene

sets is heterogeneous across pairs (some pairs are up regulated and some pairs are down regulated), however, up-regulation is more predominant for the Lupus gene set and down-regulation is more predominant for the LTM gene set. For porphyrin and chlorophyll metabolism (PCM) gene sets, p-values are <0.05 by the global z-test for up-regulation and >0.05 by the global chi-square tests. This gene set is homogenously up-regulated across pairs. The global z-test of the (Stouffer) GAGE-ZZ-GSA is similar to that of PAGE-ZZ-GSA and t-profiler-ZZ-GSA, but has lower power and smaller type I error. Thus, this test can be used confirmed the results of the global z-test of PAGE-ZZ-GSA or t-profiler-ZZ-GSA.

The pilot microarray dataset in 3 versions used in this paper is only about 10% of our planned sample size. Commonly used filtering and normalization approaches were used to process the raw data. The results presented in this paper are mainly for evaluation and demonstration of statistical approaches. Results may be somewhat different with larger sample sizes or with different approaches for processing raw microarray data of degraded mRNA or with different approaches for imputing missing values.

In summary, both the gamma GAGE and the ZZ-GSA approaches can be used for gene set analysis of matched microarray data. The combination use of the global z-test, simple chi-square test and standardized chi-square test of the proposed ZZ-GSA approach increases power in detecting differential expression both homogeneously and heterogeneously across pairs with statistically reasonable type I errors, and helps interpret the pattern of differential expression across pairs properly. The gamma GAGE and (Stouffer) GAGE-ZZ-GSA approaches produce very low false positive rates. However, the (Stouffer) GAGE has lower power than PAGE- ZZ-GSA and T-profiler-ZZ-GSA. The gamma GAGE can be over sensitive to perturbation in a few pairs, and interpretation of the results should thus take this issue into account. The test for bi-

directional regulation of the gene set of PAGE-ZZ-GSA, T-profiler-ZZ-GSA, (Stouffer) GAGE-ZZ-GSA produces considerably high false positive rates, and thus, should be used with caution.

CHAPTER6. SUMMARY, FUTURE RESEARCH APPLICATIONS AND DIRECTIONS

Gene set analysis of microarray data from uDNBS of 53 CP cases vs. matched controls of the OWL study has revealed the role of three of the four pre-hypothesized pathways of CP. Empirical gene sets representing hypoxia and inflammation are significantly down-regulated while the empirical gene set representing thyroid hormone disorders is significantly up-regulated. The differential expression of these three gene sets may suggest a co-effect or correlation of the three pathways of inflammation, hypoxia and thyroid disorders on the development of CP. Thus, a statistical method is needed to evaluate the correlation among gene sets. Of the three significant gene sets, the empirical gene set representing the fetal inflammatory response syndrome appears to be the most prominent. The analysis exploring other gene sets from the KEGG database also reveals the differential expression of inflammatory-related gene sets. This indicates that inflammation may play an important role in the development of CP during the peri-partum period. The stratified analysis shows that the empirical inflammatory gene set seems to up-regulated in uDNBS of preterm newborns while down-regulated in term newborns. This indicates an interaction effect between gestational age and this inflammatory gene set.

The results of pathway analysis suggest that uDNBS of long term storage yield considerable amount of mRNA for genome wide gene expression profiling and differential expression of gene sets may be found. However, further exploration and examination of microarray data from uDNBS has revealed significantly decreasing trend of microarray data quality over time of storage. The deterioration of microarray data quality over time may reduce the possibility of detecting differential expression of individual genes or gene sets. The initial results from 53 CP case –matched control pairs suggest that differential expression of individual

genes as well as gene sets may be preserved better in uDNBS of less than 6 years old. Therefore, we recommend prioritizing uDNBS of six years old or less for study.

The decreasing trend of detected expression signal over time of storage is found significantly in around 90% of all genes available in the arrays. This makes the criteria for selection of housekeeping genes based on stable expression signal across samples inapplicable. The proposed approach for selection of housekeeping genes based on the median of decreasing slopes of expression signal of all genes takes this decreasing trend into account and may be generalized if there are other covariates that affect the trend of detected expression signal. For microarray data used in this dissertation, this approach helps us to select six potential housekeeping genes which are selected robustly across all arrays as well as across case-control groups and laboratory batches. However, the qualification of these genes may need further evaluation by qPCR. Also, the simple conventional quantile normalization method is used to produce the processed microarray data used for the selection of these genes. Thus, the list of potential housekeeping genes may change if other methods of normalization which take the decreasing trend of expression signal into account are used.

Both the original gamma GAGE method and the proposed ZZ-GSA approach can be used for gene set analysis of matched microarray data as long as the strength and weakness of each approach is understood and is taken into account properly. The combination use of the global z-test, simple chi-square test and standardized chi-square test of the proposed ZZ-GSA approach increases power in detecting differential expression both homogeneously and heterogeneously across pairs with statistically reasonable type I error and helps interpret the pattern of differential expression across pairs properly. The gamma GAGE and (Stouffer) GAGE-ZZ-GSA approaches produce very low false positive rates. However, the (Stouffer) GAGE has lower power than

PAGE- ZZ-GSA and T-profiler-ZZ-GSA. The gamma GAGE can be oversensitive for perturbation in a few pairs, and the interpretation of the results should take this issue into account. For the test of bidirectional regulation of gene set, the gamma GAGE can produce reliable results while the PAGE-ZZ-GSA, T-profiler-ZZ-GSA, (Stouffer) GAGE-ZZ-GSA produces considerably high false positive rates, and should be used with caution.

REFERENCES

REFERENCES

1 Little WJ. On the influence of abnormal parturition, difficult labours, premature birth, and asphyxia neonatorum, on the mental and physical condition of the child, eapecially in relation to deformities. Tranacribed from The Obstetric Society of London 1861-62; 3: 293.

2 Little WJ. The classic: Hospital for the cure of deformities: course of lectures on the deformities of the human frame. 1843. Clin Orthop Relat Res. 2012 May;470(5):1252-6.

3 Freud S (1968). Infantile Cerebral Paralysis. University of Miami Press, Coral Gables, FL, USA (Original work published in 1897).

4 A. Kavcˇ icˇ and D. B. Vodusˇek. A historical perspective on cerebral palsy as a concept and a diagnosis

5 Raju TN. Historical perspectives on the etiology of cerebral palsy. Clin Perinatol. 2006 Jun;33(2):233-50. .

6 Peter Rosenbaum, Nigel Paneth et al. A report: the definition and classification of cerebral palsy, April 2006. Dev Med Child Neurol Suppl. 2007 Feb;109:8-14.

7 Paneth N, Kiely JL. The frequency of cerebral palsy: A review of population studies in industrialized nations since 1950. Clinic Dev Med 1984; 87:46-56.

8 Bhushan V, Paneth N, Kiely JL. Recent secular trends in the prevalence of cerebral palsy. Pediatrics 1993; 91:1094-1100.

9 Paneth et al. The descriptive epidemiology of cerebral palsy. Clin Perinatol. 2006 Jun;33(2):251-67.

10 Yeargin-Allsopp M, Van Naarden Braun K, Doernberg NS et al: Prevalence of cerebral palsy in 8-year-old children in three areas of the United States in 2002: a multisite collaboration. Pediatrics. 2008;121:547-54.

11 Arneson CL, Durkin MS, Benedict RE, Kirby RS, Yeargin-Allsopp M, Van Naarden Braun K, Doernberg NS. Prevalence of cerebral palsy: Autism and Developmental Disabilities Monitoring Network, three sites, United States, 2004. Disabil Health J. 2009;2:45-8.

12 Kirby RS, Wingate MS, Van Naarden Braun K, Doernberg NS, Arneson CL, Benedict RE, Mulvihill B, Durkin MS, Fitzgerald RT, Maenner MJ, Patz JA, Yeargin-Allsopp M. Prevalence and functioning of children with cerebral palsy in four areas of the United States in 2006: a report from the Autism and Developmental Disabilities Monitoring Network. Res Dev Disabil. 2011;32:462-9.

13 E Himpens et al. Prevalence, type, distribution, and severity of cerebral palsy in relation to gestational age: a meta-analytic review. Dev Med Child Neurol. 2008 May;50(5):334-40.

14 Centers for Disease Control and Prevention (CDC) (2004). "Economic costs associated with mental retardation, cerebral palsy, hearing loss, and vision impairment—United States, 2003". MMWR Morb. Mortal. Wkly. Rep. 53 (3): 57–9.

15 Wu YW, Escobar GJ, Grether JK, Croen LA, Greene JD, Newman TB: Chorioamnionitis and cerebral palsy in term and near-term infants. JAMA. 2003;290:2677-84.

16 Neufeld MD, Frigon C, Graham AS, Mueller BA: Maternal infection and risk of cerebral palsy in term and preterm infants. J Perinatol. 2005;25:108-13.

17 Croen LA, Grether JK, Curry CJ, Nelson KB: Congenital abnormalities among children with cerebral palsy: More evidence for prenatal antecedents. J Pediatr 2001;138:804-10.

18 Boyle CA, Yeargin-Allsopp M, Schendel DE, Holmgreen P, Oakley GP: Tocolytic magnesium sulfate exposure and risk of cerebral palsy among children with birth weights less than 1,750 grams Am J Epidemiol. 2000;152:120-4.

19 Gilbert WM, Jacoby BN, Xing G, Danielsen B, Smith LH. Adverse obstetric events are associated with significant risk of cerebral palsy. Am J Obstet Gynecol. 2010 Oct;203(4):328.e1-5. Epub 2010 Jul 3. PubMed PMID: 20598283.

20 Wu YW, Croen LA, Torres AR, Van De Water J, Grether JK, Hsu NN. Interleukin-6 genotype and risk for cerebral palsy in term and near-term infants. Ann Neurol.2009 Nov;66(5):663-70. PubMed PMID: 19938160.

21 Grether JK, Nelson KB, Walsh E, Willoughby RE, Redline RW: Intrauterine exposure to infection and risk of cerebral palsy in very preterm infants. Arch Pediatr Adolesc Med 2003;157:26-32.

22 Nelson KB, Dambrosia JM, Iovannisci DM, Cheng S, Grether JK, Lammer E: Genetic polymorphisms and cerebral palsy in very preterm infants. Pediatr Res 2005;57:494-499.

23 Nelson KB, Grether JK, Dambrosia JM, Walsh E, Kohler S, Satyanarayana G, Nelson PG, Dickens BF, Phillips TM: Neonatal cytokines and cerebral palsy in very preterm infants. Pediatr Res 2003;53:1-8.

24 Costantine MM, How HY, Coppage K, Maxwell RA, Sibai BM. Does peripartum infection increase the incidence of cerebral palsy in extremely low birthweight infants? Am J Obstet Gynecol. 2007;196:e6-8. PubMed PMID: 17466686.

25 Kuroda MM, Weck ME, Sarwark JF, Hamidullah A, Wainwright MS. Association of apolipoprotein E genotype and cerebral palsy in children. Pediatrics. 2007;119:306-13. PubMed PMID: 17272620.

26 Jacobsson B et al. Antenatal risk factors for cerebral palsy. Best Pract Res Clin Obstet Gynaecol. 2004 Jun;18(3):425-36.

27 Badawi N et al. Antepartum risk factors for newborn encephalopathy: the Western Australian case-control study. BMJ 1998;317(7172):1549-53.

28 Badawi N et al. Intrapartum risk factors for newborn encephalopathy: the Western Australian case-control study. BMJ 1998;317(7172):1554-8.

29 Gilbert WM, Jacoby BN, Xing G, Danielsen B, Smith LH. Adverse obstetric events are associated with significant risk of cerebral palsy. Am J Obstet Gynecol. 2010 Oct;203(4):328.e1-5. Epub 2010 Jul 3.

30 Graham EM, Ruis KA, Hartman AL, Northington FJ, Fox HE. A systematic review of the role of intrapartum hypoxia-ischemia in the causation of neonatal encephalopathy. Am J Obstet Gynecol. 2008 Dec;199(6):587-95.

31 Wu YW, Colford JM Jr. Chorioamnionitis as a risk factor for cerebral palsy: A meta-analysis. JAMA. 2000 Sep 20;284(11):1417-24.

32 Shatrov JG, Birch SC, Lam LT, Quinlivan JA, McIntyre S, Mendz GL. Chorioamnionitis and cerebral palsy: a meta-analysis. Obstet Gynecol. 2010 Aug;116(2 Pt 1):387-92.

33 O'Shea TM, Allred EN, Dammann O, Hirtz D, Kuban KC, Paneth N, Leviton A; ELGAN study Investigators. O'Shea TM, Allred EN, Dammann O, Hirtz D, Kuban KC, Paneth N, Leviton A; ELGAN study Investigators. Early Hum Dev. 2009 Nov;85(11):719-25. Epub 2009 Sep 17.

34 Leviton A, Allred EN, Kuban KC, Hecht JL, Onderdonk AB, O'shea TM, Paneth N. Microbiologic and histologic characteristics of the extremely preterm infant's placenta predict white matter damage and later cerebral palsy. the ELGAN study. Pediatr Res. 2010 Jan;67(1):95-101.

35 Romero R, Gotsch F, Pineles B, Kusanovic JP. Inflammation in pregnancy: its roles in reproductive physiology, obstetrical complications, and fetal injury. Nutr Rev. 2007 Dec; 65(12 Pt 2):S194-202.

36 Yoon BH et al. Intrauterine infection and the development of cerebral palsy. BJOG. 2003 Apr;110 Suppl 20:124-7.

37 Romero R et al. Inflammation in pregnancy: its roles in reproductive physiology, obstetrical complications, and fetal injury. Nutr Rev. 2007 Dec;65(12 Pt 2):S194-202.

38 Girard S et al. Role of perinatal inflammation in cerebral palsy. Pediatr Neurol. 2009 Mar;40(3):168-74.

39 Saliba E et al. Inflammatory mediators and neonatal brain damage. Biol Neonate. 2001;79(3-4):224-7.

40 A.L. Den Ouden, J.H. Kok, P.H. Verkerk et al. The relation between neonatal thyroxine levels and neurodevelopmental outcome at age 5 and 9 years in a national cohort of very preterm and/or very low birth weight infants. Pediatr Res, 39 (1996), pp. 142–145

41 F.B. Diamond, J.S. Parks, A. Tenore et al. Hypothyroxinemia in sick and well preterm infants. Clin Pediatr (Phila), 18 (1979), pp. 559–561 555.

42 A.J. Hadeed, L.D. Asay, A.H. Klein et al. Significance of transient postnatal hypothyroxinemia in premature infants with and without respiratory distress syndrome. Pediatrics, 68 (1981), pp. 494–498.

43 C.W. Rabin, A.O. Hopper, L. Job et al. Incidence of low free T4 values in premature infants as determined by direct equilibrium dialysis. J Perinatol, 24 (2004), pp. 640–644.

44 M.L. Reuss, N. Paneth, J.M. Lorenz et al. Correlates of low thyroxine values at newborn screening among infants born before 32 weeks gestation. Early Hum Dev, 47 (1997), pp. 223–233.

45 C. Romagnoli, V. Curro, R. Luciano et al. Serial blood T4 and TSH determinations in low birth weight infantsInfluence of gestational age, birth weight and neonatal pathology on thyroid function. Helv Paediatr Acta, 37 (1982), pp. 331–344.

46 Hong T, Paneth N. Maternal and infant thyroid disorders and cerebral palsy. Semin Perinatol. 2008 Dec;32(6):438-45.

47 M.L. Reuss, N. Paneth, J.A. Pinto-Martin et al. The relation of transient hypothyroxinemia in preterm infants to neurologic development at two years of age. N Engl J Med, 334 (1996), pp. 821–827.

48 A. Lucas, R. Morley, M.S. Fewtrell. Low triiodothyronine concentration in preterm infants and subsequent intelligence quotient (IQ) at 8 year follow up. BMJ, 312 (1996), pp. 1132–1133 discussion 1133-1134.

49 P. Chowdhry, J.W. Scanlon, R. Auerbach et al. Results of controlled double-blind study of thyroid replacement in very low-birth-weight premature infants with hypothyroxinemia. Pediatrics, 73 (1984), pp. 301–305.\

50 A.G. van Wassenaer, J.H. Kok, J.J. de Vijlder et al. Effects of thyroxine supplementation on neurologic development in infants born at less than 30 weeks' gestation. N Engl J Med, 336 (1997), pp. 21–26.

51 Haddow JE, Palomaki GE, Allan WC, et al: Maternal thyroid deficiency during pregnancy and subsequent neuropsychological development of the child. N Engl J Med 341:549-555, 1999.

52 Pop VJ, Kuijpens JL, van Baar AL, et al: Low maternal free thyroxine concentrations during early pregnancy are associated with impaired psychomotor development in infancy. Clin Endocrinol (Oxf) 50:149-155, 1999.

53 Kooistra L, Crawford S, van Baar AL, et al: Neonatal effects of maternal hypothyroxinemia during early pregnancy. Pediatrics 117:161-167, 2006.

54 Golomb MR, Garg BP, Saha C, Azzouz F, Williams LS. Cerebral palsy after perinatal arterial ischemic stroke. J Child Neurol. 2008 Mar;23(3):279-86.

55 Kenet G, Nowak-Göttl U. Fetal and neonatal thrombophilia. Obstet Gynecol Clin North Am. 2006 Sep;33(3):457-66.

56 Senbil N, Yüksel D, Yilmaz D, Gürer YK. Prothrombotic risk factors in children with hemiplegic cerebral palsy. Pediatr Int. 2007 Oct;49(5):600-2.

57 Kirton A, deVeber G. Cerebral palsy secondary to perinatal ischemic stroke. Clin Perinatol. 2006 Jun;33(2):367-86.

58 Nelson KB. Perinatal ischemic stroke. Stroke. 2007 Feb;38(2 Suppl):742-5.

59 O'Callaghan M.E., MacLennan A.H., Haan E.A., Dekker G.South Australian Cerebral Palsy Research Group: The genomic basis of cerebral palsy: a HuGE systematic literature review. Hum Genet 126. 149-172. 2009.

60 Mohr S, Liew CC: The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends in Molecular Medicine 2007; 13: 422-432.

61 Liew CC Ma J, Tang HC, Zheng R, Dempsey AA: The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J. Lab. Clin. Med. 2006; 147, 126–132.

62 Olney RS, Moore CA, Ojodu JA et al: Storage and use of residual dried blood spots from state newborn screening programs. J Pediatr. 2006;148:618-22.

63 Haak PT et al. Archived unfrozen neonatal blood spots are amenable to quantitative gene expression analysis. Neonatology. 2009; 95(3): 210–216.

64 Resau JH, Ho NT et al. Evaluation of Sex-Specific Gene Expression in Archived Dried Blood Spots (DBS). Int. J. Mol. Sci. 2012, 13, 9599-9608; doi:10.3390/ijms13089599.

65 Nelson KB, Dambrosia JM, Grether JK, Phillips TM. Neonatal cytokines and coagulation factors in children with cerebral palsy. Ann Neurol 1998;44(4):665–675.

66 K.B Nelson, J.K Grether, J.M Dambrosia, E Walsh, S Kohler, G Satyanarayana et al. Neonatal cytokines and cerebral palsy in very preterm infants. Pediatr Res, 53 (2003), pp. 600pp. .

67 Grether JK, Nelson KB, Dambrosia JM, Phillips TM 1999 Interferons and cerebral palsy. J Pediatr 134: 324-332.

68 Wu YW, Croen LA, Torres AR, Van De Water J, Grether JK, Hsu NN: Interleukin-6 genotype and risk for cerebral palsy in term and near-term infants. Ann Neurol. 2009;66:663-70.

69 Djukic M, Gibson CS, Maclennan AH, Goldwater PN, Haan EA, McMichael G, Priest K, Dekker GA, Hague WM, Chan A, Rudzki Z, VAN Essen P, Khong TY, Morton MR, Ranieri E, Scott H, Tapp H, Casey G. Genetic susceptibility to viral exposure may increase the risk of cerebral palsy. Aust N Z J Obstet Gynaecol. 2009 ;49:247-53.PMID: 19566553.

70 Gibson CS, Maclennan AH, Dekker GA, Goldwater PN, Sullivan TR, Munroe DJ, Tsang S, Stewart C, Nelson KB; Candidate genes and cerebral palsy: a population-based study. Pediatrics. 2008 ;122:1079-85.

71 Rosenbaum PL, Palisano RJ, Bartlett DJ, Galuppi BE, Russell DJ: Development of the Gross Motor Function Classification System for cerebral palsy. Dev Med Child Neurol. 2008;50:249-53.PMID:18318732.

72 Eliasson AC, Krumlinde-Sundholm L, Rösblad B, Beckung E, Arner M, Ohrvall AM, Rosenbaum P: The Manual Ability Classification System (MACS) for children with cerebral palsy: scale development and evidence of validity and reliability. Developmental Medicine & Child Neurology 2006, 48: 549–554. PMID: 16780622

73 http://www.michigan.gov/mdch/0,1607,7-132-2942_4911_4916_53246-209738--,00.html.

74 Rosenbaum P, Dan B, Fabiola R, Leviton A, Paneth N, Jacobsson B, Goldstein M, Bax M: Proposed definition and classification of cerebral palsy, April 2005. The definition of cerebral palsy. Dev Med Child Neurol 2005;47:571-6.

75 Paneth N, Hong T, Korzeniewski S: The descriptive epidemiology of cerebral palsy. Clin Perinatol. 2006;33:251-67.

76 Haak PT et al. Archived unfrozen neonatal blood spots are amenable to quantitative gene expression analysis. Neonatology. 2009; 95(3): 210–216.

77 Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006; 24: 1140-50.

78 Bolstad BM, Irizarry RA, Astrand M, Speed TP et al: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185-93

79 Smyth GK: Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiment. Statistical Applications in Genetics and Molecular Biology 2004; 1: 3.

80 Smyth GK (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397–420.

81 Storey JD. A direct approach to false discovery rates. J. R. Statist. Soc. B (2002)64, Part 3, pp. 479–498.

82 Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann. Statist. Volume 31, Number 6 (2003), 2013-2035.

83 Curtis, R.K., M. Oresic, and A. Vidal-Puig, Pathways to the analysis of microarray data. Trends Biotechnol, 2005; 23: 429-35

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

85 Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ: GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009;10:161

86 Madsen-Bouterse SA, Romero R, Tarca AL et al: The transcriptome of the fetal inflammatory response syndrome. Am J Reprod Immunol. 2010;63:73-92.

87 Chi JT, Wang Z, Nuyten DS et al: Gene Expression Programs in Response to Hypoxia: Cell Type Specificity and Prognostic Significance in Human Cancers. PLoS Med, 2006;3: e47

88 Maynard, M.A. and M. Ohh: von Hippel-Lindau Tumor Suppressor Protein and Hypoxia-Inducible Factor in Kidney Cancer. Am J Nephrol 2004; 24: 1-13

89 http://www.broadinstitute.org/gsea/msigdb.

90 Clement K, Viguerie N, Diehn M et al: In vivo regulation of human skeletal muscle gene expression by thyroid hormone. Genome Res 2002; 12: 281-91

91 Olney RS, Moore CA, Ojodu JA et al: Storage and use of residual dried blood spots from state newborn screening programs. J Pediatr.2006;148:618-22.

92 Olney RS, Moore CA, Ojodu JA et al: Storage and use of residual dried blood spots from state newborn screening programs. J Pediatr. 2006;148:618-22.

93 Karlsson H, Guthenberg C, von Dobeln U, Kristenssson K. Extraction of RNA from dried blood on filter papers after long-term storage. Clin Chem 2003;49:979–81

94 Zhang YH, McCabe ER. RNA analysis from newborn screening dried blood specimens. Hum Genet 1992;89: 311–4.

95 Matsubara Y, Ikeda H, Endo H, Narisawa K. Dried blood spot on filter-paper as a source of messenger-RNA. Nucleic Acids Res 1992;20:1998

96 Gauffin F, Nordgren A, Barbany G et al: Quantitation of RNA decay in dried blood spots during 20 years of storage. Clin Chem Lab Med 2009;47:1467–1469.

97 Haak PT et al. Archived unfrozen neonatal blood spots are amenable to quantitative gene expression analysis. Neonatology. 2009; 95(3): 210–216.

98 Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7, 3.

99 Sambrook, J., Russel, D.W., 2001. Molecular Cloning: A Laboratory Manual, 3rd ed. Cold Spring Harbord Laboratory Press, Cold Spring Harbord, NY.

100 Van de Goor T.A. The principle and promise of Labchip technology. (2003) PharmaGenomics 3:16-18.

101 Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, Kornacker K: Chipping away at the chip bias: RNA degradation in microarray analysis. Nat Genet 2003, 35(4):292-3.

102 Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, Bareil C, Candeil L, Fraslon C, Conseiller E, Granci V, Mazière P, Kramar A, Ychou M, Pau B, Martineau P, Molina F, Del Rio M: Impact of RNA degradation on gene expression profiles: Assessment of different methods to reliably determine RNA quality. J Biotechnol 2007, 127(4):549-59.

103 Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006; 24: 1140-50.

104 Gordon K. Smyth. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiment. Statistical Applications in Genetics and Molecular Biology 3 (2004), No. 1, Article 3.

105 Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397–420.

106 Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ: GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics, 2009;10:161

107 Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E,Mueller O, Schroeder A, Auffray C: Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. Nucleic Acids Research(Published online 30 March) 2005, 33:e56.

108 Miller C, Diglisic S, Leister F, Webster M, Yolken R: Evaluating RNA status for RT-PCR in extracts of postmortem human brain tissue. Biotechniques 2004, 36(4):628-633.

109 Roche applied science. Selection of housekeeping genes. Technical note. No. LC 15/2005.

110 Nanostring technology. Reference genes for normalization of expression data. Technical note. 2009.

111 Lee PD, Sladek R, Greenwood CMT, Hudson TJ (2001) Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies. Genome Res 12: 292–297

112 Schmittgen T, Zakrajsek B (2000) Effect of experimental treatment on housekeeping gene expression: validation by real-time, quantitative RT-PCR. J. Biochem. Biophys. Methods 46: 69-81.

113 Selvey S, et al (2001) β-Actin - an unsuitable internal control for RT-PCR. Molecular and Cellular Probes 15: 307-311.

114 Zhong H and Simons J W (1999) Direct Comparison of GAPDH, ß-Actin, Cyclophilin, and 28S rRNA as Internal Standards for Quantifying RNA Levels under Hypoxia. Biochemical and Biophysical Research Communications 259: 523-526.

115 Barber RD, Harmer DW, Coleman RA & Clark BJ. 2005. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. Physiol Genomics 21 389– 95 doi:10.1152/physiolgenomics.00025.2005.

116 Maron, J.L., Arya, M.A., Seefeld, K.J., Peter, I., Bianchi, D.W., and Johnson, K.L. pH but not hypoxia affects neonatal gene expression: relevance for housekeeping gene selection. J. Matern. Fetal Neonatal Med. 21:443–447, 2008.

117 Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A (2004) Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques 37: 112–119.

118 Thellin,O., Zorzi,W., Lakaye,B., De Borman,B., Coumans,B., Hennen,G., Grisar,T., Igout,A. and Heinen,E. (1999) Housekeeping genes as internal standards: use and limits. J Biotechnol., 75, 291–295.

119 Radonić A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A (2004) Guideline for reference gene selection for quantitative real-time PCR. Biochem Biophys Res Commun 313: 856–862.

120 H.K Hamalainen, J.C Tubman, S Vikman, T Kyrola, E Ylikoski, J.A Warrington, R Lahesmaa. Identification and validation of endogenous reference genes for expression profiling of T helper cell differentiation by quantitative real-time RT-PCR. Anal. Biochem., 299 (2001), pp. 63–70.

121 de Jonge, H. J., R. S. Fehrmann, E. S. de Bont, R. M. Hofstra, F. Gerbens, W. A. Kamps, E. G. de Vries, A. G. van der Zee, G. J. te Meerman, A. ter Elst. 2007. Evidence based selection of housekeeping genes. PLoS ONE 2: e898.

122 Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002 Jun 18;3(7):RESEARCH0034.

123 Andersen CL, Jensen JL, Ørntoft TF. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 2004 Aug 1;64(15):5245-50.

124 Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006; 24: 1140-50.

125 Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50

126 Efron, B. and R. Tibshirani, On testing the significance of sets of genes. Stanford Statistics Working Paper, 2006: p. http://www-stat.stanford.edu/~tibs/ftp/GSA.pdf.

127 Hosack, D.A., et al., Identifying biological themes within lists of genes with EASE. Genome Biol, 2003. 4(10): p. R70

128 Tian, L. Greenberg, SA. Kong, SW Altschuler, J. Kohane, IS. Park, PJ. Discovering statistically significant pathways in expression profiling studies Proc. Natl Acad. Sci. 2005, 102, 38:13544-13549.

129 Irizarry RA, Wang C, Zhou Y, Speed TP: Gene set enrichment analysis made simple. Stat Methods Med Res. 2009;18:565-75.

130 Goeman, J.J., et al., A global test for groups of genes: testing associated with a clinical outcome. Bioinformatics, 2004. 20: p. 93-99

131 Zhen Jiang and Robert Gentleman. Extensions to gene set enrichment. Bioinformatics. Vol. 23 no. 3 2007, pages 306– 313. doi:10.1093/bioinformatics/btl599.

132 Boorsma, A. Foat, BC. Vis, DE. Klis, F. and Bussemaker, HJ.T-profiler: scoring the activity of predefined groups of gens using gene expression data, Nucleic Acids Res.2005 33, W592-595.

133 Kim S, Volsky DJ: PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics, 2005. 6: p. 144.

134 Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ: GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics, 2009;10:161

135 Fridley BL, Jenkins GD, Biernacka JM (2010) Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods. PLoS ONE 5(9): e12693. doi:10.1371/journal.pone.0012693

136 Stouffer S.A, Suchman E.A, Devinney L.C, Star S.A, Williams R.M Jr. The American soldier: Adjustment during army life. (Vol 1) Princeton, NJ: Princeton University Press.

137 Irizarry RA, Wang C, Zhou Y, Speed TP: Gene set enrichment analysis made simple. Stat Methods Med Res. 2009;18:565-75.

138 Rosner, B. Fundamentals of Biostatistics, 4th ed. p217-221. 1995, Duxbury, New York.

139 Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol. 2006; 24: 1140-50.