Public Health Risk Assessment:
Validation of risk assessment matrix limitations and an analytical approach to gene set reduction for continuous phenotype in microarray studies.

by

Shabnam Vatanpour

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Public Health

Department of Public Health Sciences
University of Alberta

ABSTRACT

Although risk is a core element of public health practice, its definition varies greatly among various public health programs. Several methods have been developed for risk assessment and management in different contexts of public health to better understand disease progression and outcome development. My dissertation consists of two rather different approaches to risk assessment: the first part deals with a flaw in the current public health risk assessment via risk matrices, the second part addresses a methodological gap in the analysis of data measured by DNA microarray technology.

We first evaluated the risk assessment matrix which is a semi-quantitative tool for assessing risks, and setting priorities in risk management. Although the method can be useful in promoting discussion to distinguish high risks from low risks, a published critique described a problem when the frequency and severity of risks are negatively correlated. A theoretical analysis showed that risk predictions could be misleading. We explored this predicted problem by constructing a risk assessment matrix using a public health risk scenario, tainted blood transfusion infection risk that provides negative correlation between harm frequency and severity. We estimated the risk from the experiential data and compared these estimates with those provided by the risk assessment matrix. We concluded that the risk assessment matrix should not be abandoned, but users must address the source of problem in applying the matrix to inform decision makers.

We then focused on DNA microarray studies which open a new platform with an opportunity to study and compare thousands of genes at the same time, leading to early and more accurate disease risk assessment, diagnosis, as well as improved tailored treatment. Advances in DNA microarray technology have stimulated methodological research on data analysis in biomedical

studies. Using microarray data analysis, researchers are able to assess the association of a priori defined gene sets sharing a common biological theme (pathways) with an outcome of interest (phenotype) and gain insights into biological functions of genes and pathways influencing disease mechanisms.

Gene set analysis (GSA) is a popular approach to examine the association between a predefined gene set and a phenotype. Few GSA methods have been developed for continuous phenotypes. However, often not all the genes within a significant gene set contribute to its significance. While a few methods have been developed to extract core genes from gene sets in the case of binary phenotypes studies, such as diseased versus disease-free subjects, no attention has been paid to studies measuring a continuous phenotype. We developed a computationally efficient gene set reduction method to identify core subsets of gene sets associated with a continuous phenotype. Identifying the core subset enhances our understanding of the biological mechanism and reduces costs of disease risk assessment, diagnosis and treatment.

To evaluate the performance of the method, we applied our method to two real microarray data sets. First, we examined the association between pathway expressions and tumor volume in a cohort of lethal prostate cancer patients from Swedish Watchful Waiting cohort, and extracted main genes from significant pathways. Second, we assessed whether there is an association between pathways expression in newborns' blood and their birth weight in Conditions Affecting Neurocognitive Development and Learning in Early Childhood (CANDLE) study, and reduced the significant pathways to their core subsets.

PREFACE

This thesis is an original work by Shabnam Vatanpour with supervision of Dr. Irina Dinu and co-supervision of Dr. Steve Hrudey. The data analysis in Chapters 1, 4, and 5 is my original work with Dr. Hrudey and Dr. Dinu. The literature review in Chapters 1 and 2 and the concluding Chapter 6 is my original work. Data used in Chapter 4 are publically available through the Gene Expression Omnibus website. Data used in Chapter 5 are available from the Conditions Affecting Neurocognitive Development and Learning in Early Childhood Study collected by the CANDLE study team at the University of Tennesse Health Science Center.

Chapter 1 of this dissertation has been published as Vatanpour S, Hrudey S, Dinu I. Can Prevailing Public Health Risk Assessment Methodology Be Misleading? International Journal of Environmental Research and Public Health, Int. J. Environ. Res. Public Health. 2015, 12, 9575-9588. I was responsible for conducting literature review, designing the application in the tainted blood transfusion, and evaluating the quantitative risk assessment as well as manuscript drafting. Dr. Hrudey was the senior author responsible for the concept formation of the research project, study design, and supervision of the research conduct. Dr. Dinu supervised the development and evaluation of the quantitative risk assessment. Dr. Duncan Saunders and Dr. Yutaka Yasui provided critical review comments. This research was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada held by Dr. Steve Hrudey.

Chapter 4 of this dissertation will be submitted as Vatanpour S, Yasmin F, Pana M, Wang X, Pyne S, Dinu I. Core subset of gene sets associated with prostate cancer tumor size. I was responsible for the development of the study, programming and conducting the data analysis, interpretation and presentation of the results as well as manuscript drafting. Farzana Yasmin

assisted in conducting the analysis, Mara Pana assisted in interpretation of the results. Dr. Dinu was involved in the concept formation and supervised the analysis design and research conduct.

Chapter 5 of this dissertation will be submitted as Vatanpour S, Pyne S, Leite A.P, Dinu I. Gene set analysis and reduction of birth weight based on embryonic stem cells and immunologic signatures. I was responsible for the development of the study, programming and conducting the data analysis, interpretation and presentation of the results as well as manuscript drafting. Dr. Pyne provided the study data and research question. Dr. Dinu supervised the methodological development, data analysis, results interpretation and manuscript drafting. Ana Paula Leite provided the list of stem cell signatures.

This dissertation is dedicated

to my beloved parents,

and to my supportive husband, Nima.

AKNOWLEDGMENTS

Over the past four years I have received support and encouragement from many people. I would like to express my sincere gratitude to my supervisor, Dr. Irina Dinu for her patience, guidance and support throughout my PhD program, and especially for her confidence in me. Her dedication and passion for her research was motivational for me, even during tough times of conducting my research off campus. I really appreciate her constructive feedback and insight for conducting this thesis, and beyond that for helping me shape my career path. What I have learned from her over the past years is valuable.

I would like to sincerely thank Dr. Steve Hrudey for his support and assistance. His insights, unique expertise, and motivation were a great source of inspiration for me during the first part of this thesis. I really appreciate his help and effort for publishing the first manuscript in this thesis. It is a privilege to work closely with a researcher of his unique caliber.

I wish to express my appreciation to Dr. Saumyadipta Pyne for his inspiring research idea and his insightful comments. My sincere thanks go to Ana Paula Leite who provided me with the data set from her research work. Special thanks to Dr. Anita Kozyrskyj, Dr. Linglong Kong for their helpful feedback.

I wish to thank Dr. Duncan Saunders and Dr. Yutaka Yasui for their critical review and constructive feedback on my published paper. I wish to thank Farzana Yasmin for the stimulating discussions we had on this research. I am grateful to my professors and colleagues at the School of Public Health. They have been a source of professional support and development. I also wish to thank the CANDLE team for providing the data set for my research and their helpful support.

I owe my deepest debt of gratitude to my parents, Soraya and Hesam. They have supported every aspect of my life with boundless love, support and encouragement. I would also like to thank my sisters for their influential guidance in this path. I cannot thank enough my husband, Nima, for believing in me. Completing this work would have never been a success without his support, understanding and encouragement.

Lastly, I would like to acknowledge the research funding support I received from a number of organizations during the course of this work including the Queen Elizabeth II Graduate Scholarship, the School of Public Health Travel Award, the Profiling Alberta's Graduate Students Award.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Can Public Health risk assessment using risk matrices be misleading?

1.1 Introduction

Assessing and managing risk is a core element of public health practice, although explicit and detailed documentation of these processes varies among various public health programs. Use of a qualitative (semi-quantitative) risk assessment matrix is a growing practice. The comparative simplicity and apparent ease of use of this approach likely contributes to widespread adoption including a generic international standard for risk assessment techniques in support of risk management (ISO 31000, 2009). Major public institutions have adopted the risk assessment matrix in fields ranging from assessing highway construction risk, financial risk, preventing terrorist attacks, to agency-wide enterprise risk management across all of government (Ashley et al, 2006; Guide to corporate risk profile, 2013). The World Health Organization has adopted this approach for risk assessment of acute public health events (WHO, 2012) and for assuring safe drinking water (WHO, 2011). Risk matrices have also been adopted nationally in Australia for assuring safe drinking water (NHMRC, 2013) and for drinking water safety plan implementation in Alberta, Canada (Drinking water safety plan training course, 2013).

Although the various applications of this technique differ in specific details, they all involve the common structural features of a matrix with one axis representing categories of probability (likelihood or frequency) of possible hazardous events and the other axis representing categories

of severity (impact or consequences) of those events. Each intersecting cell of the matrix (i.e., row-column pair) is pre-assigned a risk such as low, medium, or high risk. This basic structure is consistent with a widely adopted, if somewhat simplified, concept of risk as being primarily a function of two variables, one representing probability and the other consequences.

The UK National Health Service (NHS) has developed detailed guidance for applying the risk assessment matrix technique, which specified the following properties as being essential for such a risk assessment matrix, "it should:

• be simple to use;

• provide consistent results when used by staff from a variety of roles or professions;

• should be capable of assessing a broad range of risks including clinical, health and safety, financial risks, and reputation; and

• should be simple for NHS trusts to adapt to meet their specific needs."(NPSA/NHS, 2008)

The ISO standard characterized this technique as offering (ISO 31000, 2009):

"Strengths:

• relatively easy to use;

• provides a rapid ranking of risks into different significance levels.

Limitations:

• a matrix should be designed to be appropriate for the circumstances so it may be difficult to have a common system applying across a range of circumstances relevant to an organization;

• it is difficult to define the scales unambiguously;

• use is very subjective and there tends to be significant variation between raters;

• risks cannot be aggregated (i.e., one cannot define that a particular number of low risks or a low risk identified a particular number of times is equivalent to a medium risk);

• it is difficult to combine or compare the level of risk for different categories of consequences."

Cox outlined a number of serious deficiencies with the risk assessment matrix approach for assessing risk, including: Poor resolution, ambiguous inputs and outputs, sub-optimal allocation of resources based on inaccurate risk estimation and outright errors in assigning higher rankings to quantitatively lower risks (Cox, 2008). In particular, for the last concern, Cox demonstrated that the prediction of risk arising from the risk assessment matrix could be worse than a random guess by using a mathematical function for which frequency and severity are negatively correlated and using the commonly adopted formulation (with frequency as a measure of probability and severity as a measure of consequence):

$$\text{risk} = \text{frequency} \times \text{severity} \qquad (1.1)$$

This definition of risk provides one value for the risk of a scenario. The notion of risk cannot be summarized in one value and a large amount of information can be lost. The most powerful definition of risk is the set of triplets (scenario, likelihood, severity) which incorporates uncertainty into estimation of likelihood and severity.

Specifically Cox proposed the following theoretical but plausible deterministic negative relationship between frequency and severity values (Cox, 2008):

$$\text{frequency} = z - \text{severity (for severity between 0 and z)} \quad (1.2)$$

He designed a simplified 2 × 2 risk assessment matrix with two categories of frequency (Low, High) and two categories of severity (Low, High), then assigned medium risk to the pairs (frequency, severity) of (Low, High) and (High, Low), high risk to the pair (High, High), and low risk to the pair (Low, Low). He demonstrated that in this risk assessment matrix, most points in the medium risk categories actually have smaller risk values from Equation (1.1) than any points in the low risk cells.

This theoretical example demonstrates that the risk category assignment by the matrix is different from the risk calculation that is intended to accurately estimate the risk and, as such, the risk matrix predictions can be, according to Cox (2008), worse than useless (i.e., worse than random).

The prospect of risk predictions being worse than random for risks having a negative correlation between frequency and severity is gravely troubling because such a negative correlation is to be expected in many, if not most, of the circumstances that risk assessment matrix is used to characterize. The wide-spread practice of risk management has reduced the occurrence of hazards causing serious consequences, making their frequency lower. Certainly, for risks being able to accurately distinguish low frequency-high consequence risks from high frequency-low consequence risks is crucial.

Despite a growing number of citations, this grave concern of the risk assessment matrix method has received little traction in applied fields such as public health since first proposed by Cox in 2008.

Given our focus on health risk, we sought a practical public health example for which we could find experiential data on risk to assess the practical implications of this concern about risk assessment matrices. Cases, such as drinking water safety, where risk assessment matrices are being widely adopted were not pursued for our analysis because, while there is no shortage of monitoring data, little of this can be readily used for assessing tangible public health risk (Rizak & Hrudey, 2006). The connection between available monitoring data and risk is complex and drinking water disease outbreaks in affluent countries are comparatively rare (Hrudey & Hrudey, 2004).

The tangible health risks associated with tainted blood transfusions, by comparison, offers a circumstance where, after the major tragedies associated with HIV and hepatitis C transmission through transfusion of tainted blood and blood products, there has been a concerted effort to estimate the frequency of blood contamination for a range of pathogens capable of causing a wide range of disease outcomes of variable severity. Quintela et al. (2008) produced a generic risk assessment matrix addressing production processes in blood banks, but this analysis did not provide the kind of risk data needed to evaluate the Cox concerns.

The objective of our study is to explore the validity of risk matrices for health risk assessment by using a public health risk scenario, tainted blood transfusion infection risk because it provides experiential frequency data estimates for which the frequency of a risk is expected to be negatively correlated with the severity of consequences. That negative correlation is a

requirement for allowing risk assessment matrix predictions to be worse than random and potentially harmful according to the analysis of Cox (2008).

1.2 Methods

To illustrate the behavior of the risk assessment matrix tool, first we constructed a risk assessment matrix for the hazards associated with infection risk from tainted-blood transfusion using only frequency and severity values. Second, we identified the relationship between frequency and severity values and estimated the risk using Equation (1.1). Then we compare the estimated risk values (quantitative values) with the risk levels in the risk assessment matrix to verify their compatibility.

Risk ranking for decision makers in the risk assessment matrix is commonly visualized by assigning colors to risk categories, which are the cells in the matrix. The assignment of risk categories to the risk assessment matrix (Figure 1.1) must be done initially by the risk assessor, with an application of judgment, before any specific risks are placed in the matrix. Misunderstanding that this color-coding approach must be restricted to risk has appeared where color-coding was also pre-assigned for both the severity and frequency categories (NPSA/NHS, 2008). The color-coding in a risk assessment matrix must only apply to the risk categories that are a product of the severity and frequency ratings that determine the location of any specific risk in the matrix. The magnitude assignment (provided by the color coding) for any risk thus results from its placement in the matrix according to its estimated severity and frequency.

Colored Cells are the Risk Categories	Low Risk	Medium Risk	High Risk

Frequency of Scenario	Severity of Consequences		
	Low Severity	Medium Severity	High Severity
High Frequency	Medium	High	High
Medium Frequency	Low	Medium	High
Low Frequency	Low	Low	Medium

Figure 1.1 Generic risk assessment matrix

Table 1.1 National Health Service criteria for severity and frequency levels

Criteria for Severity Levels	
Very Low Severity	• Minimal injury requiring no/minimal intervention or treatment • No time off work
Low Severity	• Minor injury or illness requiring minor intervention • Increase in length of hospital stay by 1–3
Medium Severity	• Moderate injury requiring professional intervention • Increase in length of hospital stay by 4–15 days • Impacts on a small number of patients
High Severity	• Major injury leading to long-term incapacity/disability • Increase in length of hospital stay by >15 days
Very High Severity	• Incidence leading to death • Multiple permanent injuries or irreversible health effects • Impacts on a large number of patients
Criteria for Frequency Levels	
Extremely Low Frequency	• Frequency between 0.000001 and 0.0000099
Very Low Frequency	• Frequency between 0.00001 and 0.000099
Low Frequency	• Frequency between 0.0001 and 0.00099
Medium Frequency	• Frequency between 0.001 and 0.0099
High Frequency	• Frequency between 0.01 and 0.099
Very High Frequency	• Will undoubtedly happen/recur, possibly frequently. Frequency greater than 0.1

We adapted the NHS criteria (2008) for assigning the severity and frequency rankings as listed in Table 1.1. To obtain estimates of frequency for our purposes, we collected the prevalence estimates of different blood infectious diseases in blood donors and the population of Canada from the reports of the Public Health Agency of Canada (2007) from 1987 to 1996 (Table 1.2). For these data we found a very wide range (6 orders of magnitude) of frequency values (0.0000008 to 0.4; Table 1.2). Because of the wide range of values involved, we adopted a logarithmic scale for both the frequency and severity categories.

Table 1.2 Severity and frequency of blood infectious diseases in Canada, 1987-1996

Infectious Diseases	Severity	Severity Category [a]	Frequency	Frequency Category [b]	Source
HIV	10^5	Very High	0.000001	Extremely Low	Blood Donors
HTLV	10^4	High	0.0000008	Extremely Low	Blood Donors
Hepatitis B	10^3	Medium	0.00001	Very Low	Blood Donors
Hepatitis C	10^3	Medium	0.000004	Extremely Low	Blood Donors
Hepatitis G	10	Very Low	0.01	High	Blood Donors
Bacterial Contamination	10^2	Low	0.000026	Very Low	Blood Donors
Cytomegalovirus	10^2	Low	0.4	Very High	Blood Donors
Epstein-Barr virus	10^2	Low	0.9	Very High	Blood Donors
TT virus	10	Very Low	0.3	Very High	Blood Donors
SEN virus	10	Very Low	0.02	High	Blood Donors
CJD/vCJD	10^5	Very High	0.000001	Extremely Low	Population
Syphilis	10^4	High	0.000006	Extremely Low	Blood Donors

[a] Categories assigned using the severity categories provided in Table 1; [b] Categories assigned using the frequency categories provided in Table 1.

Because we located no reports on the prevalence of Creutzfeldt Jakob Disease/variant Creutzfeldt Jakob Disease (CJD/vCJD) in blood donors we used the prevalence in the entire

population instead. We acknowledge that this will likely over-estimate the frequency and consequently the risk among blood donors for transmitting CJD/vCJD.

We evaluated the disease severity by assigning severity ranging from very low to very high for each blood infectious disease according to expected complications, mortality, morbidity and available treatment for the infection. While the severity ranking is clearly a judgmental input to the risk assessment matrix based on NHS criteria ranging from very low to very high, frequency is assigned a ranking (extremely low to very high) based on where the frequency evidence dictates (i.e., according to Table 1.1).

For the matrix scheme we adopted an additional color was added to deal with the wide range of values in frequency and consequences. In our scheme (Figure 1.2) red indicates very high risk that requires immediate actions and priority in decision-making, orange indicates high risk that requires attention and a control process, yellow indicates moderate risk that requires a specific monitoring program, and green indicates low risk that can be managed according to current standard controls and regulation. The expectation for a risk assessment matrix is that the semi-quantitative ranking provided will be consistent with an underlying quantitative risk ranking which could, at least in theory, be defined by a risk function.

For each infectious hazard in Table 1.2, we were able to place it in the risk assessment matrix (Figure 1.2) by considering the frequency and severity category according to the assignments we made in Table 1.2 according to the NHS scheme (Table 1.1). In addition, because we have the experience-based estimates of frequency for each hazard and we could use a mid-point of the assigned judgmental severity category from Table 1.2, we were able to calculate a risk value, using Equation (1.1). This value is shown for each infectious hazard in Table 1.2 as the number

labeled “Obs.” meaning “observed” for each hazard placed in the risk assessment matrix (Figure 1.2).

To allow us to evaluate the concern expressed by Cox (2008), we calculated Spearman’s correlation of frequency and severity in this risk assessment matrix in logarithmic scales to confirm whether the data we were using satisfied the Cox requirement for a negative correlation between severity and frequency.

Colored cells are the Risk Categories	Low Risk	Medium Risk	High Risk	Very High Risk

Frequency of Infection	Severity of Consequences				
	Very Low Severity	Low Severity	Medium Severity	High Severity	Very High Severity
Very High Frequency	TT virus [a]Obs 3 [b]Est 10	Cytomegalovirus [a]Obs 35 [b]Est 13 Epstein-Barr virus [a]Obs 90 [b]Est 79			
High Frequency	SEN virus [a]Obs 0.2 [b]Est 0.19 Hepatitis G [a]Obs 0.11 [b]Est 0.10				
Medium Frequency					
Low Frequency					
Very Low Frequency		Bacterial contamination [a]Obs 0.003 [b]Est 0.007	Hepatitis B [a]Obs 0.01 [b]Est 0.01		
Extremely Low Frequency			Hepatitis C [a]Obs 0.004 [b]Est 0.014	Syphilis [a]Obs 0.06 [b]Est 0.01 HTLV [a]Obs 0.01 [b]Est 0.05	HIV [a]Obs 0.13 [b]Est 0.03 CJD/vCJD [a]Obs 0.1 [b]Est 0.04

Figure 1.2 Risk assessment matrix providing colored risk categories plus observed and estimated risk. a Observed (Obs) risk numbers shown are based on the generic risk function (Risk = Frequency × Severity; Equation (1.1)) and using Table 1.1 entries for frequency and severity based on Table 2 data; b Estimated (Est) risk numbers shown are based on the fitted risk function Equation (1.4)

Furthermore, we determined an empirical relationship for log-severity as a function of log-frequency for these infectious disease data, as:

$$\text{log-Severity} = 0.24\ \text{log-Frequency}^2 + 1.01\ \text{log-Frequency} + 1.99 \qquad (1.3)$$

Applying the basic relationship for risk in terms of severity and frequency (Equation (1.1)) to Equation (1.3), an empirical equation for risk as a function of frequency can be determined as:

$$\text{log-Risk} = 1.99 + 2.01\ \text{log-Frequency} + 0.24\ \text{log-Frequency}^2 \qquad (1.4)$$

The relationship between this empirical function and the observed estimates of risk derived from Table 1.2 is shown in Figure 1.3.

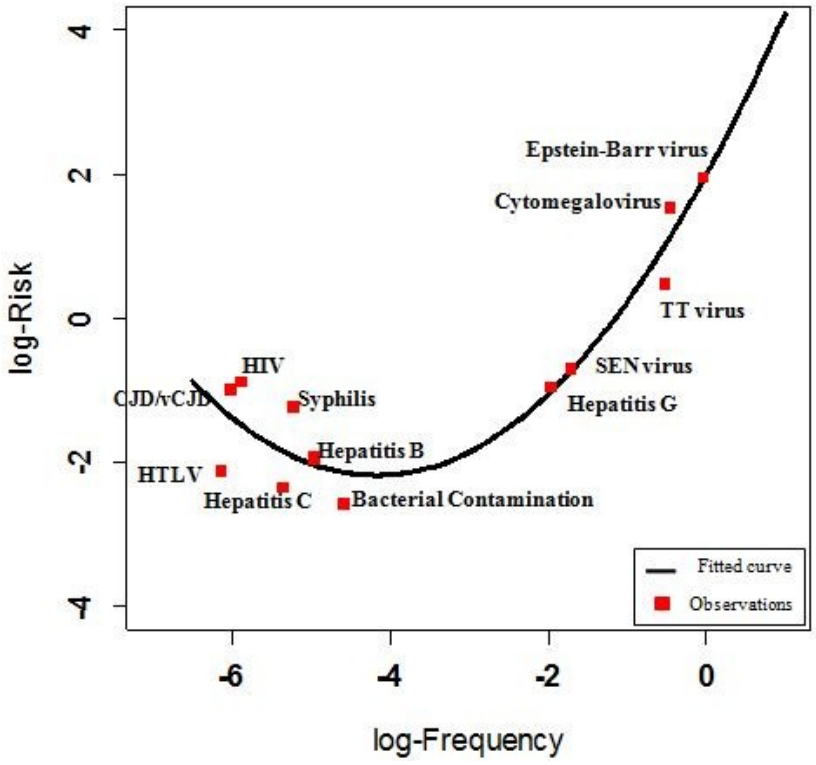


Figure 1.3 Risk estimation according to log-Risk = log-Frequency + log-Severity

The calculated risk values for each infection hazard are shown in the risk assessment matrix (Figure 1.2) for each hazard as "Est." meaning "estimated". The evidence in Figure 1.2 does not show any medium, high or very high risks most likely because risk management of blood transfusions has been focused on lowering such extreme risks. However, this lack of higher risk observations challenged our ability to fully assess the concern that Cox raised about the value of predictions raised by risk assessment matrices. Consequently, we attempted to explore this matter further by using the empirical relationship (Equation (1.4)) we found based on the observed data (Table 1.2).

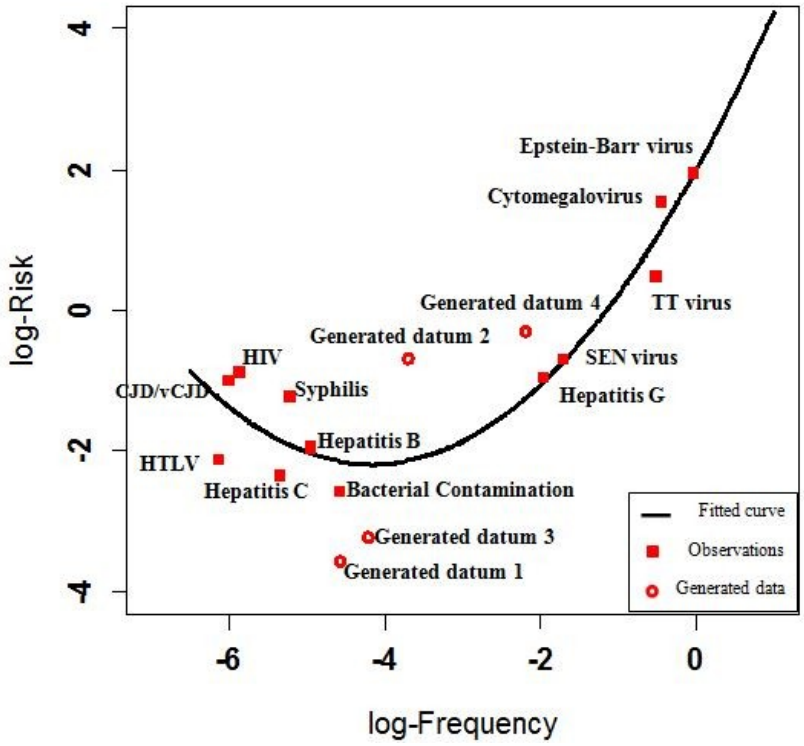


Figure 1.4 Observed and estimated risk for observations and generated data

We sought to populate the risk assessment matrix with some generated risk values that were not found in Table 1.2, but which were consistent with the empirical risk relationship (Equation (1.4). For this purpose, we generated four scenarios with frequencies from the prediction interval limits for the new risk estimation in the middle parts (log-frequency between −4.5 and −2), where there are no experiential frequency estimates for blood transfusion infections hazards and calculated their severities accordingly to populate the risk assessment matrix (Figure 1.4).

We divided the log-frequency gap (−4.5, −2) into three equal parts and selected the two cut points −2.83 and −3.67. The risk estimation for these points using Equation (4) is −1.76 (95% PI: (−3.22, −0.3)) and −2.13 (95% PI: (−3.57, −0.69)), respectively. We generated four data points according to the 95% prediction interval limits of fitted risks. We calculated the corresponding severities from Equation (1.3) and rounded the values to the nearest severity value (Table 1.3).

Table 1.3 Frequency and Severity of Generated Data

Generated Data	**Frequency**	**Risk**	**Severity**
Datum 1	0.00003	0.0003	10
Datum 2	0.00021	0.21	1000
Datum 3	0.00006	0.0006	10
Datum 4	0.005	0.5	100

We illustrated the fitted risk curve defined by product of severity and frequency of the diseases (Figure 1.4). Risks calculated from Equation (1.4) (reported to 1 significant figure to acknowledge the large uncertainty in these data) are shown on the risk assessment matrix in Figure 1.5.

1.3 Results and discussion

1.3.1 Results

The Spearman correlation between log-severity (S) and log-frequency (F) of blood infectious diseases based on PHAC reports (Table 1.2) displays a negative correlation of −0.81 which satisfies the theoretical condition prescribed by Cox for creating a fundamental problem with a risk assessment matrix.

The product of this exercise is the risk assessment matrix shown in Figure 1.2. This is populated according to the blood transfusion hazards provided in Table 1.2, using the categories proposed by the NHS (2008) in Table 1.1. As expected, given the means used for producing it, the risk assessment matrix apparently distinguishes low and medium risks, i.e., the higher colored risk categories have higher quantitative risks (i.e., the observed values as determined in accordance with Equation (1.1) for the quantitative values in Table 1.2). For example, the observed risk value for the Epstein-Barr virus in the medium (yellow) risk category is 90. This estimate is greater than all the observed risk values in the low (green) risk categories, such as TT virus with an observed risk of 3 (Figure 1.2).

The criticism about range compression for the risk assessment matrix is borne out by finding that the low risk category includes observed risks ranging from 0.003 to 3, a risk range of 1000 fold.

In order to test our primary concern, the possibility of the risk assessment matrix making a risk prediction that is worse than random, we had to resort to generating data using the empirical risk relationship (Equation (1.4)) we found for these hazards.

Colored cells are the Risk Categories	Low Risk	Medium Risk	High Risk	Very High Risk

Frequency of Infection	Severity of Consequences				
	Very Low Severity	Low Severity	Medium Severity	High Severity	Very High Severity
Very High Frequency	TT virus [a]Obs 3 [b]Est 10	Cytomegalovirus [a]Obs 35 [b]Est 13 Epstein-Barr virus [a]Obs 90 [b]Est 79			
High Frequency	SEN virus [a]Obs 0.2 [b]Est 0.19 Hepatitis G [a]Obs 0.11 [b]Est 0.10				
Medium Frequency		*generated datum* 4 [b]Est 0.50			
Low Frequency	*generated datum* 3 [b]Est 0.0006		*generated datum* 2 [b]Est 0.21		
Very Low Frequency		*generated datum* 1 [b]Est 0.0003 Bacterial contamination [a]Obs 0.003 [b]Est 0.007	Hepatitis B [a]Obs 0.01 [b]Est 0.01		
Extremely Low Frequency			Hepatitis C [a]Obs 0.004 [b]Est 0.014	Syphilis [a]Obs 0.06 [b]Est 0.01 HTLV [a]Obs 0.01 [b]Est 0.05	HIV [a]Obs 0.13 [b]Est 0.03 CJD/vCJD [a]Obs 0.1 [b]Est 0.04

Figure 1.5 Risk assessment matrix providing colored risk categories plus observed and estimated risk and generated data. a Observed (Obs) risk numbers shown are based on the generic risk function (Risk = Frequency × Severity; Equation (1.1)) and using Table 1.1 entries frequency and severity using Table 1.2 data; b Estimated (Est) risk numbers shown are based on the fitted risk function Equation (1.4); c Generated data.

The four entries in Figure 1.5, labeled "generated datum" 1 to 4, were calculated to provide us with more data observations in the medium risk category. The generated data points 2 and 4 have estimated risk values of 0.21 and 0.50 and both are categorized in Figure 1.5 as medium risks. When compared with TT virus, which was categorized as a low risk in Figure 1.5, we find that it has an estimated (according to Equation (1.4)) risk of 10. This anomaly illustrates the concern posed by Cox (2008), that the risk assessment matrix provides a risk categorization (color code) that is incorrect in relation to an empirical calculation of the risk. Although we had to resort to generating data from an empirical relationship derived from experiential frequency estimates, we have found that the theoretical concern of Cox can be demonstrated for hazard data derived from authentic experience.

1.3.2 Discussion

Given the wide-spread and apparently growing popularity of risk matrices for risk assessment purposes, the prospect of obtaining results that are worse than random is clearly a serious concern. Yet, we have found little practical uptake of Cox's concerns evident in public health relevant literature in the six years since being published. Wieland et al. (2011) referred to the Cox critique of risk assessment matrices in relation to the limited resolution of the method possibly leading to an overestimation of risk for an evaluation of qualitative risk assessment of the spread of African Swine fever. Pickering and Cowley (2010) provide an extensive critique of risk assessment matrices, including citing criticisms by Cox but they do not address the specific concern about risk assessment matrices being worse than random for cases in which there is a negative correlation between frequency and severity of risk. Hubbard and Evans (2010) present a number of arguments against all common judgmental scoring methods for risk assessment, including the steps necessary to construct risk assessment matrices, but they only refer to Cox

with respect to range compression and loss of resolution. Holt et al. (2014) took note of the limitations of risk assessment matrix structure in their review of tools for guiding decisions in relation to assessing risks from pests.

Levine (2012) referred to Cox in criticizing risk assessment matrices for failing to acknowledge uncertainty in the rating of risks according to the axes categories, ignoring information on how to best manage risks or to acknowledge the decision-maker's risk preferences. Levine's main concern was also the range compression, which he proposed to remedy by using logarithmic scales to reflect the large range of values that often exist. Regarding the flaw that Cox has described, Levine only concluded without elaboration: "When used to assess a set of hazards with a negative correlation between frequency and consequence, risk matrices are often uninformative and occasionally misleading."

Ball and Watt (2013) have provided the most complete evaluation of the practical problems with risk assessment matrices. They acknowledge the potential for erroneous risk ranking described by Cox but go further after they observe that he: "Determines that risk matrices are limited in their ability to rank risks correctly and further that they should not be used as they often are, that is, as proxies for risk management decisions by the simple device of overlaying them with colors associated with risk management priorities. This is because optimal resource allocation is quite obviously a function of far more than the two dimensions of likelihood and consequence upon which the matrix rests." Their valid concerns about over-simplification of risk are elaborated by richer, more comprehensive definitions of risk than provided by Equation (1.1) (probability × consequences), which acknowledge the inherently multidimensional character of risk and the inevitable reliance of risk assessment on subjective estimates (Kaplan & Garrick, 1981; Renn,1992; Hrudey, 2000).

In practice, the risk assessment matrix is constructed based on possible hazards, but without any prior assumption on relationship between frequency of hazard and its severity. Our illustration with a tangible public health risk scenario provides insight into limitations of the risk assessment matrix for guiding decision making for the common circumstance where the frequency of hazard and its consequence are negatively correlated. Decision makers need to identify the expected correlation between frequency and severity and recognize that where a negative correlation exists, the risk assessment matrix categorization of risk might not reflect the quantitative risk estimates in accordance with an assumed risk function and may well mislead decision-makers with a worse than random assessment of risk (Cox, 2008).

A tangible, pragmatic approach to the Cox problem for risk assessment matrices has been illustrated in a risk management approach to support the implementation of drinking water safety plans in Alberta, Canada (2013). In this approach, the rating scheme assigns numerical scores for frequency and consequence as well as the generic risk function (risk = frequency × consequence) that thereby defines where in the risk assessment matrix evaluated risks will be plotted. This predefined approach for constructing the risk matrix relies on the validity of the predetermined assigned numerical ratings, but it likely avoids the problems of creating predictions that are worse than random. Of course, all the other practical limitations and associated cautions for the risk assessment matrix that have been summarized earlier remain valid concerns for such simplified applications.

1.4 Conclusions

Our limited validation of the Cox concern, using a tangible public health risk example, suggests a need for careful reconsideration of uses of the risk assessment matrix in risk management. There is no straightforward solution to address the concerns raised about risk

assessment matrices. We do not propose a viable alternative to the risk assessment matrix tool for mapping risks that lack prior knowledge on harm frequency and its severity. However, risk analysts in all fields using the risk assessment matrix should be aware of this limitation. At least, they should investigate or contemplate the plausible correlation between frequency and severity for the hazards to be evaluated in the risk assessment matrix according to their prior knowledge in the field. When some data are available (generally not the case), they could look at data in the manner we did and try to fit a risk function and eventually compare the results with the risk assessment matrix results to identify anomalies.

We do not advocate a wholesale abandonment of risk assessment matrices for guiding risk management, particularly when applied, as they commonly are, to diverse hazards across a broad organizational portfolio. Of course, application of the risk matrix to a diverse range of hazards brings its own complications and challenges that must be acknowledged. The construction and evaluation of a risk assessment matrix can, if used wisely, stimulate a valuable discussion among operational personnel to reflect on what can go wrong and how well prepared the organization is equipped to manage various risks. Provided that the results of a risk assessment matrix exercise are treated with appropriate and healthy scepticism, they can serve a useful purpose for initiating and focusing a discussion about risk priorities within an organization. Achieving healthy scepticism may be difficult as long as risk matrix users see this technique as a simple tool and ignore the embedded complexity involved.

The primary danger revealed in this analysis, owing largely to the pioneering insight offered by Cox (2008), is to avoid allowing such over-simplified risk analyses to become the risk management decision rather than properly being only an operational input that can guide, challenge and inform decision-making to be based on a comprehensive understanding of risk.

Risk assessment matrix outputs should not be allowed primarily to drive or, in the worst case, to become the risk management decision.

Chapter 2

Introduction to microarray technology

Although risk is a core element of public health practice, its definition varies greatly among various public health programs. Several methods have been developed for risk assessment and management in different contexts of public health to better understand disease progression and outcome development, In Chapter 1, we discussed a semi-quantitative approach to risk assessment which provides an insight into risk trends across various scenarios. In the following chapters, we will discuss a quantitative approach to risk assessment using the DNA microarray technology. This platform produces important information that can be useful in understanding disease progression and identifying disease biomarkers.

2.1 DNA microarray technology

Molecular biology research evolves through the development of the technologies like DNA microarray. Researchers are able to investigate a large number of genes in an efficient manner and understand the fundamental aspects underlying traits or diseases.

DNA microarrays are assays for quantifying the types and amounts of messenger RNA (mRNA) transcripts present in a collection of cells. DNA microarray studies involve the collection of biological specimens (e.g., tumor tissue, blood) from subjects; isolation and extraction of RNA; and placement of isolated RNA on the microarray platform (Simon et al., 2003).

The microarray chip consists of a solid surface to which strands of polynucleotides (probes) have been attached in specified positions. The probes for a gene consist of complementary DNA (cDNA) so that the mRNA from a subject binds with the cDNA on the chip if both share sufficient sequence complementarities. The intensity of binding is then quantified into numerical values that represent the amount of gene expression (Simon et al., 2003). Figure 2.1 illustrates the basic principle of cDNA microarray assay of gene expression (Gibson & Muse, 2001). Researchers use one such microarray chip for each subject in their study, ordering chips from a chip manufacturer such as Affymetrix (Affymetrix, 2000).

Using microarrays, researchers are able to measure and study the expression of thousands genes simultaneously. These studies can provide insights into underlying mechanism of diseases by screening genes whose expressions are different between disease cases and controls, or between two groups of patients with and without treatment. They can be used to identify biomarkers of clinical outcome.

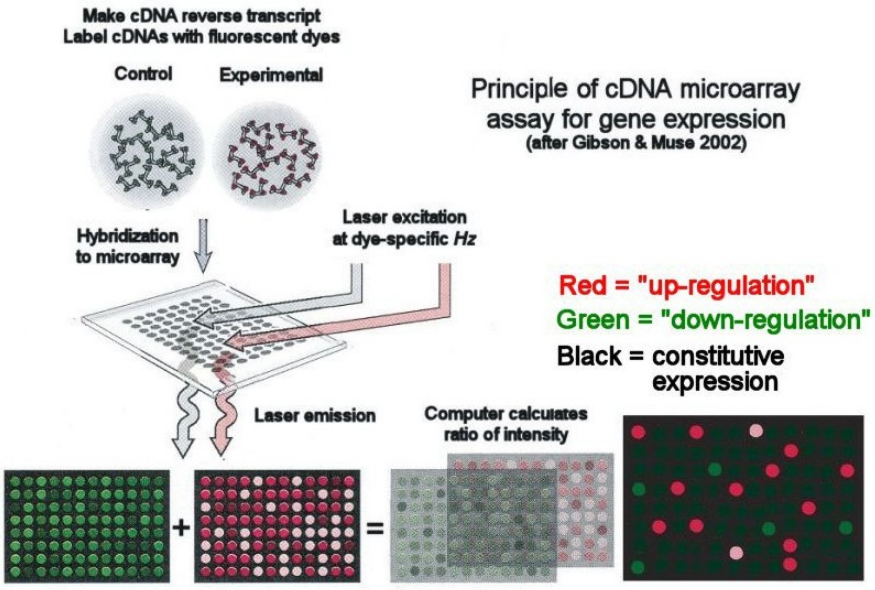


Figure 2.1 principle of cDNA microarray assay of gene expression (Gibson & Muse, 2001)

2.2 Challenges in DNA microarray studies

DNA microarray data comprise very large amount of information on gene expression. It consist of very large number of genes (p) measured on a relatively small number of samples (n). Therefore, classical analysis techniques which consider large n and small p are not applicable to DNA microarray data. This feature is referred to as the high-dimensionality problem and it presents a difficult challenge in the analysis of microarray data.

The second challenge in the analysis of microarray data is the small variability in gene expression measures for some genes. The regular test statistic (e.g., two-sample t-test statistic) gives a very large value because of the small standard deviation. The large value results in statistical significance for genes whose expression means are not differentially expressed.

The third crucial issue is adjusting for multiple testing of thousands of genes. Each statistical test reports the probability of observing a test score by chance assuming no association between gene expressions and the phenotype of interest. Among 10,000 independent tests, even if we set the threshold for p-values as low as 0.01, we will identify 100 of those as "significant" genes just by chance. Various adjustments for multiple testing in microarray data have been introduced (Benjamini, 1995; Storey, 2003). The preferred approach is to control the false discovery rate (FDR) which measures the proportions of false positives among all genes called significant.

Poor reproducibility of important gene lists yielded by independent studies is another problem (Ein-Dor et al., 2006). Most methods do not take into account the possibility of interaction between individual genes, therefore, either fail to observe or detect weak associations. This approach ignores the coordination between genes better described by a pathway structure composed of multiple genes with related biological functions.

2.3 Microarray data analysis

There are two general approaches to study associations of gene expression with diseases or phenotypes in microarray data analysis: Individual Gene Analysis (IGA) and Gene Set Analysis (GSA). IGA examines each gene individually to find differentially expressed genes associated with phenotypes or characteristics. Once a list of significant genes is assembled, we need to identify biological functions or pathways that are over-represented in a given list. An alternative is to identify sets of functionally related genes in advance and to assess whether these gene sets show differential expression.

The focus in expression data analysis has shifted from single gene to gene set level in recent years because many diseases or phenotypes are believed to be associated with modest regulation in a set of related genes rather than a strong increase in a single gene (Subramanian, 2005). However, both approaches can be effective and sometimes their combination is more powerful.

2.4 Individual gene analysis methods

Many individual gene analysis methods have been developed with respect to the characteristics of microarray data, for example Fold Change (DeRisi et al., 1996; Schena et al., 1996), Regularized t-test (Baldi & Long, 2001), Regression Modeling (Thomas et al., 2001), and Significance Analysis of Microarrays (SAM) (Tusher et al.,2001). Among these methods, SAM is the most popular one. We discuss this method in details in the following section.

2.4.1 SAM method

SAM is a popular analytical method that searches for statistically significant genes associated with phenotypes in a microarray data set. SAM is a moderated t-statistic calculated based on permutations of the group labels (e.g. case-control label) adjusted for the multiple hypothesis testing. The permutation test accounts for high dimensionality problem which is the basis of

calculating statistical significance of associations between a gene and the phenotype of interest. Once the test statistic is calculated for the original data, its significance is evaluated by calculating the test statistic for permuted versions of the data set. Under the null hypothesis of no association, the group label is interchangeable. The p-value is then calculated based on the permutation distribution of the test statistic, as the proportion of times the permuted test statistic is as extreme or more extreme than the observed test statistic.

The advantage of SAM over other IGA techniques (e.g. t-test) is that we do not need to assume equal variance and independence of genes. SAM can be applied to various types of phenotypes including continuous and binary phenotypes. Here we discuss the technical details of the SAM for continuous phenotype because this is the focus of our proposed method.

Suppose a matrix X consists of gene expression measurements x_{ij} for gene i and subject j, and y_j denotes phenotype measurement for subject j where $i = 1,2, \dots, p$ and $j = 1,2, \dots, n$. For each gene i, SAM examines the null hypothesis of H_0: there is no association between the gene expressions and the phenotype.

The test statistic d_i is defined as:

$$d_i = \frac{r_i}{s_i + s_0}, \quad i = 1,2, \dots, p, \tag{2.1}$$

where r_i is a linear regression coefficient of gene i on the phenotype, s_i is a standard error of r_i, and s_0 is an exchangeability factor. The details of SAM score components are described below.

$$r_i = \frac{\sum_j y_j(x_{ij} - \bar{x}_i)}{\sum_j (y_j - \bar{y})^2},$$

where $\bar{x}_i = \sum_j \frac{x_{ij}}{n}$, and

$$s_i = \frac{\hat{\sigma}_i}{[\sum_j (y_j - \bar{y})^2]^{1/2}},$$

where $\hat{\sigma}_i$ is the square root of residual error:

$$\hat{\sigma}_i = \left[\frac{\sum_j (x_{ij} - \hat{x}_{ij})^2}{n-2}\right]^{1/2},$$

$$\hat{x}_{ij} = \hat{\beta}_{i0} + r_i y_j,$$

$$\hat{\beta}_{i0} = \bar{x}i - r_i \bar{y}_j .$$

The exchangeability factor s_0 prevents genes whose expression is near zero and unreliable from having large d_i scores. This estimate is expressed as a percentile of the standard deviation of all the genes. Details about calculating s_0 are as follows:

Let s^α be the α percentile of all s_i values and $d_i^\alpha = r_i/(s_i + s^\alpha)$. Compute the 100 quantiles of the s_i values, denoted by $q_1 < q_2 < \cdots < q_{100}$. For each value of $\alpha \in (0, 0.05, 0.1, \ldots, 1.0)$, we calculate v_j of d_i^α as:

$$v_j = mad(d_i^\alpha | s_i \in [q_j, q_{j+1}))$$

where mad is the median absolute deviation from the median divided by 0.64. Then, we define cv(α) as a coefficient of variation of the v_j values and choose $\hat{\alpha} = \text{argmin}[\text{cv}(\alpha)]$. s_0 is fixed at

the value $\hat{s}_0 = s^{\hat{\alpha}}$. The value of s_0 is chosen such that the estimated coefficient of variation of d_i is minimized (Chu et al., 2002).

Steps of SAM procedure:

1. Compute SAM statistic for each gene i.
2. Rank the d_i values, $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(p)}$.
3. Permute the phenotype values y_j, B times. For each permutation b, compute SAM statistic d_i^{*b} and corresponding order statistic $d_{(1)}^{*b} \leq d_{(2)}^{*b} \leq \cdots \leq d_{(p)}^{*b}$.
4. Estimate the expected order statistic from the set of B permutations using $\bar{d}_i = (1/B) \sum_b d_{(i)}^{*b}$ for $i = 1,2, \ldots, p$.
5. Plot the $d_{(i)}$ values against the expected values $\bar{d}_{(i)}$.
6. Gene i is not associated with the phenotype if $d_i \cong \bar{d}_{(i)}$. Genes exhibiting differences greater than a pre-specified threshold Δ are labeled as associated with the phenotype. All genes with positive relative differences, i.e., $d_{(i)} - \bar{d}_{(i)} > \Delta$ are called 'positive significant'. Similarly, all genes with negative relative differences, i.e., $\bar{d}_{(i)} - d_{(i)} > \Delta$ are called 'negative significant'. The smallest and largest cut-points d_i among the significant genes are denoted as $cut_{up}(\Delta)$ and $cut_{low}(\Delta)$.

2.4.2 Multiple hypothesis testing

Microarray data analysis methods test associations of thousands genes with the phenotype of interest, simultaneously. Adjusting for multiple hypothesis tests is essential. A measure of error for single hypothesis test is type I error. Various methods have been developed to estimate an overall measure of error for multiple hypotheses such as family-wise error rate (FWER), Bonferroni, and False Discovery Rate (FDR).

FWER is the probability of at least one false rejection among multiple testing. Suppose we have p genes in a microarray data set and type I error is α, then FWER for p hypothesis testing is $(1-(1-\alpha)^{p})$. When p is very large which is common in microarray studies, this value becomes very high and close to one.

Bonferroni is the classic approach that controls the FWER assuming the genes are independent. To guarantee that FWER is at most α, we reject all p hypothesis tests with a type I error of α/p. Bonferroni is useful for testing small number of genes. However, this method is too conservative for large number of genes in the sense that only very few genes can be significant in this case (Storey & Tibshirani, 2003).

SAM uses FDR which is a popular approach of adjusting for multiple testing. This approach focuses on the proportion of falsely significant genes. Table 2.1 summarizes the outcomes of p hypothesis tests.

Table 2.1 Possible outcomes from p hypothesis tests

True state	Decision rule		Total
	Called not significant	Called significant	
Null	U	V	p_0
Non-null	T	S	p_1
Total	$p-R$	R	p

According to Table 2.1, FDR=V/R, type I error = V/p_0, type II error = T/p_1, and power = $1-T/p_1$. SAM reports FDR for each gene by estimating the proportion of true null genes in the data set. Details of FDR calculation in the SAM is given below.

Steps of FDR calculation

1. Compute the total number of significant genes based on Δ value (from step 6 of SAM procedure). Then we calculate the median and 90th percentile of falsely called genes by computing median and 90th percentile of values from each of the B permutation sets of d_i^{*b} that fall above $cut_{up}(\Delta)$ or below $cut_{low}(\Delta)$ values.
2. Compute $q_1, q_3 = 25\%$ and 75% of the permuted d values.
3. Compute $\hat{\pi}_0 = \#\{d_i \in (q_1, q_3)\}/(0.5p)$, where d_i are the values of the original data set and p is the total number of genes.
4. Choose $\hat{\pi}_0 = \min(\hat{\pi}_0, 1)$.
5. FDR is calculated as the ratio of median or 90th percentile of falsely called genes time $\hat{\pi}_0$ divided by the number of significant genes (Chu et al., 2002).

2.5 Gene set analysis methods

In most studies, IGA methods lead to a list of many significant genes even after multiple test adjustments have been made. The interpretation of such a large list of genes is complicated. According to IGA methods, significance of genes is highly affected by the arbitrary cut-off values chosen by researchers. Sometimes, these methods show weak to moderate associations for some genes and as a result those genes are removed from the list of significant genes (Nam & Kim, 2008). Moreover, replication of the findings from IGA in different microarray experiments is another serious challenge (Ein-Dor et al., 2005; Ein-Dor et al., 2006).

Molecular biologists have compiled lists of genes grouped by their common biological functions which are called biological pathways. There are various pathway databases that are freely available for microarray data analysis such as Kyoto Encyclopedia of Genes and Genomes

(KEGG) (Kanehisa & Goto, 2000), Gene Expression Omnibus (Edgar et al., 2002), Biocarta (Nishimura, 2001), and Molecular Signature Data Base (Liberzon et al, 2011).

A variety of GSA methods have been developed with the aim to identify gene sets associated with phenotypes in DNA microarray studies. These methods incorporate previous biological knowledge of presumably related genes within a gene set and hence are more powerful in finding associations with phenotypes. GSA methods are different in terms of the methodological assumptions related to definition of a sample and formulation of the null hypothesis. Extensive methodological discussions and reviews are given by Goeman and Buhlmann (2007), Nam and Kim (2008), and Maciejewski (2014). We briefly discuss important aspects of GSA methods.

There is a need to deal with many challenges in GSA methods due to characteristics of the data:

1. The number of gene set is far larger than the number of observations.
2. Gene expression measurements, especially within each gene set can be highly correlated.
3. Number of pathways is increasing rapidly. Efficient GSA methods are required to address the computational burden of testing thousands gene sets.

The GSA methods are broadly classified as ‘self-contained’ or ‘competitive’. Competitive methods compare the associations for genes within the gene set with associations for genes in the gene set complement to determine whether genes in a particular gene set are associated more with a phenotype as compared to genes outside the gene set. Examples of competitive gene set methods for analysis of gene expression studies are gene set enrichment analysis (GSEA) (Subraminan et al., 2005), SAFE (Barry et al., 2005), Random set methods (Newton et al., 2007), and GSA (Efron and Tibshirani, 2007).

In contrast, self-contained methods assess the association between the phenotype and expression of the gene set of interest ignoring other genes that are not in the gene set. Examples include Global test (Goeman et al., 2004), ANCOVA (Mansmann and Meister, 2005), SAM-GS (Dinu et al, 2007), and LCT (Dinu et al., 2013).

Competitive methods are based on the untenable assumption that genes are independent. Genes can be highly correlated, especially those within a gene set. These methods use expression measurements for genes outside the gene set of interest. However, self-contained methods only use expression measurements for the genes in the gene set under study, an approach following closely the statistical hypothesis testing framework.

The key methodological distinction between the two approaches is inherent to the gene-sampling versus subject-sampling concept. The term 'sampling' refers to permutation test used in GSA methods to estimate the null distribution. Competitive methods use genes as the sampling units whereas self-contained methods use subjects as sampling units. Under the self-contained null hypothesis of no association between the gene sets and the phenotype, labels are interchangeable and the null distribution is estimated based on permuting the labels of subjects. Under the competitive null hypothesis of no differential expression of genes in the gene set of interest compared with expression of genes not in the set, we assume that genes are independent and the null distribution is estimated based on permuting the genes (Geoman & Buhlmann, 2007).

Geoman and Buhlmann (2007) strongly discourage using competitive methods due to invalid statistical independence assumption across genes. Delongchamp et al. (2006) commented on how ignoring the correlations within the gene sets can overstate significance and proposed meta-

analysis methods for combining p-values with a modification to adjust for correlation. Chen et al. (2007) argue their preference for the self-contained hypothesis over the competitive one because the p-values computed under the former are consistent with the principle of statistical significance testing, while the p-values computed under the latter do not take into account correlations among genes. Our focus here is on self-contained methods which preserve correlations within gene sets.

2.6 GSA methods for continuous phenotypes

Most of GSA methods have been developed for binary or categorical phenotypes. The urge of improving methods for continuous phenotype is increasing on the ground that quite often the outcome of interest is measured as a continuous variable, for example, tumor volume, birth weight, metabolites or proteins. In such cases it is neither easy nor meaningful to dichotomize or categorize continuous phenotypes. Some specific ranges may fail to express underlying biological function for each subject. Moreover, these ranges are arbitrary defined by specialists and different specialists might use different ranges according to the patient's health condition. It would be beneficial to directly analyze continuous phenotypes in DNA microarray studies. We discuss here GSA methods for continuous phenotypes.

2.6.1 Significance Analysis of Microarrays for Gene Sets (SAM-GS)

SAM-GS is an extension of SAM which accommodates gene set analysis proposed by (Dinu et al., 2007). This method uses the sum of squares of ratio between the regression coefficient for an individual gene and its corresponding standard error. Basically, it combines moderated t-statistic of single genes into a measure of association of a gene set with the phenotype.

For a given gene set S of size s, the SAM-GS test statistic is calculated as the L_2 norm squared of the vector $d = (d_1, d_2, \dots, d_s)$:

$$SAM - GS = \sum_{i=1}^{s} d_i^2, \quad (2.2)$$

where d_i is the SAM score estimated by (2.1) for each gene i. The Permutation test is used to assess significance of the gene set S. When a collection of gene sets is tested, FDR adjustment for multiple hypothesis tests is used.

2.6.2 Global test

The Global method is based on the generalized linear regression framework in which the distribution of the phenotype is modelled as a function of the covariates. For a continuous phenotype linear regression model is used. We assume we have gene expression measurements of n subjects for p genes. Let $X = (x_{ij})$ denote the $n \times s$ data matrix containing only s genes in the gene set of interest and Y as the $n \times 1$ vector containing the phenotype. We define:

$$E(Y_i|\beta) = \alpha + \sum_{j=1}^{s} x_{ij}\beta_j \quad (2.3)$$

where α is an intercept, and β_j is the regression coefficient for gene $j = 1,2, \dots, s$. Whether there is an association between the gene expression and the phenotype is equivalent to testing the hypothesis:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_s = 0.$$

It is not possible to test this hypothesis using a classical approach because s might be large relative to n. To tackle this problem we assume that $\beta_1, \dots, \beta_s$ are samples from some common distribution with expectation zero and variance τ^2. Then the single unknown parameter τ^2 determines how much the regression coefficients deviates from zero. The null hypothesis becomes:

$$H_0\colon\ \tau^2 = 0.$$

Let $r_i = \sum_{j=1}^{s} x_{ij}\beta_j, i = 1,2, \dots, n$ be the linear predictor, the total effect of all covariates for person i, then $r = (r_1, \dots, r_n)$ is a random vector with $E(r) = 0$ and $Cov(r) = \tau^2 XX^T$. We can simplify the model (2.3) in a simple random effect model in which each subject has a random effect that influences its phenotype:

$$E(Y_i|r_i) = h^{-1}(\alpha + r_i). \quad (2.4)$$

The test statistic under the null hypothesis can be described as:

$$Q = \frac{(Y-\mu)'R(Y-\mu)}{\mu_2}, \quad (2.5)$$

where $\mu = h^{-1}(\alpha)$ is the expectation of Y under H_0, $R = (1/s)XX^T$ is an $n \times n$ matrix proportional to the covariance matrix of the random effects r, μ_2 is the second central moment of Y under H_0. There is no computational problem to estimate the distribution of the test statistic Q because it only involves the small $n \times n$ covariance matrix R between the samples and not the large $s \times s$ covariance matrix between genes (Goeman et al., 2004).

2.6.3 Linear Combination Test (LCT)

LCT incorporates the gene expression covariance matrix into the test statistic to take into account the correlation among gene expressions. Suppose the gene expression data consists of n subjects with phenotype $Y_1, Y_2, \dots, Y_n$ and a predefined gene set S contains the gene expression measurements of n subjects for p genes $\{X_1, X_2, \dots, X_p\}$. We test the null hypothesis that the gene set is not associated with the phenotype. This multivariate hypothesis can be rewritten as H_0: no linear combination of $X_1, X_2, \dots, X_p$ is associated with the phenotype of interest. The linear combination of p genes can be written as $Z(\boldsymbol{\beta}) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. For a given vector of coefficient β, H_0: can be analyzed in the framework of univariate regression:

$$Y_i = \alpha_0 + \alpha_1 Z_i(\boldsymbol{\beta}) + e_i, \quad (2.6)$$

where α0 and α1 are the intercept and slope respectively, ei ~ N(0, σ2) where i denotes subjects 1,...,n. This is a classical simple linear regression problem.

For testing H_0, we consider the linear combination with the maximum correlation with the phenotype among all possible linear combinations, i.e.,

$$\boldsymbol{\beta}^* = \text{argmax}_{\boldsymbol{\beta}} \rho^2_{\mathbf{Y},Z(\boldsymbol{\beta})} \quad (2.7)$$

where $Z(\beta) = \beta^T \mathbf{X}$, and the square of the correlation between Y and $Z(\boldsymbol{\beta})$ is:

$$\rho^2_{\boldsymbol{Y},Z(\boldsymbol{\beta})} = \frac{Cov(\boldsymbol{Y}, Z(\boldsymbol{\beta}))^2}{\sigma^2_Y \sigma^2_{Z(\boldsymbol{\beta})}}. \quad (2.8)$$

$\sigma^2_{\boldsymbol{Y}}$ is a constant value and we can ignore it in the derivation of the test statistic. Then we have:

$$\begin{aligned}\sigma^2_{Z(\boldsymbol{\beta})} &= E\left[(\boldsymbol{\beta}^T X - E\left[\boldsymbol{\beta}^T X\right])^2\right] \\ &= E[(\boldsymbol{\beta}^T (X - E[X]))^2] \\ &= \boldsymbol{\beta}^T E[(X - E[X])^2]\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T \hat{\Omega}\boldsymbol{\beta}.\end{aligned}$$

$$\begin{aligned}Cov(\boldsymbol{Y}, Z(\boldsymbol{\beta}))^2 &= E[(\boldsymbol{Y} - E[\boldsymbol{Y}])\left(\boldsymbol{\beta}^T X - E\left[\boldsymbol{\beta}^T X\right]\right)]^2 \\ &= \boldsymbol{\beta}^T E[(\boldsymbol{Y} - E[\boldsymbol{Y}])(X - E[X])]E[(\boldsymbol{Y} - E[\boldsymbol{Y}])(X - E[X])]^T \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^T Cov_{\boldsymbol{Y},X} Cov^T_{\boldsymbol{Y},X} \boldsymbol{\beta}\end{aligned}$$

and we can simplify the equation (2.8):

$$\rho^2_{Y,Z(\boldsymbol{\beta})} = \frac{\boldsymbol{\beta}^T Cov_{\boldsymbol{Y},X} Cov^T_{\boldsymbol{Y},X} \boldsymbol{\beta}}{\boldsymbol{\beta}^T \hat{\boldsymbol{\Omega}} \boldsymbol{\beta}},$$

where $Cov_{\boldsymbol{Y},\boldsymbol{X}} = (Cov(\boldsymbol{Y}, \boldsymbol{X}_1), \ldots, Cov(\boldsymbol{Y}, \boldsymbol{X}_p))^T$ and $\hat{\boldsymbol{\Omega}}$ is the gene expression covariance matrix with the hh'-th entry being:

$$\omega_{hh'} = \frac{1}{n-1}\sum_{l=1}^{n}(x_{hl} - \overline{x}_h)\left(x_{h'l} - \overline{x}_{h'}\right).$$

The optimization problem can be written as:

$$\rho^2_{\boldsymbol{Y},Z(\boldsymbol{\beta})} = \frac{\boldsymbol{\beta}^T \boldsymbol{A}\boldsymbol{\beta}}{\boldsymbol{\beta}^T \boldsymbol{B}\boldsymbol{\beta}}$$

where $\boldsymbol{A} = Cov_{\boldsymbol{Y},X} Cov_{\boldsymbol{Y},X}^{T}$ and $\boldsymbol{B} = \widehat{\boldsymbol{\Omega}}$. This optimization problem can be solved by $\boldsymbol{\beta}^{*}$, the maximal eigen vector of $\boldsymbol{AB}^{-1}$ and $\rho^{2}_{\boldsymbol{Y},Z(\boldsymbol{\beta}^{*})}$ is the corresponding eigenvalue (Johnson, 2002).

When the gene set size *p* is larger than the sample size *n* which is a common situation in GSA, the covariance matrix **B** is singular. A possible way to deal with this problem is using a shrinkage covariance matrix proposed by Schafer and Strimmer (2005). We replace the singular covariance matrix $\widehat{\boldsymbol{\Omega}}$ with a shrinkage covariance matrix $\widehat{\boldsymbol{\Omega}}^{*}$, given by $\omega_{\mathrm{hh}'}{}^{*} = \rho^{*}_{\mathrm{hh}'}\sqrt{\omega_{\mathrm{hh}}\omega_{\mathrm{h}'\mathrm{h}'}}$ with shrinkage coefficients:

$$\rho^{*}_{hh'} = \begin{cases} \rho_{hh'} min\left\{1, \max\left(0, 1-\hat{\lambda}^{*}\right)\right\}, & h \neq h' \\ 1 & h = h' \end{cases} \qquad (2.9)$$

where $\rho_{hh'}$ is the sample correlation between h-th and h'-th genes, and λ^{*} is the shrinkage intensity estimated by:

$$\hat{\lambda}^{*} = \sum_{h \neq h'} var(\rho_{hh'}) \,/\, \sum_{h \neq h'} \rho_{hh'}{}^{2}. \qquad (2.10)$$

Incorporating the covariance matrix estimator into the test statistic leads to high computational cost. To tackle this problem the orthogonal transformation of the original gene expression measurements is obtained using eigenvalue decomposition of the shrinkage covariance matrix, i.e., $\widehat{\boldsymbol{\Omega}}^{*} = \boldsymbol{UDU}^{T}$. The orthogonal basis vectors is computed by $\left(\mathrm{V}_1, \ldots, \mathrm{V}_\mathrm{p}\right) = \left(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p}\right)\boldsymbol{UD}^{-1/2}$. Hence, the square of the correlation is rewritten as:

$$\rho^{2}(\boldsymbol{\gamma}) = \frac{\boldsymbol{\gamma}^{T} Cov_{\boldsymbol{Y},V} Cov_{\boldsymbol{Y},V}^{T} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^{T}\boldsymbol{\gamma}},$$

where $\boldsymbol{\gamma} = D^{1/2}U^{T}\boldsymbol{\beta}$ and $Cov_{\boldsymbol{Y},V} = (Cov(\boldsymbol{Y}, \boldsymbol{V}_1), \ldots, Cov(\boldsymbol{Y}, \boldsymbol{V}_p))^{T}$. The coefficients of the most significant combinations are given by $\boldsymbol{\gamma}^{*} \propto Cov_{\boldsymbol{Y},V}$ (Schafer & Strimmer, 2005). Therefore,

the LCT statistic is proportional to the sum of the covariance squared between the phenotype and orthogonal transformation of gene expression measurements:

$$\rho^2(\boldsymbol{\gamma}^*) = c\sum_{j=1}^{p} Cov(\boldsymbol{Y}, \boldsymbol{V}_j)^2,$$

where c is a constant. We use a permutation test (permuting phenotype labels) to evaluate the statistical significance against the null hypothesis for this test statistic. This approach is efficient because we only need to compute $\hat{\Omega}^*$ once for the original data and not for each permuted data set.

2.7 Critical needs in GSA

A gene set can be significant only because a subset of genes within the set is actually differentially expressed, and the rest of the genes may not be contributing to the set significance. In fact, a large set may be easily identified as significant only because one gene is associated with the phenotype. It is very important to assess significant gene sets to identify only those core members that are associated with the phenotype, as a core subset.

Identifying core subsets provides an efficient way to gain biological insights into the disease mechanism. Reduction to the most predictive genes is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis, intervention strategies and tailored treatment. Limiting the number of genes can lead to a change of platform from high-dimensional microarray technology to alternate methods, such as real time polymerase chain reaction (PCR) assays that are cheaper and faster. The alternate methods are easily applicable to a routine clinical setting for diagnosis purposes (West et al., 2006; Pittman et al., 2004).

Dinu et al. (2008) developed a gene set reduction method, referred to as SAM-GS reduction (SAM-GSR) for identifying the core subset for a binary phenotype. No methods have been introduced to address gene set reduction for a continuous phenotype yet. In this section, we review the SAM-GSR analysis for a binary phenotype and in the next chapter, we address the problem of finding differentially expressed core genes for a continuous phenotype.

2.7.1 Gene set reduction for binary phenotype

We discuss here the procedure of gene set reduction for a binary phenotype. The gene set reduction process follows two main parts:

1. identifying significant gene sets associated with the phenotype of interest,
2. extracting the core subsets from significant gene sets.

Dinu et al. (2008) extended SAM-GS analysis to extract the core subsets of gene sets that are differentially expressed by a binary phenotype. For a given gene set S, SAM-GS statistic is the L_2 norm of the t-like statistics,

$$SAM - GS = \sum_{i=1}^{s} d_i^2,$$

where $d_i = (\bar{x}_1(i) - \bar{x}_2(i))/(s(i) + s_0)$ is estimated for each gene i, $\bar{x}_1(i)$ and $\bar{x}_2(i)$ are the sample average of each group of the phenotype, $s(i)$ is a pooled standard deviation over the two groups and s_0 is a small positive constant that adjusts for the small variability in microarray measurements. Permutation test is used to obtain the statistical significance of gene set S (Dinu et al, 2007).

Given a statistically significant gene set S, we use the following principle to extract core members: for a pair of genes (i, j) in S, $|d_i| > |d_j|$ suggests that gene j belongs to subsets only if

gene i belongs to the subset. This principle is motivated by the fact that d_i^2 is the contribution of each gene to the test statistic and the core subset must consist of genes with larger contributions (Dinu et al., 2008).

We follow the next steps to gradually partition the set S into subsets:

1. Calculate the SAM statistic d_i for each within the gene set S.
2. Select the first k genes ($k = 1, \dots, s-1$) with the largest statistic $|d|$ to form a reduced set R_k. Let $\bar{R}_k$ be the complement of R_k in S, and c_k be the SAM-GS p-value of $\bar{R}_k$.
3. The reduced set R_k corresponds to the least k such that c_k is larger than a threshold c, chosen by analyst.

By removing genes with joint statistical significance, as a set, above a threshold $c_k > c$, we ensure that we keep member of a set that are not significant by themselves, but collectively form a set that becomes significant (Subramanian et al., 2005).

We do not use criteria such as the FDR cut-off to extract core subsets because FDR corresponds to each gene while using this approach we combine the contribution of each gene into an overall measure of association. Hence, we take into account correlations among genes and their tendency to work together towards the significance. A set consisting only of moderately associated genes can still be significant (Dinu et al, 2008).

The rationale behind using c_k over p_k for selecting core members is that even only one significant gene can make the reduced subset significant. The p_k value can be very small, in some scenarios all close to zero, even if the $\overline{R}_k$ contains genes that are associated with the phenotype. Hence, using p_k as a cut-off is not useful in partitioning the gene set into two subsets

of core genes and redundant genes. On the other hand, using c_k we are able to choose different cut-off values from more conservative such as 0.01 to more liberal such as 0.1. Therefore, we have more flexibility to choose members of the core subset (Dinu et al., 2008).

Chapter 3

Methods

In this chapter, we describe our proposed method of gene set reduction for microarray gene expression studies with continuous phenotypes. First, we assess the association of gene set expressions with a continuous phenotype. Given significant gene sets, we apply our procedure to identify core subsets that chiefly contribute to the association. We analyzed the performance of the LCT-GSR method using two real microarray gene expression data.

3.1 Identification of significant gene sets for continuous phenotypes

Genes within gene sets are expected to be correlated because they share similar biological functions and the same chromosomal locations. Among GSA methods for continuous phenotypes, the LCT method efficiently incorporates the gene expression covariance matrix into the test statistic. This characteristic is desired in the GSA method because it leads to a powerful and computationally efficient approach for evaluating the association of a gene set with a continuous phenotype (Dinu et al., 2013).

We use the LCT method to evaluate associations of gene sets with continuous phenotypes. Since the number of genes in the gene sets is much larger than the number of subjects the covariance matrix is singular. We overcome this problem by using a shrinkage covariance matrix estimator. Then, we perform eigenvalue decomposition of the shrinkage covariance matrix for the original data to reduce the high computational cost of integrating this estimator. If the covariance matrix is $\widehat{\Omega}^{*} = \mathrm{UDU}^{\mathrm{T}}$ then the orthogonal basis vectors are $(\mathrm{V}_1, \ldots, \mathrm{V}_\mathrm{p}) = (\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p})\mathrm{UD}^{-1/2}$. Therefore, the LCT statistic is defined by:

$$\rho^2(\boldsymbol{\gamma}^*) = c\sum_{j=1}^{p} Cov(\boldsymbol{Y}, \boldsymbol{V}_j)^2,$$

where $\gamma = D^{1/2}U^T\beta$ and β is the vector of regression coefficients. Permutation test is used to assess the statistical significance against the null hypothesis. We described details of this method in section (2.6.3).

We have a matrix of p gene expression measurements for n subjects as shown in Table 3.1. To incorporate gene sets information we need to link the gene set data set to this matrix. We create a new matrix M refer to as 0/1 matrix to check whether a gene from the gene expression data exists in the gene set data set. The rows of the 0/1 matrix represent p genes and the columns represent l gene sets. M_{ij} is defined as 1 if the i-th gene from the list of microarray gene is part of the j-th gene set, and 0 otherwise. This matrix shown in Table 3.2 is used as an input to the LCT analysis.

Table 3.1 An example of microarray gene expression data set

Gene name	Subject 1	Subject 2	…	Subject n
Gene 1	14.16	13.95		14.55
Gene 2	9.41	11		11.25
⋮				
Gene p	9.89	9.95		8.82

Table 3.2 An example of 0/1 matrix

Gene name	Gene set 1	Gene set 2	…	Gene set l
Gene 1	0	1		0
Gene 2	1	0		0
⋮	⋮	⋮	⋮	⋮
Gene p	0	1		0

3.2 Identification of core genes for continuous phenotypes

We apply gene set reduction method to the list of genes identified as significant by LCT analysis to obtain core genes. To the best of our knowledge, there are no methods for reducing gene sets to their core subsets for continuous phenotypes. In this section, we discuss our proposed method for gene set reduction for continuous phenotypes. We develop the method referred to as LCT-GSR based on the concepts used in the SAM-GSR. We use SAM values to measure the magnitude of association between each gene and the phenotype of interest.

3.2.1 LCT-GSR algorithm

For each significant gene set, we repeat the following steps. Given the significant gene set S with s genes,

1. Apply SAM to all individual genes and calculate SAM statistic d_i.
2. For $k = 1,2, \dots, s-1$, select the first k genes with largest statistic $|d_i|$ to form a reduced set R_k. Let $\bar{R}_k$ be the complement gene set of R_k in S, and c_k be the corresponding LCT p-value of the complement gene set.
3. Select the reduced set when c_k is larger than a pre-specified threshold c, chosen by the analyst.

We compute SAM statistic d_i defined by:

$$d_i = \frac{r_i}{s_i + s_0}, \qquad i = 1,2, \dots, p,$$

where r_i is the linear regression coefficient of expression measurements for gene i on the phenotype, s_i is the pooled standard error of r_i, and s_0 is the exchangeability factor or a small positive constant that adjusts for the variability in the microarray measurements.

We order the genes within the gene set according to the absolute value of their SAM values, $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(p)}$. We gradually remove the gene with the largest $|d_i|$ and apply the LCT analysis to the complement gene set $\overline{R}_k$ to calculate its p-value c_k. If $c_k < c$, we still have significant members within the complement gene set that are associated with the phenotype which make the whole set statistically significant. If $c_k > c$, there are no significant genes remained contributing to the significance of the complement gene set and we stop the procedure. When we reach the threshold, the genes within R_k represent the core subset.

The threshold value can be arbitrary chosen by the researcher based on the biological importance of the genes associated with the phenotype. This value can be flexible for each gene set, i.e., we can use different cut-off values for different gene sets. We used $c = 0.1$ as previously used by Dinu et al. (2008) for gene set reduction with a binary phenotype. We used a threshold slightly more conservative to ensure we included genes that individually may not be associated with the phenotype but collectively have a biological impact on the phenotype of interest.

Since we test the significance of multiple gene sets, we calculate FDR to adjust for multiple hypothesis testing as described by Storey (2002).

In chapters 4 and 5, we apply our method LCT-GSR to two real microarray studies to evaluate its performance. We describe each study in detail and test the association between the gene expression measurements and the continuous phenotype of interest. We report significant gene sets and their core subsets, accordingly.

Chapter 4

Prostate cancer: data description & results

4.1 What is prostate cancer?

Prostate cancer is a disease where some prostate cells have lost normal control of growth and division, and as a result, do not function as healthy cells. It can be very slow-growing and some men who develop prostate cancer may live many years without ever having the cancer detected. However, the chance of survival with prostate cancer is greatly increased by early detection of the disease. The prostate cancerous cells have uncontrolled growth, abnormal structure or the ability to spread to other parts of body (invasiveness) (Prostate Cancer Canada, 2015). Prostate cancer is described as clinically localized disease when cancerous cells are located completely inside the prostate gland.

Prostate cancer is the most common cancer in men. One in eight men will be diagnosed with the disease in their lifetime. It is estimated that in 2015, 24000 Canadian men will be diagnosed with prostate cancer and 4100 will die from the disease according to Prostate Cancer Canada. A major dilemma in prostate cancer management is how to treat patients with clinically localized disease. The death rate can be significantly reduced by improved testing and better treatment options.

4.2 Testing and diagnosis

Imaging Technology

Imaging technology such as CT scan, bone scan and MRI is increasingly used for prostate cancer diagnosis. Computed Tomography (CT) Scan uses x-ray to capture cross-sectional images

of organs, tissues, bones and blood vessels. These images are usually useful in men with prostate cancer to determine whether the cancer has spread to nearby structures such as lymph nodes.

Bone is the most common site for prostate cancer spread. A bone scan is done in men where there is clinical possibility of cancer having spread to the bone. A bone scan uses radiopharmaceuticals and a computer to create an image of the bones.

Magnetic Resonance Imaging (MRI) uses strong magnets and radio waves to create 3D images of organs. MRI is useful for identifying abnormal areas within the prostate that are suspect for cancer and MRI shows how invasive the cancer is (Prostate cancer Canada, 2015).

These approaches are mostly helpful for identifying whether prostate cancer is spreading to other organs but they may not be helpful for early detection of the disease.

Prostate Specific Antigen (PSA)

Prostate Specific Antigen (PSA) is a protein produced by cells in the prostate gland. PSA is secreted into seminal fluid and is measured in nanograms per milliliter of blood (ng/ml). There are two types of PSA, free PSA that moves freely in the blood and complex PSA that is attached to other proteins in the blood. Prostate cancer cells produce more complex PSA. We can measure the amount of PSA protein in the blood using a simple blood test referred to as the PSA test. Higher levels of PSA may indicate the presence of cancer (Prostate Cancer Canada, 2015).

There are some benefits in using the PSA test but it also has some limitations. For example, PSA may be an indicator of the presence of cancer in its early stages but can also lead to unnecessary tests and treatment. The PSA test is a simple blood test but it cannot distinguish between slow growing and aggressive cancer. A high level of the PSA test can only tell us if

there is a problem with the prostate but can not necessarily diagnose prostate cancer. The PSA test is used as red flag for follow-up. In Canada, the PSA test is used to monitor responses to cancer treatment or to monitor disease recurrence or progression rather than useing it widely as a screening tool (Prostate Cancer Canada, 2015).

Prostate biopsy

A prostate biopsy is conducted to determine whether suspicious looking cells and tissues are cancerous or not. A biopsy needle is inserted into the rectum using ultrasound as a visual aid to guide the needle through the rectum using a local anesthetic to allow removal of a tissue samples. About eight to twelve samples will be taken depending on the area to be examined (ProstateC Canada, 2015).

Grading

Pathologists examine biopsied tissue samples of the prostate under a microscope and compare the cancer tissue pattern with the normal tissue cells to determine the grade of prostate cancer for each biopsy sample. There are two systems for grading cancers: the General Grading System and the Gleason Grading System (Prostate Cancer Canada, 2015).

The General Grading System classifies prostate cancer cells as low, intermediate or high grade based on the cell appearance in relation to healthy prostate cells, abnormal or extremely dissimilar prostate cells.

The Gleason Grading System is a rating ranging from 2 to 10 that attempts to predict the aggressiveness of prostate cancer. A higher value means more aggressive cancer which is more likely to spread to other parts of body. The Gleason score is regarded as the best predictor of

cancer progression and growth. Overall, the Gleason score is the sum of primary and secondary grade, each ranging from 1 to 5.

To determine the primary Gleason grade, pathologists look at the most predominant tumor pattern to identify the grade of cancerous cells. They assign a score from 1 to 5 to the pattern based on the difference between the healthy and cancerous cells, i.e., larger differences will imply larger Gleason grades. The secondary Gleason grade is determined in a similar way by pathologists looking at the second most common pattern.

Figure 4.1 illustrates the schematic diagram of Gleason grading system. Grade 1 is assigned to the mass of evenly spaced and uniform shaped glands with no evidence of invasion of the tissue. Grade 2 is assigned to some invasion into the surrounding tissues and more variation in gland size and spacing. Grade 3 is the most common grade with less defined boundaries and more variation in shape, size and space between glands. Grade 4 characterized by gland formation with a ragged invasive edge. Grade 5 is given to a pattern with complete absence of gland formation versus clusters of cells. Grade 1 and 2 are defined as well differentiated while Grade 3 is moderately differentiated, Grade 4 is poorly differentiated and Grade 5 is undifferentiated.

The scores break down is shown below:

- Scores from 2 to 4 are very low on the cancer aggression scale.
- Scores from 5 to 6 are mildly aggressive.
- A score of 7 indicates moderately aggressive.
- Scores from 8 to 10 are highly aggressive.

The Gleason score usually is reported as (primary Gleason grade, secondary Gleason grade). Both Gleason grades of (3,4) and (4,3) give Gleason total scores of 7, however, not all Gleason

scores are equivalent, i.e., $3 + 4 \neq 4 + 3$. Someone with Gleason grades of (3,4) is actually in a little better condition than a grade (4,3). When a primary grade is 3, it means the cancer has not advanced as far with cellular deterioration (i.e., less aggressive) versus cancer with a primary grade of 4 in the predominant cancerous area.

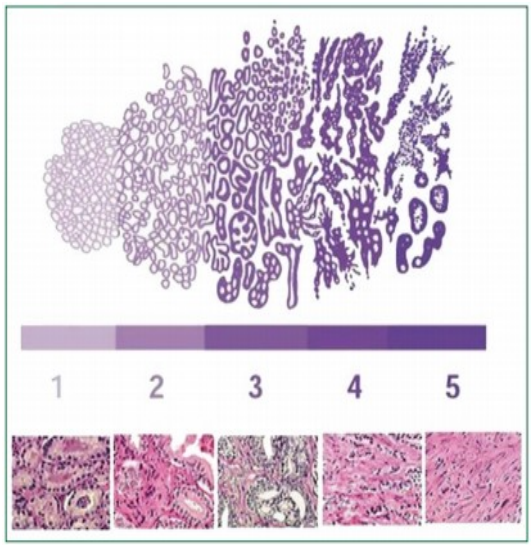


Figure 4.1 Schematic diagram of Gleason Grading System. Lower grades are associated with small, closely packed glands. As grade increases cells spread out and lose glandular architecture

Tumor volume

Tumor volume is defined as the percentage of the prostate occupied by the tumor. Tumor volume assessment was conducted with the aid of a grid, a plastic strip or ruler with squares of 3.0 mm as described by Humphrey and Vollmer (1990). During a microscopic examination, the areas of the gland that were invaded by a tumor are outlined using a pen with permanent ink. The marked slides are then put on top of a grid and the percentage of squares that are occupied by the tumor is calculated in relation to the whole area covered by the specimen. The tumor volume corresponds to the gland area occupied by the tumor and its absolute value is calculated by multiplying the tumor percentage by the gland's total weight (Kato et al., 2008). The tumor

volume value is a significant predictor of cancer risk closely tied to the likelihood of tumor progression and to survival time (Humphrey & Vollmer, 1990).

4.1.2 Challenges in prostate cancer management

A major dilemma in prostate cancer management is how to treat patients with clinically localized disease, prostate cancer that appears to be completely inside the prostate gland. The current prostate cancer prognostic models are based on prostate specific antigen (PSA) levels, Gleason score, and clinical staging. In practice, these models are inadequate to accurately predict disease progression specifically for men who fall within an intermediate range characterized by a PSA level between 4-10 ng/ml and a Gleason score of 6 or 7 (Sboner et al., 2010).

The benefit from radical prostatectomy, surgery that completely removes the prostate gland and surrounding tissue is often modest (Bill-Axelson et al., 2008). Specifically, the 5- to 10-year mortality following the diagnosis of prostate cancer is relatively low, regardless of the type of treatment (including radical prostatectomy) that patients receive (Bill-Axelson et al., 2005). This finding suggests that watchful waiting is an important approach for many localized prostate cancer patients. In practice, such an approach is only effective if we can identify a subset of patients who have high risk of disease progression and could benefit from active treatments. There is a need for identifying patients who must be treated and who can safely be monitored for disease progression. We reason that by investigating the gene expression measurements of prostate cancer patients, we would be able to gain insights into underlying mechanism of prostate cancer disease progression.

4.2 Data Description

The prostate cancer data set is part of the Swedish Watchful cohort study nested in a cohort of men with localized prostate cancer (1977-1999) with up to 30 years of clinical follow up (Sboner

et al., 2010). The study design was approved by the Ethical Review Boards in Örebro and Linköping. The cohort consists of 255 patients' expression measurements on 6,014 genes and histopathologic features such as Gleason score and tumor volume. The patients were categorized into lethal and indolent prostate cancer. We selected 145 patients with lethal cancer to create a homogenous cohort based on the phenotype. We downloaded the expression data file as well as histopathologic features from Gene Expression Omnibus with accession ID GSE16560 (Edgar et al., 2002).

4.3 C2 curated gene sets

In order to perform our GSA method, we need a list of pre-defined gene sets. We downloaded the C2 catalog, an extensive collection of metabolic and signaling pathways and gene sets from the Molecular Signature Database of Broad Institute of MIT and Harvard (http://www.broadinstitute.org/gsea/msigdb). The C2 catalog consist of 1892 gene sets (accessed on June 2011) collected from online pathway databases, gene sets from biomedical literature including 786 PubMed publications, gene sets compiled from published mammalian microarray studies, and knowledge of domain experts. Sources of the gene sets are provided with gene set files in the C2 catalog (Liberzon et al., 2011).

We screened the C2 catalog for associations with tumor volume which has been found to be associated with development of prostate cancer. We restricted the size of the gene sets in the C2 catalog between 15 and 500 following Subramanian et al. (2005). There were 1263 gene sets within this range. In the C2 catalog, rows represent gene sets containing a pre-defined number of genes and columns represent genes. Table 4.1 shows an example of theC2 catalog. We created a 6013x1263 matrix with 0/1 entries based on the gene expression data and the C2 catalog.

Table 4.1 An example of C2 curated gene set

Gene sets	Genes																
TCAPOPTOSISPATHWAY	TNFSF6	CD3G	CD3D	CCR5	CD3E	CD4	TNFRSF6	TRA	CD3Z	TRB	CD28						
BIOSYNTHESIS_OF_STEROIDS	MVD	HMGCR	FDPS LOC	FDPS	LSS	PMVK	FDFT1	SQLE	DHCR7	VKORC1	MVK	IDI1	NQO1	SC5DL	NQO2		
PMLPATHWAY	TNFSF6	HRAS	TNF	SP100	CREBBP	PML	TP53	PRAM-1	UBL1	PAX3	RB1	TNFRSF6	DAXX	SIRT1	TNFRSF1A	TNFRSF1B	RARA
CHEOK_MP_DN	GSS	TRA TRDV	SOCS6	SERPING1	RBBP8												
ALTERNATIVEPATHWAY	BF	C8A	C7	C9	PFC	C3	C6	C5	DF								

4.4 Results

In this section, we first report results of individual gene analysis against the continuous phenotype, tumor volume, as an initial step to identify differentially expressed genes, and results of gene set analysis obtained by the LCT analysis. Then, we describe the gene set reduction process and the core genes. We performed a logarithmic transformation on the gene expression values to get closer to a normal distribution across individuals.

4.4.1 Results from SAM analysis

Initially, we performed individual gene analysis SAM as an explanatory step before running LCT analysis. There are 346 genes among 6013 total genes with p-values ranging from 0 to 0.05 that are associated with tumor volume. Figure 4.2 illustrates the histogram of p-values from the SAM analysis of six thousand and thirteen genes. Y axis represents frequency of genes and X axis represents SAM p-values or FDR.

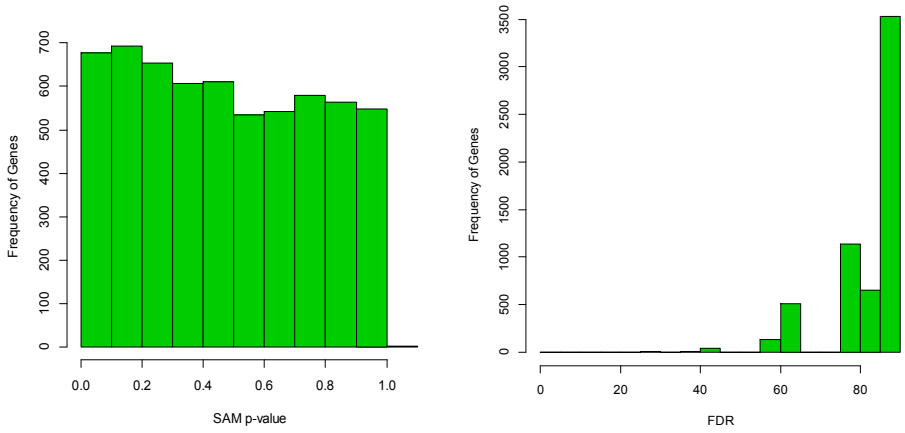


Figure 4.2 Histogram of SAM p-value and false discovery rate

4.4.2 Results from LCT analysis

We applied LCT analysis to a microarray dataset from Swedish Watchful cohort database using the generated 0/1 matrix as an input and the tumor volume as a continuous phenotype. There were 145 patients with lethal prostate cancer. LCT analysis revealed 17 gene sets among 1263 in the C2 catalog that are significantly associated with tumor volume at a cut-off p-value of 0.01 (FDR value of 0.35). Figure 4.3 illustrates the histogram of p-values from the LCT analysis of one thousand two hundred sixty three gene sets. Y axis represents frequency of gene sets and X axis represents LCT p-values. The list of gene sets associated with the tumor volume is described in Table 4.2.

Table 4.2 Gene sets associated with tumor volume phenotype based on the LCT analysis

Gene set name	Gene set Size	p-value
CARBON_FIXATION	16	0.001
HSA00710_CARBON_FIXATION	16	0.002
INNEREAR_UP	19	0.002
XPB_TTD-CS_UP	19	0.003
GALE_FLT3ANDAPL_UP	26	0.005
METASTASIS_ADENOCARC_DN	32	0.005
BCNU_GLIOMA_MGMT_48HRS_DN	123	0.005
UVC_HIGH_D5_DN	23	0.006
GH_EXOGENOUS_ALL_UP	22	0.007
NGUYEN_KERATO_UP	23	0.007
ALKPATHWAY	27	0.007
FALT_BCLL_DN	36	0.007
ET743_SARCOMA_72HRS_UP	49	0.007
AGED_RHESUS_DN	101	0.007
FALT_BCLL_UP	33	0.008
ZHAN_MMPC_EARLYVS	45	0.008
ELECTRON_TRANSPORTER_ACTIVITY	89	0.008

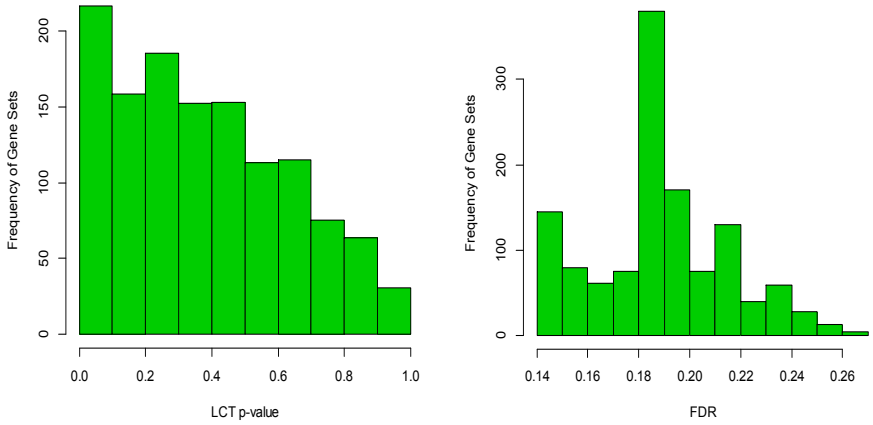


Figure 4.3 Histogram of LCT p-value and false discovery rate

4.4.3 LCT gene set reduction for continuous phenotype

The next step is to use the list of significant gene sets and perform gene set reduction. Given a significant gene set, we used the SAM statistic as a measure of association between each gene within the gene set and the tumor size. SAM is a popular analytical tool for DNA microarray data analysis at individual gene level. We presented a histogram of SAM p-values in Figure 4.2.

For reducing the significant gene set, we ranked the absolute values of the SAM statistic in a decreasing order for genes within the gene set to gradually discover the core genes associated with the tumor size. We used the SAM-R package available in R to compute the SAM statistic values. We can get both FDR values and p-values from the SAM output. However, the FDR values and p-values can be similar for most of the genes and the ranking process of genes based on their significance would be a problem. We prefer to use the SAM statistic values d which are the scores assigned to each gene on the basis of change in gene expression relative to the standard error.

We demonstrate the gene set reduction method for the significant gene sets Carbon Fixation pathway composed of 16 genes as defined in the C2 catalog. We rank the absolute value of SAM statistic for these 16 genes. First, we select the gene with the largest absolute value, ME3 with $|d_{(1)}| = 3.04$ to form the core subset and the rest of the genes within the gene set form the complement set. We apply the LCT analysis to the complement set and evaluate the LCT p-value whether it reaches the pre-specified cut-off value of 0.1. Since the p-value is smaller than 0.1, we select the gene with the second largest absolute value of SAM statistic, i.e., TKT with $|d_{(2)}| = 2.10$. We sequentially add the gene to the core subset and test the complement set until we reach the cut-off threshold. The p-value of the complement set is greater than 0.1 after taking out the third gene PKM2 with $|d_{(3)}| = 1.69$. Genes within the complement set, collectively are not associated with the phenotype and represents the redundant set. Therefore, the core subset contains three genes ME3, TKT and PKM2. Figure 4.4 shows each step of the linear combination test gene set reduction.

Table 4.3 shows the summary of the LCT-GSR including the list of gene sets along with the gene set size, core set size, percent reduction and the core pathway members. Core set size indicates the number of core genes obtained from each significant gene set applying LCT-GSR method. Percent reduction is computed by number of genes eliminated (in the complement set) divided by the total number of genes in a set multiplied by 100. Core pathway shows the core genes collectively contributing to the association with tumor volume excluding the redundant genes from the significant gene sets.

On average, we were able to reduce the number of genes in the 17 gene sets by 90% using the threshold value of 0.1. We observed a situation where a whole gene set is reduced to a single gene. That suggests the genes within the complement subset are not associated with the

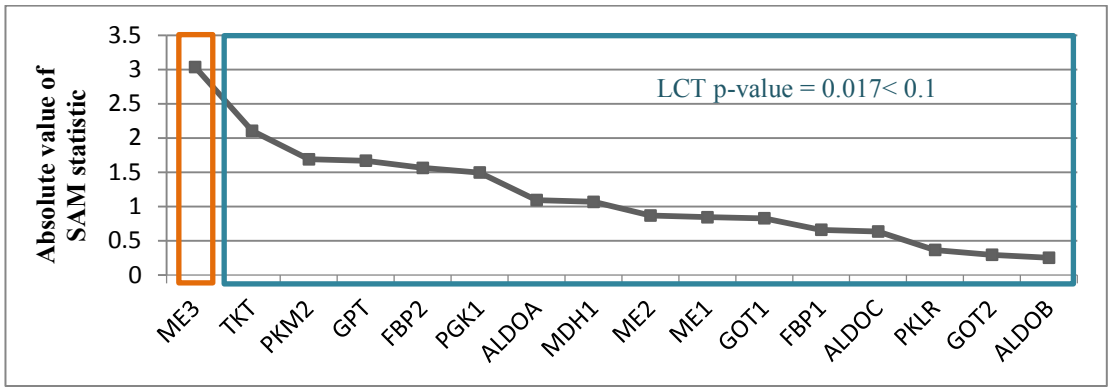
phenotype. If the significance of a set is due to only one gene, the set should be investigated with caution. Biological functional role of the significant gene within the gene set may be considered.

There are 47 core genes obtained from the LCT-GSR method. We report the statistic values of the ten most frequent core genes, their p-values and FDR values from the SAM analysis. The core gene *Malic Enzyme 3* (ME3) is the most frequent gene appearing in the reduced subset of three significant gene sets. The genes *Axis Inhibition Protein* (AXIN1), *Insulin-Like Growth Factor Binding Protein 6* (IGFBP6), *Arachidonate 15-Lipoxygenase, Type B* (ALOX15B), *Upstream Binding Transcription Factor* (UBTF), *High Mobility Group Nucleosomal Binding Domain 4* (HMGN4), *Pyruvate Kinase* Muscle (PKM2), *Cell Division Cycle 16* (CDC16), and *Transketolase* (TKT) appeared two times. The rest of thirty eight core genes appeared once in the

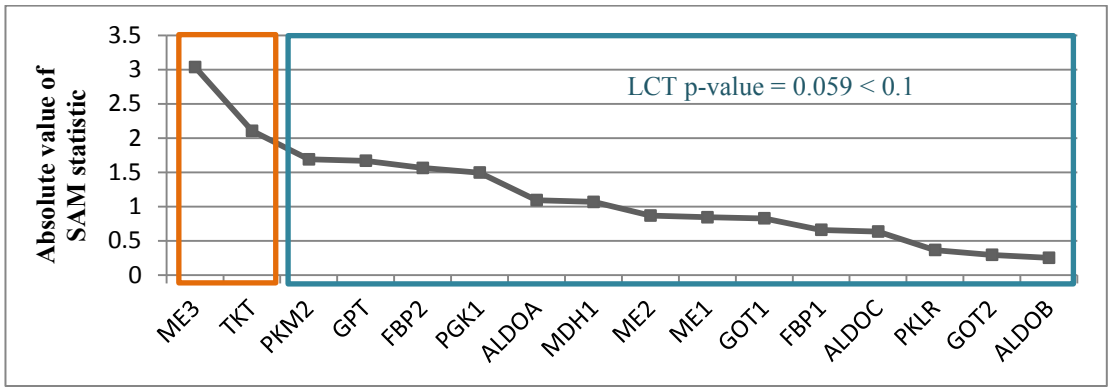
We can observe that some core genes are not statistically significant or partially significant at individual gene level analysis. However, together working with other genes they contribute to the significance of the gene set.

4.4.4 Biological interpretation of findings

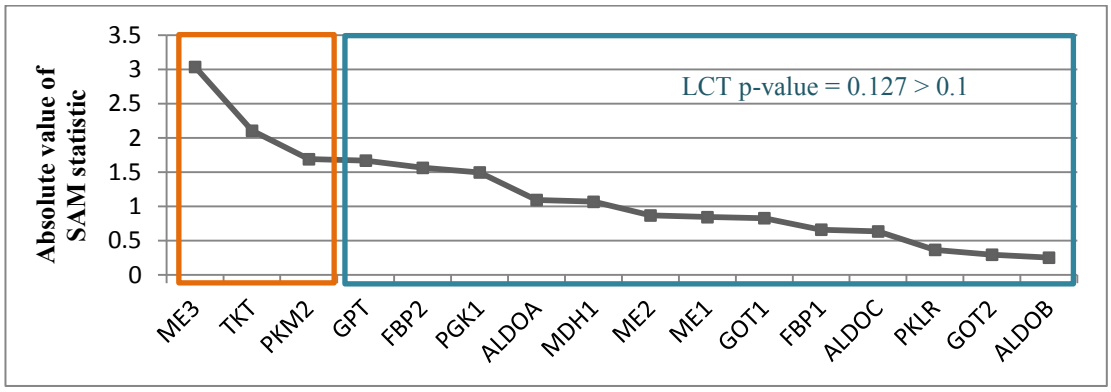
Biological interpretation of statistically significant genes is an essential step in the gene set analysis. It can help researchers to understand underlying mechanism of the disease or trait. Our method identified pathways and genes that were previously discovered to be associated with the tumor volume as well as new markers that need to be further validated. *Malic Enzyme 3*, a gene known to have an important role in cancer cell proliferation (Zheng FJ, et al., 2012), appears most frequently in the three core subsets. Some well-characterized regulators of tumor volume showing up in the core subsets include: *Insulin-Like Growth Factor Binding Protein 6* (Koiko et al., 2005)*, Cell Division Cycle 16, Axis Inhibition Protein, Transketolase* and *Pyruvate Kinase Muscle* (The Human Protein Atlas).



(a)



(b)



(c)

Figure 4.4 An example of linear combination test gene set reduction. We used CARBON FIXATION gene set, identified to be significant by LCT. Each plot shows the absolute value of SAM statistic for genes within this gene set in a decreasing order. In this example we required three consecutive iterations of the gene set reduction method
significant gene sets.

Table 4.3 Extracting core subsets for tumor volume

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
ELECTRON_TRANSPORTER_ACTIVITY	89	3	96.6	TSTA3, ME3, ALOX15B
CASPASEPATHWAY	19	1	94.7	BIRC2
GNATENKO_PLATELET	30	2	93.3	RGS10, SPARC
CARBON_FIXATION	16	3	81.3	ME3, TKT, PKM2
ZHAN_MMPC_EARLYVS	45	3	93.3	SPIB,SNRPC, SLC7A6
GNATENKO_PLATELET_UP	30	2	93.3	RGS10, SPARC
FALT_BCLL_DN	36	6	83.3	HEBP2,IFI6,HMGN4,SERP1,NPC2, PUM1
TPA_RESIST_EARLY_DN	65	3	95.4	ME3,POMZP3, DPP6
METASTASIS_ADENOCARC_DN	32	2	93.8	DLG3, RNASE1
AGED_RHESUS_DN	101	8	92.1	AXIN1, UBE2D2, DPP4, HMGN4, CDC16, RARRES2, JARID1C, SPARC
UVC_HIGH_D5_DN	23	3	87.0	SFRS3, DYRK1A, UBTF
XPB_TTD-CS_UP	19	2	89.5	PRKCZ, PTN
INNEREAR_UP	19	3	84.2	IGFBP6,RPS5, VAMP5
BCNU_GLIOMA_MGMT_48HRS_DN	123	5	95.9	ALOX15B,CRABP1,EPHX1,KIF5A, GP1BB
GH_EXOGENOUS_ALL_UP	22	2	90.9	NOS1, POU2F2
HSA00710_CARBON_FIXATION	16	3	81.3	ME3,TKT, PKM2

Table 4.4 Frequency of the genes within core pathway with SAM p-values and FDR

Gene name	Frequency	SAM p-value	SAM FDR
ME3	3	0.00	0.42
AXIN1	2	0.00	0.25
IGFBP6	2	0.00	0.42
ALOX15B	2	0.01	0.87
UBTF	2	0.02	0.60
HMGN4	2	0.02	0.60
TKT	2	0.02	0.60
CDC16	2	0.03	0.60
PKM2	2	0.06	0.88

Chapter 5

Birth Weight: data description & results

5.1 Background

Low birth weight (LBW) is defined as birth weight of less than 2,500 grams (5 pounds 8 ounces) regardless of gestational age. A baby may be born too early before 37 weeks of pregnancy (preterm birth) or unable to grow enough before delivery (small for gestational age) leading to LBW. Babies with LBW are more likely to have health and developmental problems including learning difficulties, hearing and visual impairments, chronic respiratory problems such as asthma and chronic diseases later in life (Cole et al., 2002).

As adults, individuals born small for gestational age (SGA) are at greater risk of multiple chronic illnesses (Gillman et al., 2007). The link between low birth weight and adult illness might be explained by uteroplacental insufficiency that alters organ function and hormonal milieu to make the individual more susceptible to disease (Barker, 1998). In addition, genetic or epigenetic factors may exist that both reduce fetal growth and increase predisposition to disease later in life (Basso et al., 2006).

It is now widely recognized that methylation of cytosine in CpG dinucleotides is a mechanism for downregulating gene expression for at least a third of human genes, and there is substantial variation in methylation among individuals and tissues (Eckhardt et al., 2006; Rakyan et al., 2004; Song et al., 2005). The change in DNA methylation is known as the cause of some newborn illnesses and growth disorders. While DNA methylation is important in developmental processes, and its variation in blood lymphocytes has been associated with adult body mass

index (BMI) (Feinberg et al., 2010), analysis of DNA methylation patterns with respect to birth weight have produced mixed results.

DNA Methylation ultimately exerts its biological consequences via its regulatory effects on mRNA production and resultant protein production, both of which are complex processes. Therefore, variation in gene expression levels is one step closer to a direct biological effect than DNA methylation and might exhibit a stronger association with birth weight variation (Adkins et al, 2012). For instance, in candidate gene studies significant associations with birth weight have been published for placental expression levels of 11b-HSD1, 11b-HSD2, DLX4, LEP, PHLDA2, FTO, IGF-I, IGFBP-1, MEST, MEG3, GATM, GNAS, PLAGL1, and the growth hormone like cluster of genes (Apostolidou et al., 2007; Bassols et al., 2010; Koutsaki et al., 2011; Männik et al., 2010; McMinn et al., 2006; McTernan et al., 2001; Mericq et al., 2009; Murthi et al., 2006; Sheikh et al., 2001; Struwe et al., 2007; Tzschoppe et al., 2009; Struwe et al., 2009). Many more significant associations between birth weight and gene expression have been published over the last decade relative to DNA methylation suggesting the need for further investigation at gene expression level.

In a recent study of 201 newborns ranging in birth weight from 2.1 to 5 kg, Adkins et al. (2012) did not identify strong genome-wide association of birth weight with gene expression. The analysis in this study was focused on identifying individual genes that are associated with birth weight among a set of clinically normal newborns. We reason that correlation among genes especially those within biological pathways might impact the association with birth weight. Therefore, in this real microarray study we investigated the association between a priori defined sets of genes and the continuous phenotype birth weight using the LCT-GSR method. The

ultimate aim of this investigation is to identify biomarkers that contribute to variation in birth weight.

5.2 Data Description

The birth weight data set is part of a larger longitudinal cohort study of human development from pregnancy to age 3, the Conditions Affecting Neurocognitive Development and Learning in Early Childhood (CANDLE). CANDLE was performed in Shelby County, Tennessee. Written informed consents were obtained from all mothers, and this study was approved by the institutional review boards of all the participating hospitals (Adkins et al., 2012). Data on maternal age, gestational age, race, and baby's gender are also available. We obtained approval from the University of Tennessee Health Science Center for accessing data on continuous phenotype birth weight measured on newborn blood.

The selection criteria for the cohort were: maternal age 18–40 years, singleton pregnancy, complete data on birth weight and maternal prepregnancy weight, and absence of several complications, specifically sexually transmitted disease, diabetes, oligohydramnios, preeclampsia, placental abruption, tocolytics, and cervical cerclage. We selected gestational ages of 35–42 weeks and mother whose self-declared race was only Caucasian or only African-American. After applying these additional criteria, the final sample size was 114. This data set consists of 24,924 gene expression measurements from blood sample for 114 newborns, 67 African-American and 47 Caucasion, with mean birth weight of 3340 (SD: 490) grams. The mothers mean age is 27 years old and the mean gestational age is 39 weeks. Table 5.1 shows the characteristics of the participants in the selected cohort.

Table 5.1 Characteristics of the participants (n=114)

Variable	Mean (SD)	Range
Race African-American [n] Caucasian[n]	 47 67	
Female [n]	50	
Gestational age [weeks]	39 (1.2)	35-42
Mothers' age [years]	27 (5.1)	18-39
Birth weight [gr]	3340 (490)	1931-4954

Rates of low birth weight vary among women of different origins. It has been long observed that the rate of low birth weight among African-American mothers is twice that of Caucasian women (Collins et al., 2004). On the other hand, birth weight has consistently been shown to be higher in males than in females (Van Vliet, 2009). Table 5.2 suggests lower birth weight for the African-American mothers and the difference is statistically significant (t=-4.2, p-value=0.0001). There is no significant difference between birth weight of male and female newborns though (t=0.09, p-value=0.927). We examined whether the effect of race on birth weight is modified by gender and the interaction was not significant (t=1.01, p-value=0.314). Since gender and race are important characteristics influencing the birth weight we adjust for both variables in the model.

Table 5.2 Birth weight of the participants by race and gender

Race	Gender	Frequency	Birth wight (SD)	Total birth weight (SD)
Caucasian	Male	24	3615.2 (542)	3553.7 (470)
	Female	23	3489.5 (382)	
African-American	Male	40	3168.9 (485)	3190.1 (449)
	Female	27	3221.4 (395)	

5.3 Pre-defined gene sets

C7 immunologic signatures gene sets

We need a list of pre-defined gene sets to perform our method. We downloaded the most recent list of gene sets in the Molecular Signature Database C7 catalog (accessed on May 2015) from Broad Institute (http://www.broadinstitute.org/gsea/msigdb). The C7 catalog contains 1910 gene sets representing immunologic signatures collected from immunologic studies.

Stem cell signatures

We also used another source of pre-defined gene sets, stem cell signatures. This list contains 457 gene sets collected from manuscripts (Leite & Pyne, manuscript in preparation) and others from the Differentiation Map portal (Novershtern et al., 2011), Ingenuity Pathway Analysis tool (http://www.ingenuity.com/), and ChIP-X database (Lachmann et al., 2010).

In these lists, row represent gene sets containing a priori defined genes and columns represent entrez gene IDs. We restricted the size of gene sets in both lists to be between 15 and 500. There are 251 gene sets within this range for the stem cell signatures and 1910 gene sets for the C7 catalog.

5.4 Results

We applied LCT-GSR to the gene expression data set from CANDLE Study adjusting for race and gender. We first report results of individual gene analysis for the continuous phenotype birth weight and results of the LCT analysis. Then we describe the gene set reduction process and the core genes. We performed a logarithmic transformation on the gene expression values to increase the normality of the distribution across individuals.

5.4.1 Results from SAM analysis

Initially, we performed SAM as an exploratory step before running LCT analysis. Figure 5.1 illustrates the histogram of 24,924 p-values from the SAM analysis. Y axis represents frequency of genes and X axis represents SAM p-values or FDR. There are 1,675 significant genes among 24,924 total genes with p-value smaller than 0.01 that are associated with birth weight.

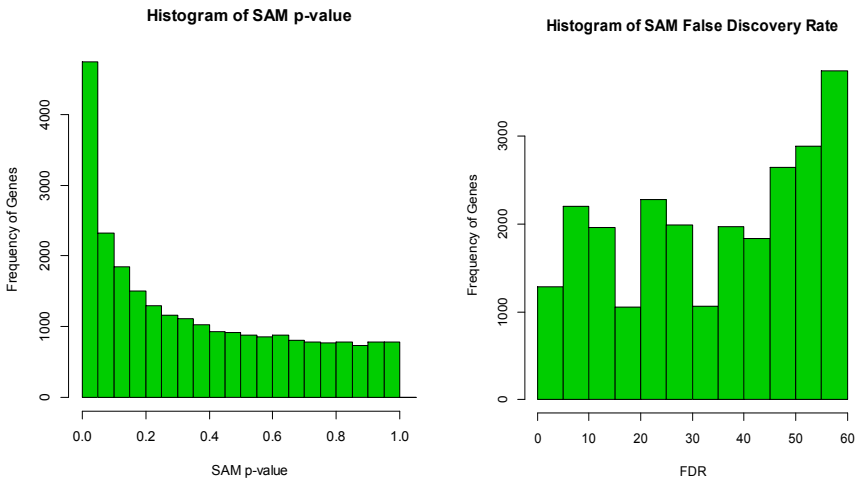


Figure 5.1 Histogram of SAM p-values and false discovery rate

5.4.2 Results from LCT analysis

We applied LCT analysis to a microarray dataset from CANDLE study using the generated 0/1 matrix as an input and the birth weight as a continuous phenotype. The LCT analysis revealed 33 gene sets in the stem cell signatures (FDR<0.003) and 210 gene sets in the C7 catalog (FDR<0.004) that are associated with birth weight at a cut-off p-value of 0.01.

Figures 5.2(a) and (b) illustrate the histogram of p-values from the LCT analysis and distribution of FDR values for the stem cell signatures and C7 catalog, respectively. Y axis

represents frequency of gene sets and X axis represents LCT p-values or FDR values. The list of gene sets associated with birth weight is described in Table A and B in the appendix.

5.4.3 LCT gene set reduction for continuous phenotype

The next step is to use the list of significant gene sets and perform gene set reduction. Given a significant gene set, we used the SAM statistic as a measure of association between each gene within the gene set and the birth weight.

For reducing the significant gene set, we rank the absolute values of the SAM statistic in a decreasing order for genes within the gene set to gradually discover the core genes associated with the birth weight. We apply the LCT analysis to the complement set and evaluate the LCT p-value whether it reaches the pre-specified cut-off value of 0.1. When we reach the complement LCT p-value threshold we ensure that the genes within the complement set collectively are not associated with the birth weight. The remaining genes form the core subset. We discussed the LCT-GSR in details in Chapter 4. Here we report the summary of the results for each category.

Table 5.5 shows the summary of the LCT-GSR for stem cell signatures including the list of gene sets along with the gene set size, core set size, percent reduction and the core pathway members. Core set size indicates the number of core genes obtained from each significant gene set applying LCT-GSR method. Percent reduction is computed by number of genes eliminated (in the complement set) divided by the total number of genes in a set multiplied by 100. Core pathway member shows the core genes collectively contributing to the association with birth weight excluding the redundant genes from the significant gene sets.

There are 33 significant gene sets within stem cell signatures (p-value<0.01) associated with variation in birth weight after adjusting for the race and gender. There are 228 genes identified to

be significantly associated with variation in birth weight from these gene sets after adjusting for the race and gender variables. On average, we were able to reduce the number of genes in the 33 significant gene sets of stem cell signatures by 84.3% using the cut-off value of 0.1.

Table 5.6 shows the summary of the LCT-GSR for C7 catalog. There are 210 significant gene sets within C7 catalog (p-value<0.01) associated with variation in birth weight after adjusting for the race and gender. There are 1604 genes identified to be significantly associated with variation in birth weight from these gene sets after adjusting for the race and gender variables. On average, we were able to reduce the number of genes in the 210 significant gene sets of C7 catalog by 89% using the cut-off value of 0.1.

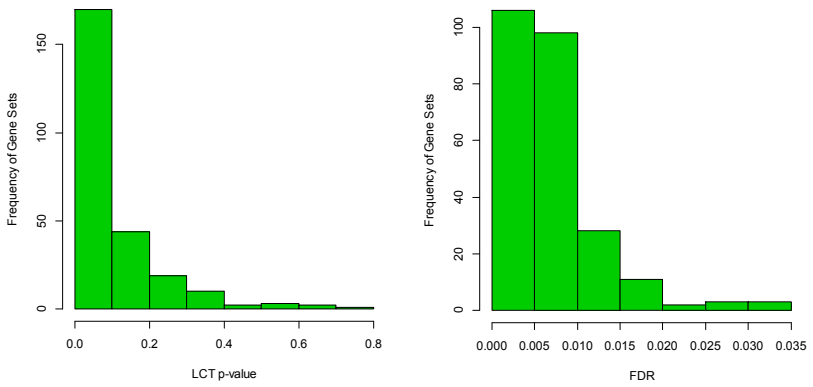
Table C and D in the appendix illustrates the core genes obtained from the LCT-GSR method with frequency greater than two and the corresponding p-values and FDR values form the SAM analysis. In the stem cell signature, the core genes *Kruppel-Like Factor 6* (KLF6), *Diazepam Binding Inhibitor* (DBI), *Early Growth Response 3* (EGR3), and *Jun Proto-Oncogene* (JUN) are the most frequent gene appearing in the reduced subset of four significant gene sets. There are total 229 unique genes identified in the reduced subsets.

In the C7 catalog, the core genes *Lectin, Galactoside-Binding, Soluble, 3* (LGALS3) and *G0/G1 Switch 2* (G0S2) are the most frequent gene extracted from 17 significant gene sets. The core gene *Endothelial PAS Domain Protein 1* (EPAS1) appeared in 16 significant gene sets and *Iduronate 2-Sulfatase* (IDS) and *Chemokine (C-X-C Motif) Ligand 8* (CXCL8) appeared in 15 significant gene sets. There are total 1603 unique genes identified in the reduced subsets.

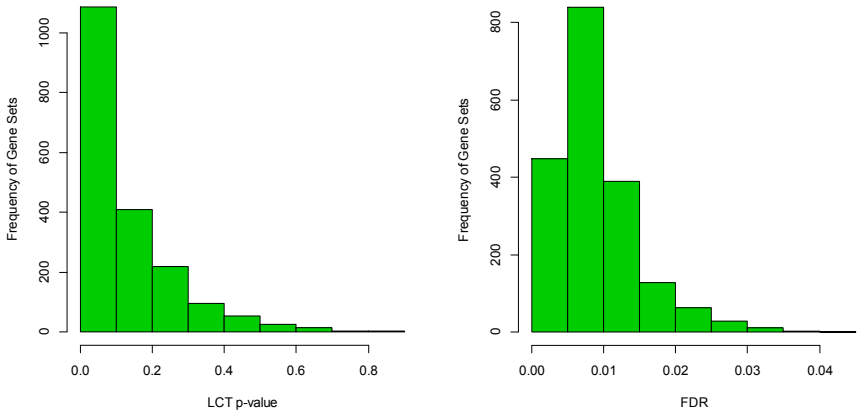
There are 180 unique core genes extracted from the significant gene sets in the stem cell signatures which overlapped with the core genes extracted from the list of significant gene sets in C7 catalog. This shows reproducibility of our method across databases.

5.4.4 Interpretation of findings

There are genes among core pathway members that are not associated with the birth weight at individual level analysis; for example, *N(Alpha)-Acetyltransferase 35* (NAA35) and *GABA(A) Receptor-Associated Protein-Like 2* (GABARAPL2) with the SAM p-value 1.0 and FDR 59.6%, *Heparan Sulfate (Glucosamine) 3-O-Sulfotransferase 3A1* (HS3ST3A1) and *Par-3 Family Cell Polarity Regulator* (PARD3) with the SAM p-value 1.0 and FDR 54.6% in the C7 catalog. However, they contribute to the significant association with birth weight jointly with other genes within the core subset. We can observe that the gene GABARAPL2 was identified in 4 different significant gene sets. Biological interpretation of these genes is essential in understanding potential biomarkers influencing birth weight.



(a)



(b)

Figure 5.2 Histogram of LCT p-values and false discovery rate (a) using stem cell signatures, (b)using C7 catalog

Table 5.3 Extracting core subsets of stem cell signatures associated with birth weight

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
IPA_affects differentiation of embryonic stem cells	41	5	87.8	RNF2, ANGPT1, TLN1, NANOG, SOX2
StemCell_Kasper06_30genes_16880536-table1	30	3	90.0	ULK1, EGR3, IDS
DMAP_MEGA_UP	46	13	71.7	NUDT6, LOC55338, AGT, CALD1, SIX3, POLH, SSX1, TNP2, TFAP2A, PCP4, LAMB4, TBCE, LOC57399
DMAP_MONO1_DN	47	21	55.3	BRD8, GIMAP5, BTG1, ZCCHC6, MAP7D1, MICB, PREP, IQSEC1, ZFP36L2, ACOX1, IRF2, RNASEL, SARS, GEMIN6, HLA-A, DUSP10, KCNJ2, APOL2, TM2D3, SELPLG, TLR1
DMAP_PRE_BCELL2_UP	44	8	81.8	ZNF124, SERPINA5, MFSD6, 654056, PHF20L1, GNG11, ARHGEF17, CSPP1
DMAP_PRE_BCELL3_DN	44	8	81.8	ZNF124, SERPINA5, MFSD6, 654056, PHF20L1, GNG11, ARHGEF17, CSPP1
StemCell_Lim08_50genes_18510698-Table1	47	9	80.9	GP5, PLEK, GABRE, LRP12, SLC44A1, CALD1, SCD, PDE5A, CXCL3
Ben-Porath_MYC_TARGETS_WITH_EBOX	226	24	89.4	APP, BAX, GSTP1, MNX1, EGR3, JUN, MST1, DBI, RHOG, CD79B, SNHG5, CD2, HDAC3, PRTN3, MUC1, HSPA8, HMBS, MPO, HIST1H4E, SERPINE1, TXN, NBN, PPID, BCL3
DB_ESR1-15608294	88	14	84.1	CRCP, SIRT3, SERPINB9, BRCA1, TRIP10, BRIP1, SERPINE1, ZNF600, ENSA, CASP8AP2, AGT, LTF, DCC, PGR
StemCell_Kocer08_87genes_18667080-TableS6	71	6	91.5	HSPA1B, CTSB, MCC, ACTR2, BTG1, KIAA0020
StemCell_Shim04_25genes_15246160-table6	22	3	86.4	KLF6, JUN, IDS
StemCell_Fruehauf06_110genes_16863911-table1	97	9	90.7	CLK4, HSPA1B, MS4A3, RNASE3, EGR3, HIST1H2BK, RNASE2, MPO, ELL2
DMAP_ERY_UP	45	9	80.0	XK, TRAK2, ARHGEF12, RHCE, TMCC2, GYPE, ACSL6, ANK1, HBBP1
DMAP_GM_EARLY_DN	42	10	76.2	DMP1, KCNH6, NAG18, ASCC2, EPB41L4A, LOC55338, SIX3, POLH, SEMA3C, SSX1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
DMAP_PRE_BCELL_UP	39	7	82.1	LOC55338, CLDN14, POLR3G, POLH, DPYS, TFAP2A, LAMB4
DMAP_BCELL_DN	44	5	88.6	ACTN1, DMP1, NUCB2, BLZF1, ASCC2
DMAP_TCELLA6_DN	45	7	84.4	KLF6, CD58, DBI, FAS, SYT11, YWHAQ, AUTS2
StemCell_Tondreau08_52genes_18405367-Table2b	41	6	85.4	IGFBP7, MFAP5, COL8A2, HAS3, CALD1, PAWR
DMAP_BCELLA2_UP	49	6	87.8	CTSB, EGR3, CD1C, GIMAP5, DSE, IDH3A
DMAP_TCELLA6_UP	44	5	88.6	FKTN, GP5, CEPT1, IGF1R, NET1
IPA_affects differentiation of stem cells	72	5	93.1	RNF2, ANGPT1, GATA2, TLN1, NANOG
DMAP_ERY4_DN	47	5	89.4	HLA-DPB1, LILRA6, ACSM5, HLA-DMA, C4BPA
IPA_decreases differentiation of stem cells	18	5	72.2	JUN, DKK1, LIF, IL6ST, NEUROG1
StemCell_Colombo09_111genes_19123479-TableS1	92	8	91.3	OTUD1, HBP1, MGAT1, MTMR3, CHIC2, MIS12, TRIB1, FIP1L1
StemCell_Lim08_25genes_18510698-Table2	25	3	88.0	MS4A3, JUN, ALOX5
DMAP_ERY_DN	46	13	71.7	EIF4B, ACSL5, GMFG, PDCD4, DBI, TES, RPL39, RPS3A, ZFP36L2, TRIM44, SMAD3, DICER1, RPL13A
DMAP_GM_EARLY_UP	40	10	75.0	GRHPR, TMEM156, EHD4, CR2, DYRK4, MRPS18B, GTF2H5, QTRTD1, BET1, SHMT2
DMAP_HSC1_DN	48	6	87.5	PLEK, ARAP3, TIMP3, PRKAR2B, DNAJA1, DNAJC6
DMAP_HSC3_UP	48	6	87.5	PLEK, ARAP3, TIMP3, PRKAR2B, DNAJA1, DNAJC6
DB_PPARG-19300518	194	17	91.2	MCM2, ATP1A2, NDUFV1, SMARCA4, DBI, CHIC2, G0S2, SDHC, LEP, COX15, RCL1, PDZRN3, FGF10, S100A8, UBE2I, ALDH3A1, ACADVL

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
StemCell_Bhattacharya05_2843genes_16207381-Table1Sa	312	24	92.3	IDI1, MCM2, WDR18, SOAT1, KLF6, YIPF1, FAR2, KIAA0020, ZCCHC6, GGCT, CD79B, TCEB3, GYG2, MAP4K4, MSMO1, CHMP2B, MTHFD2, HEATR5B, SNRPA, PICALM, THRAP3 STON1-GTF2A1L, JARID2, PREP
DMAP_MONO2_DN	40	7	82.5	ATP5J2, PRIM2, ZNF43, CUL7, TCIRG1, MYO15B, NDUFS6
DMAP_TCELLA2_DN	47	2	95.7	KLF6, CD58

Table 5.4 Extracting core subsets of C7 catalog associated with birth weight

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
KAECH_NAIVE_VS_DAY8_EFF_CD8_TCELL_UP	198	15	92.4	CLK4, EML5, APP, FRMD8, EVI5, WDR74, KMT2A, RPL5, BCKDHB, EGR2, SLC44A1, HMP19, LIPA, IGF1R, CD72
KAECH_NAIVE_VS_DAY8_EFF_CD8_TCELL_DN	194	25	87.1	GABARAPL2, F2R, CAPNS1, KRTCAP2, SERPINB9, DBI, ATP5J2, CAPZB, MAP7D1, RSU1, ITGA4, LGALS3, DHRS1, CASP1, CTLA4, TXN, E2F8, GLRX, SEC61G, EFHD2, DLGAP5, ABRACL, GZMA, TACC3, SH2D1A
KAECH_DAY15_EFF_VS_MEMORY_CD8_TCELL_UP	192	22	88.5	PHF13, ARPP19, GMFG, KRTCAP2, MBD4, KMT2A, CD79B, GJA3, RSU1, CCR6, RHD, LGALS3, SORBS1, S100A8, EGR2, JARID2, IGF1R, IL1B, FCGR2B, MTM1, ALDH2, GZMA
GOLDRATH_EFF_VS_MEMORY_CD8_TCELL_UP	197	32	83.8	IDI1, CDC6, SERPINB9, DBI, KIAA0101, CKS2, SMC2, BRCA1, H1F0, RHD, LGALS3, DHRS1, FPR2, S100A8, SYPL1, FDFT1, TXN, TSPAN32, E2F8, MRPL18, TMEM14C, BUB1, MTM1, DEGS1, DLGAP5, EGR1, GZMA, RAD51, TACC3, CKS1B, TMPO, CHAF1A
GSE10094_LCMV_VS_LISTERIA_IND_EFF_CD4_TCELL_UP	196	35	82.1	FKTN, CREB1, RSRP1, CEPT1, FEZ2, BDP1, NDUFB3, ZNF318, HINFP, ARHGAP30, RPL5, ANKRD44, NUFIP2, FAM134C, MORC3, ZNF623, MED20, STK38, PEAK1, EPM2AIP1, SAMHD1, SPN, TOLLIP, FAM69A, BCDIN3D, POLK, C2orf68, ZDHHC7, VAMP4, USP47, ACSS1, APPL1, ACSF3, CGGBP1, UIMC1
GSE10239_NAIVE_VS_MEMORY_CD8_TCELL_UP	199	12	94.0	NUBPL, IDI1, RNF19A, TXNDC15, HNRNPK, CEPT1, ACSL3, TMEM131, USP1, MPP6, DIRC2, AMPD1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE10239_NAIVE_VS_KLRG1INT_EFF_CD8_TCELL_DN	197	13	93.4	ZNHIT3, SCN8A, BANF1, SERPINB9, TRIM37, RPS27L, HMBS, MPZL3, LGALS3, EHD4, SLAMF7, MPP6, PDSS2
GSE10239_NAIVE_VS_KLRG1HIGH_EFF_CD8_TCELL_UP	195	11	94.4	RSRP1, IDS, BTG1, UBLCP1, RPL11, RPL35, VEGFB, FAS, DIRC2, AMPD1, THYN1
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_BCELL_UP	195	25	87.2	ACTN1, GMFG, STAT4, RARRES3, ZMYM6, GIMAP5, MAP7D1, CD2, KDSR, ADTRP, CTLA4, LDLRAP1, USP20, MEOX1, WWP1, SIRPG, ATP2B4, FHIT, TIMP1, SKAP1, CLUAP1, FAM134B, PRKCQ, TESPA1, IL7R
GSE10325_CD4_TCELL_VS_LUPUS_CD4_TCELL_UP	189	14	92.6	ZNF212, DHPS, TACSTD2, BTG1, RPL11, VEGFB, CCT8L2, SYPL1, PFDN5, NET1, EIF3H, PIGQ, ARF4, TSFM
GSE10325_CD4_TCELL_VS_LUPUS_CD4_TCELL_DN	198	29	85.4	C18orf25, BMP8B, HIST1H2BK, CEP97, HERC2P3, FHL2, SNAPC3, ATP5J2, FAR2, RNASE2, SP110, FAS, CASP1, HDGFRP3, SPATS2L, LGALS3BP, MX1, MRPL42, SYT11, B3GNT2, IFITM1, EIF2AK2, CLUAP1, IL7, ADGRG6, MT1H, DDC, ICA1, SLC50A1
GSE11057_NAIVE_CD4_VS_PBMC_CD4_TCELL_DN	189	13	93.1	EPAS1, AHR, LILRB3, JAK2, PDLIM1, GSTP1, GNLY, FEZ2, CAPNS1, NAGK, GSR, LILRA6, PEA15
GSE11057_PBMC_VS_MEM_CD4_TCELL_UP	189	16	91.5	IL18, LILRB1, LILRB3, IGFBP7, MS4A3, CSF2RA, CD1C, NAGK, RNASE2, LILRA6, FPR2, ITGA2B, SPI1, SLC46A2, GNG11, SLC15A3, SLC15A3
GSE11864_UNTREATED_VS_CSF1_IN_MAC_UP	191	24	87.4	GMFG, HNRNPK, C4orf33, HIPK2, RARRES3, PSMA1, C6orf48, CCL17, RPL35, RELA, ENSA, WIBG, SPI1, NET1, KLF13, KLHDC4, RYK, PIWIL4, HERPUD1, C7orf62, TAP2, LOC399900, RASGRP4, TMC6
GSE11864_UNTREATED_VS_CSF1_PAM3CYS_IN_MAC_DN	185	19	89.7	SLC44A4, SPATA9, KLHL23, GRWD1, APTX, LEO1, MTHFD2L, PSMA1, MUCL1, PLAGL2, EIF5B, CPSF2, CKAP2L, KLHL28, SEC61G, SPATS2L, IMMT, PHF23, MCCC2
GSE11864_CSF1_IFNG_VS_CSF1_IFNG_PAM3CYS_IN_MAC_DN	184	17	90.8	DOCK3, TNFSF14, ACSL5, ATP6V1C1, USP49, CD2, TP53BP2, FBXO46, RAB40AL, R3HCC1L, NCAM1, SLAMF7, TBC1D7, ZFAND2B, TRIB1, ETF1, SPSB1
GSE12845_IGD_POS_BLOOD_VS_PRE_GC_TONSIL_BCELL_DN	199	23	88.4	SHCBP1, SPC25, YWHAE, BANF1, CDC6, SMARCA4, ESPL1, GRHPR, SMC2, CBX5, HSPA8, EHD4, TERF2, MTHFD2, LGMN, FAM120A, RFX7, ARPC2, SNTB2, ENO2, DLGAP5, TACC3, TSFM

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE12845_IGD_NEG_BLOOD_VS_NAIVE_TONSIL_BCELL_UP	195	22	88.7	G6PC3, SSR1, STOML2, ZNF706, TM9SF1, NDUFB3, G0S2, NKG7, ACBD3, SUMO1, ALG8, HTATIP2, FAS, MCUR1, SNRPA, NDUFA5, TXN, GLRX, NDUFA4, PSMB7, DYRK4, DNAJC3
GSE12845_NAIVE_VS_PRE_GC_TONSIL_BCELL_DN	197	33	83.2	NUBPL, CASP2, SPC25, STOML2, NDUFV1, CDC6, ACTR2, SDHC, DHRS7B, OSBPL9, HTATSF1, TERF2, DENR, P2RX3, NDUFA4, C1orf112, ATP5H, TRIB1, NDUFS6, ENO2, ALDH2, ST14, ATP2A2, SAE1, DNAJC7, GPR137B, CHCHD2, MSH6, CSTB, GTF2A2, SMPDL3A, CD81, BARD1
GSE13306_TREG_VS_TCONV_SPLEEN_DN	196	14	92.9	AHR, OTUD1, POT1, DMP1, ZDHHC2, NAP1L1, GRIK3, TJP1, ZNF318, CDV3, PSAT1, EDEM3, MBNL2, MFSD6
GSE13411_NAIVE_BCELL_VS_PLASMA_CELL_UP	193	25	87.0	APP, SNX29P2, WDR74, CD1C, ZNF318, ZNF273, GRK5, LAIR1, TP53BP2, HLA-DPB1, FBXL14, IMPACT, ALOX5, CWC25, PFDN5, OSER1, KIF16B, SLC15A3, DCLRE1C, RNASET2, STK17A, FAM192A, RNF187, PIK3CD, AUTS2
GSE13484_UNSTIM_VS_3H_YF17D_VACCINE_STIM_PBMC_DN	193	30	84.5	CXCL8, MMADHC, PITPNB, ESPL1, DEAF1, GRK5, IRF9, ELK4, PDPN, MIS12, IDO1, PCDHA3, R3HCC1L, DSCAM, FAS, SOCS5, BST2, RAD21, NBN, BCL3, LAMTOR3, JARID2, AKR1C1, FCGR2B, FICD, MYH10, PLEC, SIM2, DSE, CRADD
GSE13484_12H_UNSTIM_VS_YF17D_VACCINE_STIM_PBMC_UP	197	24	87.8	EIF4B, FCN1, SPTLC1, HIPK2, BLVRB, ZNF124, PIH1D1, CDV3, KXD1, LSM14A, MBNL2, BLVRA, IMPAD1, FOCAD, HLA-DMA, CLEC5A, CUTA, PABPC3, PCYOX1, INPP4A, EIF2D, SH2D1A, PLBD1, ATP1A1
GSE13484_12H_VS_3H_YF17D_VACCINE_STIM_PBMC_UP	194	14	92.8	LILRB1, CRIPT, RSRP1, ZMIZ2, SELT, ABCA5, MICU1, PMS2P1, CBFA2T2, PLAGL2, LRP12, H1F0, CCL17, LGALS3
GSE13485_CTRL_VS_DAY7_YF17D_VACCINE_PBMC_UP	172	17	90.1	MRPS27, EIF4B, ZNF652, CHAMP1, OLR1, CYP4F3, TMEM121, RAX2, RPL11, GPR75, RPL5, MPZL3, GABRB1, C9orf135, ZNF669, DAPK2, LAMTOR3
GSE13485_DAY1_VS_DAY21_YF17D_VACCINE_PBMC_DN	190	35	81.6	RPA3, PPA2, CD24, POT1, GMFG, CD58, NAA30, GRWD1, NDUFB3, RPP30, RPS27L, GNPNAT1, METTL5, HSPA8, ALG8, PARPBP, EIF2S1, MRPL51, ZCCHC9, HMP19, GLRX, PPID, CLLU1, ZNF189, DEGS1, MED20, TRMT44, CCDC126, CHCHD1, GTF2H5, STYX, VTA1, ECHDC1, RPL38, C1GALT1C1
GSE13738_TCR_VS_BYSTANDER_ACTIVATED_CD4_TCELL_DN	182	26	85.7	AHR, TNFSF14, RAB18, HACD4, IDS, NIPA1, PPP2R2B, UHMK1, FAR2, NFKBIZ, ESPN, GSR, 42071, CCR6, CHDH, TGFBR1, ETV7, FAS, BLVRA, ALOX5, CASP1, GLYR1, TXN, CYSTM1, WWP1, DNAJC3

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE14000_4H_VS_16H_LPS_DC_TRANSLATED_RNA_DN	194	17	91.2	AHR, RNF2, ENTPD7, PHF13, POT1, KLF6, BRD8, BDP1, THAP11, C10orf10, HEXIM2, RBM4B, PSTK, ZNF43, ADTRP, HSD17B1, CYP27B1
GSE14308_TH2_VS_INDUCED_TREG_UP	194	13	93.3	CLK4, DDX6, RPE, IDS, ZC2HC1A, DDX5, VPS37C, SURF1, MAP4K4, IMPACT, NUDT6, ARPC2, FAM127C
GSE14308_INDUCED_VS_NATURAL_TREG_DN	197	26	86.8	HBP1, SNAPC1, SELT, TERF1, BCO2, DCP1B, RNF145, UBXN2A, GBF1, TAF8, DNAJB14, AMY2A, DIRC2, MORC3, OPA1, MTMR10, DNAJC3, CCDC47, PAPOLG, STK38, SLMAP, AP1G2, GZMA, GAB3, ALKBH6, HOOK3
GSE1448_CTRL_VS_ANTI_VBETA5_DP_THYMOCYTE_UP	196	18	90.8	YIPF4, ANGPT1, ZNF644, NAP1L1, CDKN2AIPNL, G0S2, DMRTB1, GATA5, ANTXR2, TAF8, POLR2I, SRP68, HMP19, PREP, HLA-E, MEOX1, ZNF326, DEGS1
GSE1448_ANTI_VALPHA2_VS_VBETA5_DP_THYMOCYTE_UP	196	13	93.4	CLK4, NKIRAS1, CRIPT, SDC4, NR1D2, HIGD2A, PACRG, MCEE, FABP1, MST1, CFHR2, IL10, GRIN2D
GSE1460_INTRATHYMIC_T_PROGENITOR_VS_THYMIC_STROMAL_CELL_UP	197	19	90.4	MGMT, ADCY8, MCM5, SCN8A, APBB3, ZMYND10, MICU1, ABCF1, THAP11, ZNF606, TIMELESS, TUBGCP4, ZXDC, FBXL14, TUBD1, PMS2P5, ANAPC15, C2orf54, SMA4
GSE1460_DP_THYMOCYTE_VS_NAIVE_CD4_TCELL_ADULT_BLOOD_UP	197	29	85.3	RPA3, PDLIM1, SHCBP1, SPC25, EGR3, FEZ2, CEP97, DBI, C3orf52, SMC2, CBX5, BRCA1, EXOG, E2F8, C1orf112, MYH10, BTG3, EGR1, TP53BP1, PPAP2B, HIST1H2AE, EIF4A3, NUP214, COL6A3, ARHGAP32, IDH3A, HIST1H4H, ECHDC1, MYB
GSE1460_DP_THYMOCYTE_VS_THYMIC_STROMAL_CELL_DN	197	16	91.9	OSMR, FEZ2, RGS13, DOLK, GRHPR, KANK2, CASP1, NBN, SNTB2, TIMP3, LGALS3BP, FZD3, PARVB, SSR4, DNAJC6, NDUFA1
GSE15659_NAIVE_VS_PTPRC_NEG_CD4_TCELL_DN	193	20	89.6	YAP1, RNF19A, ZNF644, TOP3A, ZMIZ2, RHBDL2, SCN8A, HACD4, ZNRD1, VEGFB, ST7-AS1, SPATA22, RAB35, SAP30BP, SPATS2L, RAB9A, ZNRF2, TIMM17A, RPE65, SPACA7
GSE15659_NAIVE_CD4_TCELL_VS_ACTIVATED_TREG_DN	195	28	85.6	RNF19A, TOP3A, EMC4, ZMIZ2, RHBDL2, SCN8A, ZNF557, SLC16A11, ZNRD1, ST7-AS1, UBL3, ZC3H13, SAP30BP, ZNF28, SPATS2L, TIMM17A, RPE65, THBS2, SPATA5L1, TP63, TTL, PRAMEF12, SH2D1A, ZFAND5, PRKAG3, ZNF142, RUSC1, SHB
GSE15659_CD45RA_NEG_CD4_TCELL_VS_RESTING_TREG_UP	186	19	89.8	DNAJC28, AP3M1, CTAG2, CCDC33, COMMD7, CEP164, CHPF, FBXL14, RHNO1, FDCSP, ETV7, DNAJB14, C4orf36, AP1M1, BCL3, LOC55338, CR2, DNASE1L2, DPYS

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE15659_CD45RA_NEG_CD4_TCELL_VS_ACTIVATED_TREG_DN	194	21	89.2	PNPT1, TOP3A, EMC4, ZMIZ2, RHBDL2, SCN8A, HACD4, TUBGCP3, ZNRD1, ZCCHC9, ST7-AS1, SPATA22, UBL3, PLIN5, ZC3H13, SAP30BP, TIMM17A, THBS2, TTL, TTTY13, SH2D1A
GSE15659_RESTING_TREG_VS_NONSUPPRESSIVE_TCELL_DN	193	10	94.8	YAP1, PNPT1, RNF19A, EMC4, ZMIZ2, ZCCHC9, ST7-AS1, UBL3, SAP30BP, RAB9A
GSE15750_WT_VS_TRAF6KO_DAY10_EFF_CD8_TCELL_UP	198	7	96.5	PDLIM1, DDX51, ACSL5, SPTLC1, DYRK3, NPAS3, ZNF823
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	199	29	85.4	CRCP, IDI1, CDKN2AIPNL, SERPINB9, PSMD12, PRIM2, ARFGAP3, MSMO1, HMBS, LGALS3, MTHFD2, CTLA4, FDFT1, CENPK, GLRX, NBN, MRPL18, PREP, MRPL17, HIP1R, TACC3, HERPUD1, SULT2B1, TG, TMEM159, SCD, HMGCR, TNFRSF9, PGLYRP1
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IFNAB_CD8_TCELL_UP	197	9	95.4	HCRT, EML5, CRY2, APP, NR1D2, SLC25A51, BTC, BRWD3, CD79B
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IFNAB_CD8_TCELL_DN	199	27	86.4	RPN1, CRCP, C8orf37, RPA3, IDI1, MCM2, MCM5, CDKN2AIPNL, NDUFV1, SERPINB9, PSMD12, KIAA0101, PRIM2, PSAT1, TRIM37, BRCA1, MSMO1, HMBS, LGALS3, MTHFD2, FDFT1, CENPK, GLRX, NBN, MRPL18, PREP, NDUFS6
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_TRICHOSTATINA_CD8_TCELL_DN	198	16	91.9	RPN1, CRCP, C8orf37, RPA3, IDI1, MCM2, ZMIZ2, MCM5, ORC6, CDKN2AIPNL, NDUFV1, MBD4, PSMD12, KIAA0101, PRIM2, TRIM37
GSE16522_MEMORY_VS_NAIVE_CD8_TCELL_DN	195	22	88.7	EML5, HSPA1B, JAK2, DSP, NAP1L1, GPR18, IL10, TMC7, USP37, DYDC2, PPP2R3A, FAS, CASP1, BST2, LRCH1, ZNF703, WDR83OS, IGF1R, EXO1, OR7C1, GARS, PIK3CA
GSE16522_ANTI_CD3CD28_STIM_VS_UNSTIM_NAIVE_CD8_TCELL_DN	199	12	94.0	F2R, C5orf28, BPHL, AP3M1, TM9SF1, ESPL1, PFN2, CCDC124, ESM1, TPCN1, CBX5, ALG1
GSE17580_TREG_VS_TEFF_S_MANSONI_INF_UP	196	21	89.3	IL18, FARSB, ITIH5, MCM5, YWHAE, TNFRSF13B, SERPINB9, CD79B, CD2, RIPK3, FBXW11, BRCA1, CCR6, TGFBR1, EHD4, DCLRE1A, CTLA4, TXN, GSTO1, CCR8, SLC52A3
GSE17721_CTRL_VS_POLYIC_1H_BMDM_UP	197	17	91.4	SLC40A1, RNF2, FAM213A, SPC25, FRMD8, EVI5, SDHC, CYP4F3, LMBRD1, NKAIN1, HEATR5B, WDR20, E2F8, HEYL, SLC7A1, SNX15, CUL4B
GSE17721_CTRL_VS_POLYIC_6H_BMDM_UP	195	15	92.3	NR1D2, HIGD2A, ATP6V1C1, AKR1C3, SPC25, SIRT3, EVI5, NAGK, PLSCR3, SEC16A, OSBP, MRPL51, SRP68, UBE2I, ZNF703

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE17721_CTRL_VS_POLYIC_24H_BMDM_UP	200	25	87.5	MCEE, ADCY8, FANCG, NUCB2, EDEM3, MUC1, GSR, PDK4, HAS3, MRPL51, PDYN, HEATR5B, HDGFRP3, GSTO1, HSCB, CUL4B, FKBP9, MYOZ1, CROT, ADK, ASCC1, RPL13, SSR4, NEDD8, RNF187
GSE17721_CTRL_VS_PAM3CSK4_0.5H_BMDM_DN	195	23	88.2	MAST1, PHF13, STXBP1, PACRG, SOAT1, MMADHC, CTSG, KANSL1L, BTC, BRWD3, SMOX, WNT2, COX6A2, MAPK13, ADAM33, DMC1, CNNM3, FCHO1, DCC, RPL39, RDH13, BTG3, AMELX
GSE17721_CTRL_VS_PAM3CSK4_8H_BMDM_UP	199	10	95.0	MRPL14, EML5, MCEE, IDS, TM9SF1, C19orf12, GRK5, TUBGCP3, MAPRE3, MAP4K4
GSE17721_CTRL_VS_CPG_1H_BMDM_DN	199	14	93.0	CHORDC1, PEX5L, CYP2B6, OMP, SERPINB9, CHIC2, HSPA8, DSCAM, CNTN2, SCML2, FAS, NKX6-1, MTHFD2, CBX4
GSE17721_CTRL_VS_GARDIQUIMOD_0.5H_BMDM_UP	198	20	89.9	ROCK2, MRPL14, CEP350, CRY2, MCM2, EIF4B, RPE, GPC6, GATA2, NAA30, TNFRSF13B, PITPNB, GRK5, MUC1, KLF16, SOCS5, ADTRP, CASP1, OPA1, EFHD2
GSE17721_CTRL_VS_GARDIQUIMOD_12H_BMDM_UP	198	14	92.9	SLC40A1, ADIPOQ, MCEE, SPC25, IDS, ZNF124, TNFRSF13B, CYP4F3, CENPV, KXD1, HTATIP2, MRPL51, FRRS1, PDYN
GSE17721_LPS_VS_POLYIC_24H_BMDM_UP	195	24	87.7	APP, GMFG, PSMD12, IL10, KIAA0020, PFKFB1, VPS41, HAS3, UFM1, TMED4, FASTKD2, SWI5, GSTO1, SEC61G, RDH10, METTL22, ABRACL, PPAP2B, MRPS18B, YWHAQ, RPP14, NIF3L1, IPO9, VTA1
GSE17721_POLYIC_VS_PAM3CSK4_4H_BMDM_DN	190	30	84.2	HSPA1B, SSR1, RAB6A, MRPL44, OLR1, PDCD4, RNF145, CAB39, CDV3, GDAP2, SBNO1, EDEM3, RCL1, UBE2G1, CTNND1, LMBRD1, SUOX, WDR45B, SLC44A1, ACADVL, RDH10, LAMTOR3, FCHO1, CD72, PHF12, PDE12, PLEC, TTL, RGS5, CLEC10A
GSE17721_PAM3CSK4_VS_CPG_1H_BMDM_DN	196	19	90.3	DAPP1, TNFSF14, RSRP1, TBC1D15, APBB3, BTG1, PSMA1, NFKBIZ, MPO, FAS, ZDHHC4, SLAMF7, MTHFD2, CBX4, SAP30BP, DNASE1L2, RPE65, SPSB1, ABRACL
GSE17721_PAM3CSK4_VS_CPG_4H_BMDM_UP	196	19	90.3	POT1, ORMDL1, METTL20, MRPL44, MPZL2, DNMT3B, FEM1A, COLGALT1, GDAP2, CHPF, PEA15, FAM120A, ZNF600, LAMTOR3, FCHO1, FANK1, TBX5, TMEM14C, DNAJC3
GSE17721_CPG_VS_GARDIQUIMOD_16H_BMDM_UP	198	26	86.9	YAP1, ACTN1, BPHL, ANGPT1, HBZ, GNB3, DDX5, PLA2G4F, UBE2G1, CD34, CCR6, TSKS, DSCAM, VEGFB, TSPO, GSTO1, SLC5A11, IL17RD, CR2, FKBP9, SLC5A9, HLA-DMA, MX1, CWH43, PKDCC, PGLYRP1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE17721_LPS_VS_PAM3CSK4_12H_BMDM_DN	195	23	88.2	MAST1, ORMDL1, BANF1, DBI, ATP5J2, SDHC, RMND5B, GDAP2, YPEL3, RPL5, PEA15, MPO, MTIF2, SCN7A, PROSC, SWI5, IL1B, HSD17B7, CLEC5A, PBDC1, ARHGEF12, NEDD8, MFF
GSE17721_PAM3CSK4_VS_GADIQUIMOD_4H_BMDM_UP	197	27	86.3	AKR1D1, ORMDL1, SPC25, DHPS, EVI5, THAP11, MIS12, PEA15, ADGRD1, SOX2, DSCAM, IMPACT, HELQ, SERPINE1, POLR2I, BCL3, TMEM51, FCHO1, FANK1, TBX5, TPM3, ENO2, CD72, HLA-DMA, MRPS18B, RNF181, ARHGEF12
GSE17721_PAM3CSK4_VS_GADIQUIMOD_6H_BMDM_DN	198	18	90.9	BPHL, PHF13, STXBP1, FEZ2, LRR1, COX15, HINFP, SP110, FBXW11, TOMM70A, EHD4, NKX6-1, TSPO, SLAMF7, LIPA, SLC25A25, RAB9A, OSER1
GSE17721_LPS_VS_CPG_1H_BMDM_UP	198	16	91.9	NKIRAS1, ITIH5, ACSL5, MCM5, EVI5, SEMA6C, SELT, CDC6, BRWD3, COX19, EIF5B, KCNJ9, PDCL2, FBXL14, MANBAL, MPP6
GSE17721_LPS_VS_CPG_4H_BMDM_UP	199	19	90.5	PNPT1, ZNF644, C19orf12, PRIM2, PSAT1, HINFP, ARHGAP8, SP110, XK, CHPF, TOMM70A, CPSF2, CTLA4, BST2, AP1M1, JARID2, RAB9A, SLC52A3, KIAA1033
GSE17721_POLYIC_VS_GARDIQUIMOD_24H_BMDM_UP	197	29	85.3	KLF6, ZMIZ2, FRYL, UBR4, COMMD7, IDS, SMOX, ISY1, TTC36, OSBPL9, DSCAM, EHD4, S100A8, CCNA1, SRP68, HMP19, RELA, LIPA, HSCB, EFHD2, CNOT3, SNTB2, GFRA3, LGALS3BP, AP1G2, PRRG2, ZFP36L2, MCL1, DPYS
GSE17721_LPS_VS_GARDIQUIMOD_24H_BMDM_DN	196	23	88.3	IL18, FARSB, EIF4B, FAM213A, SPC25, DHPS, BANF1, SELT, LAIR1, ATP5O, TMEM263, MORF4L1, SMC2, ACSL4, S100A8, ZCCHC17, PICALM, SLC44A1, TMEM51, LAMTOR3, PHTF2, ABRACL, SYT11
GSE17721_0.5H_VS_4H_LPS_BMDM_UP	199	25	87.4	CRCP, NR1D2, MCM2, STOML2, ZMIZ2, EVI5, NDUFB3, ABCF1, SEC16A, RHOG, KIAA0020, OSBP, TIMM10B, TGFBR1, KLF16, HTATIP2, MRPL51, FRRS1, NDUFA10, WDR20, LDLRAP1, SUPT16H, TMEM14C, DPH2, LIMK1
GSE17721_0.5H_VS_24H_POLYIC_BMDM_DN	197	29	85.3	STXBP1, CSF2RA, RPE, ATP6V1C1, SLC2A5, SPTLC1, PLEKHA3, NAA30, CHIC2, SLC29A3, QSER1, EDEM3, SMOX, FBXW11, ITGA4, TIMM10B, HTATIP2, AIG1, SYF2, UBE2I, THRAP3, ATP2C1, CDK5RAP2, ASB3, IQSEC1, DDX17, GDI1, GNG11, CLEC5A
GSE17721_0.5H_VS_12H_PAM3CSK4_BMDM_UP	199	12	94.0	MGMT, FAM213A, HELB, KLF6, EVI5, SDHC, COX15, STEAP3, DMRTB1, ELK4, DPYSL5, EHD4
GSE17721_0.5H_VS_8H_PAM3CSK4_BMDM_UP	198	11	94.4	MGMT, RPN1, SLC25A51, EVI5, YIPF1, 42065, SDHC, TUBGCP3, ATP5O, PQLC1, EHD4

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE17721_0.5H_VS_24H_GARDIQUIMOD_BMDM_DN	196	23	88.3	RAB18, UBR4, TJP1, GNPTAB, SERPINB9, DYNC1LI1, TMEM263, FBXW11, DENR, PTGIR, SYPL1, PSMB7, FCHO1, IL18BP, LOXL3, ABRACL, CMTR1, CDK5, ARHGEF12, PGAM2, TMEM199, RABIF, ANXA6
GSE17721_4_VS_24H_GARDIQUIMOD_BMDM_UP	198	17	91.4	KLF6, NAP1L1, SELT, KIAA0101, TRAPPC6A, SDHC, ZNF318, RNASE2, TIGD5, UFM1, DENR, PLEKHF1, SRP68, AP1M1, RELA, TRIB1, IL1B
GSE17974_0H_VS_0.5H_IN_VITRO_ACT_CD4_TCELL_UP	176	19	89.2	LILRB1, CEPT1, TEPP, SCN8A, RPS6KL1, SYCP2L, CCDC116, G0S2, EIF4ENIF1, KXD1, ALG12, TRIP10, ZNF337, GABRB1, CCDC83, SCGB3A1, SUSD4, GJC2, C4BPA
GSE17974_0H_VS_4H_IN_VITRO_ACT_CD4_TCELL_UP	182	25	86.3	LINC00936, LOC646870, G0S2, TFDP2, ITGA4, PSTK, R3HDM2, RFX3, SUSD4, CORO1B, RAP1GAP2, GJC2, C4BPA, COMMD9, SAMHD1, FHIT, DIS3L2, MT1H, IGBP1, DIAPH2, PCNT, FCGRT, LPP, PIK3R5, NPC2
GSE17974_0H_VS_4H_IN_VITRO_ACT_CD4_TCELL_DN	192	12	93.8	TNFSF14, EGR3, ZMYM6, GRWD1, NAGK, LEO1, C3orf52, GADD45GIP1, DHRS7B, ACSL4, ALG1, MPP6
GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_UP	185	20	89.2	OTUD1, NR1D2, PLEK, PNMA3, TMEM156, NFKBIZ, MAP4K4, ITGA4, PSTK, MBNL2, SUSD4, NET1, LOC440704, RAP1GAP2, C4BPA, HCST, FHIT, RNF125, AUTS2, CRAMP1L
GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_DN	195	35	82.1	LDHC, MCM5, DHPS, EGR3, BANF1, LRR1, ZNF557, COLGALT1, KIAA0020, GGCT, ACSL4, MTHFD2, POLR2I, SLC7A1, GNG8, ATP6V1F, PREP, SLC27A2, IMMT, MRS2, PGBD2, MRPL17, PEAK1, NFXL1, MITD1, ACOX1, MLYCD, PSMB3, HSD17B10, TMEM120A, MMACHC, FAM98A, FEN1, TMEM5, C1GALT1C1
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_1H_CD4_TCELL_UP	178	13	92.7	JUN, G0S2, PRR12, CCDC124, TUBGCP4, KXD1, FDCSP, OCLM, THAP10, SNTB2, RAP1GAP2, TMEM145, C4BPA
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_4H_CD4_TCELL_UP	186	15	91.9	LINC00936, SGTB, JUN, 42065, G0S2, ITGA4, PSTK, AIG1, TBC1D5, RFX3, SUSD4, MPP7, RAP1GAP2, C4BPA, COMMD9
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_72H_CD4_TCELL_DN	197	38	80.7	IDI1, STXBP1, MGAT1, ZNF826P, YWHAE, DHPS, DCP1B, DBI, GRHPR, C4orf3, FAM89A, SLC25A43, IGSF8, DHRS1, BRIP1, THYN1, ACADVL, MRPL18, EXO1, ATP6V1F, IMMT, MPG, MEOX1, JPH1, USP28, ASCC1, SYT11, NFXL1, NIF3L1, POLD3, RUSC1, C12orf75, CPNE2, TMEM120A, ZNF410, FAM98A, C1GALT1C1, HOMER1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE17974_IL4_AND_ANTI_IL12_VS_UNTREATED_12H_ACT_CD4_TCELL_UP	187	22	88.2	EPAS1, SDC4, RNF19A, LOC642852, HIPK1, STAT4, NIPA1, MPZL2, MOSPD3, NCR3LG1, LEO1, BUD13, GPALPP1, ANTXR2, MFSD6, UBL3, PPP1R14A, TRIB1, NET1, PHF20L1, UFD1L, FOXL1
GSE17974_IL4_AND_ANTI_IL12_VS_UNTREATED_48H_ACT_CD4_TCELL_UP	186	22	88.2	#N/A
GSE17974_1.5H_VS_72H_IL4_AND_ANTI_IL12_ACT_CD4_TCELL_DN	194	27	86.1	BAX, DHPS, MPZL2, PIH1D1, DBI, FSD1, HEXIM2, GRHPR, TCEAL3, BRIP1, FAM120A, ZNF589, THYN1, ACADVL, ATP6V1F, ARAP3, ZNF692, WDR76, PEAK1, PARVB, DHTKD1, CDK5, OXSM, POLD3, STX10, C12orf75, HOOK3
GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_18H_UP	192	23	88.0	ROCK2, AP3M1, TRPV2, POLR3B, FBXW11, RMDN1, HCFC2, SUOX, FLVCR2, EIF4EBP2, ARHGAP12, PREP, SMIM15, KIF16B, TTI1, TMEM245, FBXO21, MTMR12, METTL7A, HSD17B10, RNF141, NDUFS2, TSHZ1
GSE20366_EX_VIVO_VS_HOMEOSTATIC_CONVERSION_TREG_UP	197	41	79.2	F2R, CLK4, CEP350, TASP1, STAU1, PODXL, CHIC2, AHI1, TUBGCP4, GNPNAT1, MSMO1, CREB3L2, APOBEC3B, STX5, PIGU, SLC52A3, CD72, OTUD4, SPACA1, SPATA5L1, TDP2, DNAJA1, TNFRSF9, RSRC2, CBLB, NSMF, DXO, PDK1, HOMER1, TCERG1, TRPT1, TBCE, UCP2, OTUD6B, GRHL1, TBCCD1, NR2F6, SEPP1, RBM33, NUDT19, TK2
GSE20366_EX_VIVO_VS_DEC205_CONVERSION_UP	197	19	90.4	EPAS1, CHORDC1, PNPT1, AHR, SDC4, IDI1, EGR3, GPR15, H6PD, RUNDC3A, IL10, FAM46C, MMEL1, SMC2, ANTXR2, ALG8, ZNF839, APOBEC3B, CKAP2L
GSE20366_EX_VIVO_VS_HOMEOSTATIC_CONVERSION_NAIVE_CD4_TCELL_UP	196	44	77.6	EML5, TASP1, BRD8, TAF4, DNMT3B, ST6GALNAC3, TUBGCP3, TUBGCP4, VPS41, MSMO1, TGFBR1, ITGA1, MTIF2, GLCCI1, FDFT1, ZNF330, GAK, ARHGAP12, ZUFSP, PAPOLG, TDP2, HERPUD1, ZMYM5, ZNF566, FAM160A2, EFTUD1, SLC16A6, SFSWAP, ZNF180, ZNF790, PDE5A, GIN1, PDK1, HOMER1, CPSF1, INPP1, DNAJC13, TBCE, SQLE, SPG11, CRBN, TAF1C, MYCBP2, KIAA1191
GSE20366_EX_VIVO_VS_DEC205_CONVERSION_NAIVE_CD4_TCELL_UP	194	32	83.5	EPAS1, UNC5CL, CEPT1, AP3M1, FRYL, FRA10AC1, TTC21B, CCR6, NDC80, MTIF2, SLAMF7, PLEKHF1, AMPD1, HMP19, FAM20A, BUB1, GZMA, MITD1, METTL7A, SLC16A6, UBXN7, TRAF3IP1, DNA2, PRC1, INPP1, TTPAL, ANKRD6, TBCCD1, FAM109B, CLSPN, RCBTB2, GEMIN6
GSE20366_CD103_POS_VS_NEG_TREG_KLRG1NEG_UP	195	25	87.2	EPAS1, CYLC2, SGTB, CCDC33, MARS2, ZMYND10, GADD45GIP1, KCNJ9, C10orf76, LGALS3, ZMAT5, DKK1, GPR63, GABRA5, TXNDC2, HOXD9, ITPKA, ITGA2B, C2CD4B, GNRHR, SLC25A42, FAM170A, SPACA7, OPN1LW, MX1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE20715_0H_VS_6H_OZONE_TLR4_KO_LUNG_DN	199	25	87.4	SLC44A4, DSP, KLF6, SAR1B, TACSTD2, PSMD12, RNF145, DDX5, TCEB3, RMDN1, GSTA3, FERD3L, ELL2, LARP6, ANKRD44, CTGF, ALDH3A1, USP20, CROT, OTUD4, PDE12, ADRB2, ESRP2, HERPUD1, PSMD7
GSE22045_TREG_VS_TCONV_UP	179	22	87.7	LILRB1, F2R, SDC4, C6orf165, ALDH9A1, CD58, PPP2R2B, ZC2HC1A, RBM42, ADGRD1, CCR6, HSPA8, ADAM33, HTATIP2, MFSD6, FAM160B1, KIAA1841, RDH10, CORO1B, CMYA5, RNF181, SPRN
GSE22886_NAIVE_CD8_TCELL_VS_MEMORY_TCELL_DN	198	23	88.4	AHR, CASP2, CTSB, APP, HIF1AN, CAB39, CAPZB, ZC2HC1A, GSR, UFM1, FAS, BLVRA, EIF4EBP2, SNX15, CORO1B, HNRNPA0, PHTF2, WDR76, MTM1, CCDC47, DEGS1, RNASET2, SESN1
GSE22886_NAIVE_TCELL_VS_NKCELL_DN	197	18	90.9	IGFBP7, AKR1C3, MTMR3, FRYL, HIPK2, FEZ2, YIPF6, ZCCHC6, DHRS7B, NCAM1, CREB3L2, JARID2, PREP, LTF, CSNK1D, CLASP1, CCDC47, IST1
GSE22886_NAIVE_CD8_TCELL_VS_NKCELL_DN	196	13	93.4	HNRNPK, AKR1C3, HIPK1, MTMR3, HIPK2, SELT, ZCCHC6, NCAM1, CREB3L2, NKAIN1, MFSD6, JARID2, PREP
GSE22886_CD8_TCELL_VS_BCELL_NAIVE_UP	197	28	85.8	GNLY, B4GALT3, STAU1, ZMYM6, NUCB2, PLSCR3, NKG7, MAP7D1, CD2, PLEKHF1, THYN1, TSPAN32, MRPL57, LDLRAP1, KLRC3, EFHD2, NDFIP1, APOL3, ENO2, MRPL17, ATP13A2, GZMA, TACC3, ATP2B4, TIMP1, INPP4A, GBAP1, KIAA0391
GSE22886_NAIVE_VS_IGG_IGA_MEMORY_BCELL_DN	192	18	90.6	PPA2, BLVRB, CD58, RARRES3, NDUFB3, CAB39, PLSCR3, CDV3, UBE2G1, DEPDC5, HSPA8, PYCARD, NDUFS6, APOL3, ENO2, GARS, TSFM, PPA1
GSE22886_IGA_VS_IGM_MEMORY_BCELL_DN	196	23	88.3	FKTN, ULK1, JAK2, PLEK, AFM, GRIK3, HIST1H2BK, CD1C, GPR18, GRWD1, KMT2A, BTG1, CBFA2T2, CD79B, OSBP, DNASE1L3, MYOZ2, ZXDC, DEPDC5, CREB3L2, ZNF589, RAB35, FCGR2B
GSE22886_IGG_IGA_MEMORY_BCELL_VS_BM_PLASMA_CELL_DN	189	17	91.0	KLF6, GRWD1, TM9SF1, NUCB2, RUNDC3A, FER1L4, ARFGAP3, MUC5B, CHPF, HSPA13, LGALS3, EHD4, CREB3L2, MBNL2, BLVRA, BST2, TMCO1
GSE22886_IGM_MEMORY_BCELL_VS_BM_PLASMA_CELL_UP	197	29	85.3	APP, CD1C, SLC24A1, DYNC1LI1, CD79B, RPL11, SP110, SYPL1, PPID, SUPT16H, MORC3, OPA1, PMS2P5, CR2, CASP8AP2, NOTCH2NL, RPL39, HMBOX1, CD72, HLA-DMA, NUP43, TDP2, RPSA, LRRC31, RALYL, POLD3, PIK3CD, RBM5, SF3A2
GSE22886_IGM_MEMORY_BCELL_VS_BM_PLASMA_CELL_DN	192	14	92.7	YIPF3, MAST1, MGAT1, B4GALT3, GNB3, ZNF706, TM9SF1, GADD45GIP1, FER1L4, ARFGAP3, PFKFB1, MUC5B, SURF1, EHD4

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE22886_DAY0_VS_DAY7_MONOCYTE_IN_CULTURE_DN	200	24	88.0	EPAS1, CTSB, ATP6V1C1, ALDH9A1, TRPV2, ACSL3, NDUFB3, GOT1, DBI, SLC29A3, SLC17A5, STEAP3, GGCT, MSMO1, COL8A2, MAPK13, VEGFB, AFG3L2, MCUR1, DCSTAMP, FOCAD, TIMM17A, DENND4C, PARVB
GSE22886_DAY1_VS_DAY7_MONOCYTE_IN_CULTURE_DN	198	23	88.4	EPAS1, MRPS27, HIGD2A, RPE, FEZ2, SLC29A3, GGCT, RSU1, OSBPL9, METTL5, ALG8, COL8A2, VEGFB, TBC1D5, FDFT1, HLA-DMA, ALDH2, ATP5B, MRPS18B, CUTA, DENND4C, CDK5, SESN1
GSE22886_NEUTROPHIL_VS_DC_DN	200	35	82.5	GSTP1, SOAT1, BANF1, TRPV2, NOL7, ANKRD17, PEA15, FBXL14, TIMM10B, EHD4, MTHFD2, SYPL1, TXN, ATP5H, ATP6V1F, CCDC47, HLA-DMA, ATP5B, SSR4, FBXO21, GTF2H5, GBAS, SUCLG1, AKIP1, PEBP1, GPR137B, PFDN1, GPX3, SEC31A, APPL1, GTF2A2, MDFIC, CD81, MPV17, NDUFS1
GSE22886_NAIVE_TCELL_VS_DC_DN	200	14	93.0	EPAS1, RPE, BLVRB, FEZ2, ADAM17, LRP12, PEA15, EHD4, SUOX, MTHFD2, FLVCR2, LIPA, ATP6V1F, PREP
GSE22886_NAIVE_CD4_TCELL_VS_DC_DN	198	21	89.4	CTSB, GSTP1, CD58, CSNK2B, TM9SF1, SMC2, RAB21, VPS41, MTHFD2, FLVCR2, TMCO1, LIPA, LAMTOR3, FOCAD, TIMM17A, HLA-DMA, ATP5B, SLC30A1, ST14, RAB31, MFAP1
GSE22886_UNSTIM_VS_IL15_STIM_NKCELL_DN	198	19	90.4	RPA3, MCM2, STOML2, MCM5, CDC6, GGCT, TUBGCP3, SLCO4A1, CKS2, MSMO1, EIF2S1, USP1, DCLRE1A, MCUR1, PSMD6, SLC7A1, MED20, MRPL17, GZMA
GSE24081_CONTROLLER_VS_PROGRESSOR_HIV_SPECIFIC_CD8_TCELL_DN	190	32	83.2	RPN1, CEACAM8, JAK2, DSP, PLEK, AFM, GART, IRF9, MLANA, C1GALT1, TRIP10, APOBEC3B, SLAMF7, CHRNA2, SLC22A8, C1orf112, PYCARD, THAP10, EXO1, CASP8AP2, SNTB2, RAP1GAP2, MSX2, HIST1H2AC, EIF2AK2, UNG, POF1B, PCOLCE2, METTL7A, PARD6B, PTGS2, ZNF695
GSE24634_NAIVE_CD4_TCELL_VS_DAY3_IL4_CONV_TREG_DN	198	31	84.3	RPA3, NDUFB3, ESPL1, TFCP2, SMC2, BLVRA, UBE2I, SUPT16H, EXO1, TRIB1, IMMT, ETF1, CCDC47, MED20, PPAP2B, COPS7A, MRPL17, TACC3, MELK, POLD3, TMPO, METTL7A, HSD17B10, MAPK1, MYB, KIF3A, MRPL11, RFC5, MED22, WARS, PSME2
GSE24634_NAIVE_CD4_TCELL_VS_DAY10_IL4_CONV_TREG_DN	199	28	85.9	CTSB, ACSL5, TXNDC15, ATP6V1C1, ALDH9A1, CD58, HIST1H2BK, TM9SF1, KIAA0101, AHI1, CD79B, GPALPP1, DHRS7B, CD2, SP110, NDC80, LGALS3, CREB3L2, DHRS1, ABHD5, SMARCD2, SPATS2L, PYCARD, ARPC2, TRIB1, IMMT, PHTF2, CCDC47
GSE24634_TREG_VS_TCONV_POST_DAY3_IL4_CONVERSION_DN	199	12	94.0	IL18, IGFBP7, CTSB, CSF2RA, ATP6V1C1, PLEK, NAGK, IL10, LAIR1, C10orf76, LGALS3, EHD4

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE24634_TEFF_VS_TCONV_DAY7_IN_CULTURE_UP	195	24	87.7	MCM2, CD58, SNAPC3, FANCG, ESPL1, TIMELESS, CD79B, CD2, CKS2, ITGA4, NDC80, CREB3L2, DCLRE1A, CTLA4, CCR8, PMS2P5, TRIB1, SLC27A2, MCCC2, MYH10, DLGAP5, EGR1, SIRPG, GZMA
GSE25087_TREG_VS_TCONV_ADULT_DN	185	7	96.2	APP, FAM213A, KIF9, CEPT1, HIPK2, GIMAP5, MTHFD2L
GSE26669_CTRL_VS_COSTIM_BLOCK_MLR_CD4_TCELL_DN	195	23	88.2	N4BP2, ZNF799, RSPRY1, MBD4, CHIC2, YPEL3, WBSCR27, TTC36, RHOT1, ITGA4, CCR6, PTK2B, MTIF3, TBC1D5, MCUR1, DIRC2, FAM134C, TES, SEC11A, IQSEC1, PPAP2B, SLC30A1, PTGFRN
GSE26669_CTRL_VS_COSTIM_BLOCK_MLR_CD8_TCELL_DN	199	6	97.0	DDX6, EML5, PRKAB2, SDC4, HBP1, IDS
GSE26669_CD4_VS_CD8_TCELL_IN_MLR_COSTIM_BLOCK_DN	196	14	92.9	EML5, NR1D2, HIPK1, BCO2, MYOZ2, GDAP1L1, RHD, FRRS1, GSTO1, CDK5RAP2, NLRC3, NDUFS6, IQSEC1, RAP1,GAP2
GSE26928_CENTR_MEMORY_VS_CXCR5_POS_CD4_TCELL_DN	180	16	91.1	CREB1, NPAS3, TAS1R2, QSER1, GANC, ZCCHC6, TFCP2, OR10H3, GPR75, VPS41, GLCCI1, RAD21, HMHB1, MBOAT1, SEC31B, CLEC5A
GSE2706_R848_VS_R848_AND_LPS_8H_STIM_DC_UP	178	16	91.0	AHR, SGTB, RAB6A, MFAP5, BCO2, NAGK, SLC29A3, PSAT1, ZNF786, MPZL3, PARPBP, SPOPL, CTLA4, ERO1L, EGR2, NOS1AP
GSE27786_LSK_VS_BCELL_UP	197	19	90.4	ACTN1, IDI1, DDX56, HBZ, MCEE, FAM73A, KANSL1L, DNMT3B, PDE9A, PRIM2, KIAA0020, LAIR1, EHBP1, GSR, ADGRD1, UTP20, PQLC1, PDK4, BOD1
GSE27786_LSK_VS_ERYTHROBLAST_UP	198	19	90.4	PDLIM1, PPHLN1, IDS, DBI, ZC2HC1A, RMND5B, LAIR1, EI24, TSKS, URGCP, FRRS1, ZDHHC4, NUDT6, MTM1, KIAA1033, ZNF623, MUM1L1, STAT5A, TTI1
GSE27786_LIN_NEG_VS_BCELL_UP	197	9	95.4	MGMT, IDI1, EIF4B, ACSL5, KRT28, ATP1A2, ORC6, ZHX3, ACSL3
GSE27786_CD4_TCELL_VS_NKTCELL_DN	199	16	92.0	ENTPD7, KRT28, GRWD1, ZNF606, AKT1S1, MAP7D1, ARHGAP30, EDEM3, EI24, CENPV, METTL5, RFX7, HOXD9, ARPC2, EXOC6, ARAP3
GSE27786_NKTCELL_VS_ERYTHROBLAST_UP	199	19	90.5	TCF25, CAPNS1, ZNF606, GRK5, SURF1, CLDN12, ZDHHC4, LYPLA1, ZC3H13, GSTO1, OPA1, MBTPS2, CSNK1D, SLC30A6, SNTB2, PHF23, EGR1, SLMAP, MRPL42
GSE2826_WT_VS_XID_BCELL_DN	198	15	92.4	EPAS1, APP, ENTPD7, SLC2A5, ALDH9A1, VCAM1, EMCN, CFHR2, SERPINA5, SCN7A, UBL3, ZNF703, P4HA2, CELF4, SLC30A1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE2826_WT_VS_BTK_KO_BCELL_DN	199	17	91.5	EPAS1, IL18, SLC44A4, ACTN1, CTSB, EPO, PDLIM1, YWHAE, VCAM1, ELK4, MAP7D1, PEA15, MAPK13, KLF17, S100A8, GLRX, NBN
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	170	14	91.8	AIRE, FSHR, C6orf62, GNLY, ZDHHC2, CXorf36, FAM73A, TJP1, ESPNL, GABRB1, SYF2, ALDH3A1, CHRM5, LOC55338
GSE29618_BCELL_VS_MONOCYTE_UP	179	28	84.4	DDX6, PDLIM1, CD24, ARPP19, RPS23, TNFRSF13B, FAM46C, BTG1, TMEM156, CD79B, CCR6, ZNF43, MPP6, SYPL1, PHTF2, CD72, RABEP1, AEN, RPSA, STK17A, DDX24, ZNF93, SHMT2, CBLB, CD22, S1PR1, ABLIM1, ZNF675
GSE29618_BCELL_VS_MONOCYTE_DN	200	16	92.0	AHR, ACTN1, LILRB3, IGFBP7, CTSB, MGAT1, GMFG, PLEK, BLVRB, RHOG, RNASE2, PEA15, RAB20, BLVRA, CASP1, PID1
GSE29618_BCELL_VS_PDC_UP	186	23	87.6	PDLIM1, CD24, RARRES3, JUN, TNFRSF13B, MBD4, ZNF318, FAM46C, BTG1, CD79B, CCR6, ALOX5, SYPL1, PHTF2, FCGR2B, APOL3, CD72, TRAK2, EGR1, KIAA0922, RYK, RHOB, CHST2
GSE29618_BCELL_VS_MDC_UP	183	48	73.8	CD24, GPR18, PDCD4, TNFRSF13B, KMT2A, ZNF430, FAM46C, GRK5, BTG1, CD79B, MAP4K4, ZNF43, ELL2, MPP6, SYPL1, PHTF2, KIAA1033, CD72, EGR1, RYK, ZFP36L2, STK17A, SGCE, RBM5, CBLB, CD22, S1PR1, ABLIM1, CD81, P2RY10, TPD52, KLF2, ITPR1, KIAA1551, DMXL1, RAB30, COBLL1, ZNF107, MYC, RRAS2, CD47, POU2AF1, P2RX5, PIK3IP1, RASGRP2, TSPYL1, PIK3C2B, ARID5B
GSE29618_MONOCYTE_VS_MDC_UP	200	22	89.0	LILRB1, ARHGEF40, LILRB3, CTSB, CXCL8, GMFG, FCN1, GIMAP5, G0S2, PLAGL2, VPS37C, LILRA6, MAP4K4, MPO, ELL2, BLVRA, ABHD5, S100A8, EGR2, EGR1, BEST1, ZDHHC7
GSE29618_BCELL_VS_MONOCYTE_DAY7_FLU_VACCINE_UP	185	37	80.0	DDX6, PDLIM1, NPM1, CD24, RPS23, RPL9, GPR18, PDCD4, TNFRSF13B, MBD4, CCNB1IP1, ZNF430, FAM46C, ZNF273, BTG1, CD79B, SP110, CCR6, ZNF43, SYPL1, PHTF2, CD72, ZMYND8, RABEP1, AEN, RPSA, VPREB3, SLC25A38, SKAP1, CD22, ABLIM1, KAT6A, SLC50A1, BARD1, SETBP1, P2RY10, TPD52
GSE29618_BCELL_VS_MDC_DAY7_FLU_VACCINE_UP	182	41	77.5	DDX6, PDLIM1, CD24, GPR18, PDCD4, TNFRSF13B, MBD4, ZNF273, BTG1, CD79B, SP110, MAP4K4, ZNF43, SYPL1, PHTF2, ZMYND8, VPREB3, STK17A, SKAP1, PAWR, ZNF24, CD22, S1PR1, RNF141, ABLIM1, SIPA1L1, KAT6A, CD81, SETBP1, P2RY10, TPD52, ZNF665, EHD1, COBLL1, CD69, MYC, RRAS2, CD47, SMAGP, POU2AF1, P2RX5

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE29618_MONOCYTE_VS_MDC_DAY7_FLU_VACCINE_UP	200	28	86.0	CTSB, CXCL8, FCN1, MTMR3, GIMAP5, G0S2, GRK5, LAIR1, LGALS3, ELL2, BLVRA, S100A8, SYPL1, GLRX, FOLR2, ARAP3, EGR1, CLIP4, SASH1, STAB1, PRRG4, BEST1, ZDHHC7, PLBD1, NPC2, SIPA1L1, FCAR, SMPDL3A
GSE29618_PRE_VS_DAY7_POST_LAIV_FLU_VACCINE_MONOCYTE_UP	194	35	82.0	EPAS1, TMPRSS15, ADCY8, RNASE2, PAFAH1B2, FBXO46, ASCC2, GSTA3, MFSD6, SAP30BP, CR2, MBTPS2, CLDN14, DCC, TIMP3, IQSEC1, LIMK1, TRPV5, NTN1, ASB6, CWF19L1, KIFC3, TTYH1, ATG9A, GOLT1B, PRPF8, LCE2B, MORC2, COPB1, CST1, KAT2A, TMCC2, TRAF6, GUK1, DEPDC1
GSE30083_SP1_VS_SP3_THYMOCYTE_DN	197	26	86.8	GATSL3, CEP97, RBMXL2, C3orf52, OR5D18, YPEL3, GJA3, RAB21, HAS3, TSPO, LDLRAP1, PYCARD, KCNC2, RNF213, PCED1B, RP1L1, PHF20L1, TRIM34, SGK1, ANKRD26, IL6R, IRGM, KCNE5, IL17RA, S1PR1, SMPDL3A
GSE30083_SP1_VS_SP4_THYMOCYTE_DN	196	26	86.7	APP, C19orf12, KMT2A, FAM46C, RAB21, LDLRAP1, ADAMTS20, RNF213, SNTB2, PCED1B, PIGQ, TRIM34, NLRC5, AP1G2, DSE, ESRP2, SESN1, SGK1, SPN, IL6R, IRGM, MLYCD, IL17RA, S1PR1, AHNAK, SMPDL3A
GSE30083_SP2_VS_SP3_THYMOCYTE_DN	195	15	92.3	CEP97, RBMXL2, FBXL14, ITGA4, IGSF8, SLC22A3, SLA2, PLEKHF1, FOXC1, CERCAM, LDLRAP1, PALM, LTF, SLC5A9, ANGPTL2
GSE30083_SP3_VS_SP4_THYMOCYTE_DN	193	21	89.1	SDC4, APP, TNFSF14, PLEK, STAT4, NUCB2, C19orf12, NKG7, IRF9, LAIR1, PEA15, ZXDC, ITGA4, TBC1D5, CASP1, ERO1L, LDLRAP1, SNTB2, LGALS3BP, DSE, SGK1
GSE30962_PRIMARY_VS_SECONDARY_ACUTE_LCMV_INF_CD8_TCELL_DN	196	15	92.3	NR1D2, PHF13, TNFSF14, RSRP1, ZDHHC2, DYRK3, FBXO27, HIF1AN, PDZRN3, HMX1, PEA15, ZXDC, SUOX, NSUN6, SOWAHB
GSE30962_ACUTE_VS_CHRONIC_LCMV_PRIMARY_INF_CD8_TCELL_UP	194	38	80.4	GABARAPL2, PRKAB2, NR1D2, PDLIM1, PHF13, ZNF652, CEP97, KRTCAP2, G0S2, SMOX, YPEL3, RSU1, ESM1, PEA15, ANTXR2, EVA1B, METTL23, SYF2, RFX3, FCGR2B, RAP1GAP2, KIAA0922, STK38, ADRB2, PEAK1, GZMA, INSL6, GAB3, SESN1, SGK1, ARHGEF2, FAM160A2, FAM104A, TNIP1, IP6K1, C10orf54, FCGRT, TRAF3IP1
GSE31082_DN_VS_DP_THYMOCYTE_DN	198	18	90.9	YIPF4, ENTPD7, N4BP2, FRYL, IDS, GMEB2, BRWD3, KIF3B, NFKBIZ, TAF8, TSPO, GPR146, SYF2, AMPD1, EPHB6, ZNF646, DNAJC3, TRAK2
GSE31082_DP_VS_CD4_SP_THYMOCYTE_DN	193	18	90.7	ACTN1, SNAPC1, CTSB, JAK2, RPS23, FRMD8, CAPZB, HECTD2, PEA15, CCR6, PTK2B, ABHD5, CBX4, NDFIP1, LGALS3BP, EGR1, PARP9, DSE

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE32423_CTRL_VS_IL4_MEMORY_CD8_TCELL_UP	196	23	88.3	RNF2, C8orf37, APP, NR1D2, ZDHHC2, RSPRY1, DYRK3, TRIM63, OSBPL9, HSPA13, MAP4K4, RFX3, SYPL1, SYT10, LDLRAP1, HEY2, TES, PGAP1, MTMR10, THBS2, RPS3A, ZFP36L2, RHOB
GSE32423_IL7_VS_IL4_MEMORY_CD8_TCELL_UP	197	15	92.4	LRFN5, OTUD1, INHBC, DDX51, TEPP, TRPV2, GMEB2, COLGALT1, TIMELESS, ARHGAP30, OSBPL9, TPCN1, HSPA13, ACTL6B, ADD1
GSE3337_4H_VS_16H_IFNG_IN_CD8POS_DC_UP	196	16	91.8	MCM2, MCM5, YWHAE, CDK11B, POLR2I, METTL3, SLC44A1, ENSA, CSNK1D, ADPRHL2, ENO2, DBT, SPSB1, ADK, COPS7A, EIF4A3
GSE339_CD4POS_VS_CD8POS_DC_UP	194	14	92.8	ROCK2, TCF25, SYT2, GART, PITPNB, UHMK1, MARS, TUBGCP3, CPSF2, CCR6, BCL3, FOLR2, CD72, PHF12
GSE34205_HEALTHY_VS_RSV_INF_INFANT_PBMC_DN	200	38	81.0	CEACAM8, RNASE3, SPC25, CAMP, BLVRB, TRAF3IP2-AS1, KIAA0101, ESPL1, RNASE2, TTC36, PRTN3, XK, H1F0, MPO, C9orf66, S100A8, E2F8, TRIB1, CYSTM1, HBD, LTF, ZNF326, SERINC2, AMELX, ARF4, RAD51, GSPT1, DNAJC6, MELK, TCN2, FAM104A, PGLYRP1, HBQ1, DEFA4, CCNL1, PLBD1, TMEM52B, RHEB
GSE34205_HEALTHY_VS_FLU_INF_INFANT_PBMC_DN	199	26	86.9	RNASE3, FCN1, C4orf33, MICU1, IL10, RHOG, RNASE2, H1F0, MPO, MICB, FAS, C9orf66, MTHFD2, ERO1L, S100A8, SPATS2L, NUDT15, LTF, C10orf71, ZNF326, LGALS3BP, MX1, BTG3, AMELX, ARF4, PIWIL4
GSE34205_RSV_VS_FLU_INF_INFANT_PBMC_UP	177	14	92.1	GABARAPL2, DDX6, HBZ, ZNF461, RUNDC3A, FAM46C, RAX2, ESPN, 42071, KANK2, CNTN2, GPR146, MAGEB4, HBD
GSE360_CTRL_VS_L_MAJOR_MAC_DN	195	23	88.2	AHR, TOP3A, CTAG2, TRIM37, PLAGL2, GSR, CNTN2, HTATIP2, DCT, MBNL2, ELL2, ELAVL3, PROSC, EXOC6B, UPK1B, TWISTNB, IGF1R, MLN, DVL3, RAMP3, SAMM50, PSMD7, ANGPTL7
GSE360_DC_VS_MAC_T_GONDII_DN	195	13	93.3	TMPRSS15, IGFBP7, PIP4K2B, FCN1, SLC2A5, ZMIZ2, IDS, KRT75, VEGFB, ALOX5, CHMP2B, S100A8, FOLR2
GSE360_DC_VS_MAC_B_MALAYI_LOW_DOSE_DN	200	25	87.5	CEACAM8, WDR18, STAT4, IDS, HIST1H2BK, HERC2P3, YIPF1, G0S2, CAPZB, YAF2, SLC10A3, TRIP10, ZNF337, ADD1, BST2, RELA, CNOT3, DDX17, KIAA0922, COPS7A, MRPS18B, PARVB, PFDN4, RPP14, ANGPTL7
GSE360_DC_VS_MAC_M_TUBERCULOSIS_DN	195	8	95.9	PIP4K2B, RPA3, INHBC, TNFSF14, STOML2, ALDH9A1, IDS, ALOX5
GSE360_T_GONDII_VS_B_MALAYI_HIGH_DOSE_DC_DN	198	8	96.0	PIP4K2B, C18orf25, ZMIZ2, EVI5, NPAS3, ZNF124, MPO, ADD1

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE360_HIGH_VS_LOW_DOSE_B_MALAYI_DC_DN	194	22	88.7	PDLIM1, EIF4B, MNX1, ATP1A2, SFTPC, EGR3, MICU1, RUNDC3A, RLBP1, CEACAM7, SLC10A3, KRIT1, DSCAM, PDK4, APOBEC3B, CASP1, PSMB7, SPI1, ETF1, SEC31B, RABEP1, COL6A3
GSE360_LOW_DOSE_B_MALAYI_VS_M_TUBERCULOSIS_DC_UP	199	31	84.4	PIP4K2B, EIF4B, CAPZB, MARS, PFKFB1, RBM4B, HDAC3, RPL5, ADD1, ALOX5, CUL7, CNOT3, ALDH2, ERF, STK38, MDM2, RPS3A, HBE1, ZFP36L2, RNASE1, FOLH1, SGK1, IDH3A, STAB1, HMGA1, FCGRT, ADORA3, RASSF1, WDR43, FAM131A, SPAG7
GSE360_L_DONOVANI_VS_B_MALAYI_HIGH_DOSE_MAC_UP	196	18	90.8	CXCL8, ATP6V1C1, ADCY8, MCM5, CD1C, G0S2, BTG1, PFKFB1, SERPINE1, ZNF292, SEC61G, P4HA2, SLC27A2, DDX17, OTUD4, PRSS16, DVL3, SLC30A1
GSE360_L_MAJOR_VS_T_GONDII_MAC_UP	192	22	88.5	YIPF4, IL18, ACTN1, CXCL8, GNLY, STAT4, AFM, PPP2R2B, LEP, ZNF273, PDPN, CTNND1, MYOZ2, SLC25A26, CHPF, DMC1, FAS, GABRB1, LBX1, CTGF, TMCO1, HOXD9
GSE360_T_GONDII_VS_B_MALAYI_HIGH_DOSE_MAC_UP	195	11	94.4	IGFBP7, CXCL8, MCM2, RGS16, TUBB7P, MCM5, EGR3, CD1C, TM9SF1, KIAA0101, BTG1
GSE360_HIGH_DOSE_B_MALAYI_VS_M_TUBERCULOSIS_MAC_DN	195	14	92.8	IGFBP7, FKTN, RPA3, CXCL8, RGS16, TUBB7P, OLR1, MCC, TM9SF1, RNASE2, IDO1, CEACAM7, TNR, SERPINE1
GSE36392_EOSINOPHIL_VS_MAC_IL25_TREATED_LUNG_DN	196	22	88.8	ZNF799, PRR12, TRIM37, CD79B, SMC2, SLC25A26, FCF1, CREB3L2, TERF2, BOD1, THRAP3, LIPA, HSCB, SEC11A, NET1, ATP13A2, EPM2AIP1, TACC3, GCN1L1, NDUFA1, KIAA0391, TRAF3IP1
GSE36476_CTRL_VS_TSST_ACT_16H_MEMORY_CD4_TCELL_OLD_UP	196	11	94.4	TCL6, JUN, NUCB2, KRT75, FAM46C, CBFA2T2, QSER1, LAIR1, 42071, LILRA6, TBC1D5
GSE36476_CTRL_VS_TSST_ACT_72H_MEMORY_CD4_TCELL_OLD_UP	195	19	90.3	CRY2, SFTPC, JUN, IDS, FAM46C, CBFA2T2, PFN2, 42071, TBC1D5, UBL3, HMHB1, NOS1AP, SLC5A5, OSER1, TIMP3, IQSEC1, RAP1GAP2, ESRP1, C4BPA
GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_UP	188	18	90.4	#N/A
GSE37416_CTRL_VS_3H_F_TULARENSIS_LVS_NEUTROPHIL_UP	184	19	89.7	SLC40A1, RPL7L1, PRKAB2, HIPK2, CDKN2AIPNL, HCG27, EDEM3, RIPK3, MPZL3, MTIF3, C1orf168, GLRX, FAM134C, ARPC2, MED31, DCLRE1C, DEPDC4, PGLYRP1, ZNF180
GSE37416_CTRL_VS_12H_F_TULARENSIS_LVS_NEUTROPHIL_DN	196	20	89.8	DOCK3, SNAPC1, G0S2, TCAF2, C4orf3, NFKBIZ, ADAM17, SLC25A26, WDR54, RAB21, USP37, VEGFB, SERPINE1, CSRNP1, ERO1L, WDR45B, RELA, ZNF292, CYSTM1, IL18BP

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE37416_CTRL_VS_24H_F_TULARENSIS_LVS_NEUTROPHIL_UP	187	25	86.6	PDE4C, PRKAB2, CEACAM8, HRH4, BRF2, CDV3, GGCT, UBXN2A, ITGA4, PDK4, GABRB1, PARG, PDSS2, CLC, ADAMTS20, ZNRF2, KLF13, VPS13C, DBT, ZFP36L2, GAB3, SEMA3C, INPP4A, TCN1, PGLYRP1
GSE37416_0H_VS_24H_F_TULARENSIS_LVS_NEUTROPHIL_DN	196	26	86.7	ATP6V1C1, STAT4, OLR1, PDCD4, ABCF1, GBA3, CBFA2T2, PLAGL2, SLC25A26, VEGFB, ELL2, ERO1L, WDR45B, ZNF292, ATP6V1F, ADAMTS20, GPAA1, PHF23, CCDC101, USP28, ASCC1, HCST, LIF, HOOK3, MFAP1, HPCAL1
GSE3982_EOSINOPHIL_VS_MAC_UP	192	13	93.2	ARHGEF40, RSRP1, TNFRSF11B, MTMR3, HIPK2, LINC01565, LLPH, REPS2, ZNF250, ITGA10, METTL3, SLC44A1, HEYL
GSE3982_EOSINOPHIL_VS_NEUTROPHIL_UP	195	9	95.4	EPAS1, TCF25, EIF4B, RPL9, SMARCA4, LINC01565, C2orf47, TMEM131, RNASE2
GSE3982_EOSINOPHIL_VS_NKCELL_DN	197	16	91.9	PIP4K2B, EIF4B, AKR1C3, STAT4, MFAP5, TERF1, YIPF1, ANKRD17, GADD45GIP1, SMC2, NCAM1, PTK2B, NPY1R, PALLD, ZNF473, ABCD4
GSE3982_MAST_CELL_VS_MAC_DN	192	18	90.6	EPAS1, DOCK3, TNFSF14, MCM5, JUN, ZNF706, MPZL2, PSMD12, GIMAP5, QSER1, EIF5B, TGFBR1, PPP2R3A, DNAJB14, RAB17, ADK, CAPZA2, CLIP4
GSE3982_MAST_CELL_VS_BASOPHIL_DN	193	5	97.4	EPAS1, HBP1, RSRP1, HBZ, TNFRSF11B
GSE3982_MAST_CELL_VS_TH1_UP	198	22	88.9	HBP1, PDCD4, ZHX3, YIPF1, S100A14, NAG18, CTNND1, ACBD3, 42071, SLC10A3, IMPACT, ZNF337, TBC1D5, SOCS5, MORC3, SPI1, TIMP3, ZMYND8, TDP2, CRYGC, TRIM44, MYF5
GSE3982_MAST_CELL_VS_TH2_DN	196	17	91.3	EPAS1, SDC4, DDX51, FRYL, ORC6, CSNK2B, MPZL2, QSER1, TMEM156, TIMELESS, APOBEC3B, THRAP3, KLRC3, MYH10, LRCH3, ERF, MRPL42
GSE3982_DC_VS_EFF_MEMORY_CD4_TCELL_UP	199	15	92.5	APP, CXCL8, STXBP1, ZNF124, TACSTD2, PSMD12, PEA15, H1F0, VPS41, LGALS3, EHD4, MCUR1, DCSTAMP, CYP27B1, PREP
GSE3982_MAC_VS_BASOPHIL_DN	195	37	81.0	RSRP1, HBZ, STAR, LINC01565, THAP11, FAM124B, 42071, RPL5, LSM14A, SLC10A3, TUBD1, H1F0, REPS2, ZDHHC4, ELAVL3, FAM134C, JARID2, OSER1, PHTF2, KIAA0922, SIRPG, GZMA, RYK, ATP2B4, SH2D1A, SMURF1, C2orf68, CD55, RBM5, KRT7, MYB, ANAPC5, PDLIM3, TRADD, VEGFC, DICER1, MAPK14

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE3982_MAC_VS_EFF_MEMORY_CD4_TCELL_UP	198	23	88.4	DOCK3, IGFBP7, STXBP1, SLC2A5, EVI5, TRPV2, DBI, CKS2, BRCA1, H1F0, COL8A2, REPS2, PPP2R3A, IMPACT, SOCS5, S100A8, PICALM, OPA1, TIMM17A, TPM3, SLC12A8, PROC, GSPT1
GSE3982_MAC_VS_TH2_DN	197	26	86.8	DDX51, STAP2, PMS2P1, RMND5B, CD79B, RPL11, NAA35, POLR2I, ITGA10, SMARCD2, EPB41L4A, KLRC3, CLASP1, ZNF692, MEOX1, MED20, KIAA0922, EXOSC5, OXSM, CAPN10, CKS1B, ERCC2, RASSF1, GALNT3, ADGRA3, TMEM39B
GSE3982_BASOPHIL_VS_EFF_MEMORY_CD4_TCELL_UP	196	36	81.6	ROCK2, ACTN1, HBP1, CXCL8, BRD8, HRH4, PSMD12, DBI, CAB39, RPS27L, FAM124B, LRP12, DEPDC5, LMBRD1, H1F0, REPS2, IMPACT, SOCS5, SUOX, ABHD5, RAP2C, NDUFA4, JARID2, ARPC2, CR2, OSER1, NDFIP1, PHTF2, MTM1, ZMYND8, TLX2, RYK, SEMA3C, CD55, ZDHHC7, SHB
GSE39820_CTRL_VS_IL1B_IL6_CD4_TCELL_UP	197	15	92.4	IDI1, C9orf41, PDCD4, DNMT3B, DTWD2, ITGA4, KRIT1, HSPA8, ALG8, ELL2, SOCS5, CBX4, ANKRD44, GSTO1, C1orf112
GSE39820_CTRL_VS_TGFBETA3_IL6_CD4_TCELL_DN	197	32	83.8	SNAPC1, CTSB, BLVRB, GPR18, NAGK, PLSCR3, ARFGAP3, SMOX, YPEL3, CREB3L2, DHRS1, AFG3L2, METTL23, SERPINE1, CSRNP1, RAB2A, PICALM, STOML1, IFT43, PYCR1, NICN1, PHF20, SYT11, RNF181, ZFP36L2, TIMP1, FURIN, SMAD3, ECI2, SEC31A, RABGGTA, PRRC1
GSE5960_TH1_VS_ANERGIC_TH1_UP	198	17	91.4	MGMT, CHRM4, DEAF1, PSMA1, IRF9, TRIM37, ZNRD1, GNPNAT1, ACTL6B, SYPL1, SMR3A, NBN, TCIRG1, FCGR2B, CCL7, RPSA, C11orf31
GSE6269_HEALTHY_VS_E_COLI_INF_PBMC_DN	166	17	89.8	CXCL8, RNASE3, KLHL29, PLEK, KRT75, C10orf10, FAM124B, PRTN3, EPB41L1, HSPA13, H1F0, MPO, RAB20, PTGIR, RFX3, SUSD4, CORO1B
GSE6269_HEALTHY_VS_STREP_AUREUS_INF_PBMC_DN	167	29	82.6	CTSB, CSF2RA, FCN1, PLEK, BLVRB, 42065, RNASE2, STEAP3, TCEB3, H1F0, LGALS3, RAB20, S100A8, FLVCR2, SPI1, ELF4, EGR1, SLC15A3, ARF4, DSE, TIMP1, TOLLIP, ZFAND5, STAB1, PRRG4, MICAL2, PLBD1, NPC2, RHEB
GSE6269_FLU_VS_STREP_PNEUMO_INF_PBMC_DN	173	22	87.3	MS4A3, CD24, EIF4B, OMP, NAP1L1, AMELY, TACSTD2, RBMXL2, STEAP3, ST8SIA3, REPS2, TCEB2, HEYL, BCAS3, SNX15, GPAA1, MLN, LTF, SERINC2, GJC2, CES2, FXR1
GSE6269_E_COLI_VS_STREP_AUREUS_INF_PBMC_DN	172	11	93.6	YIPF3, RNASE3, IDS, TM9SF1, CYP4F3, LGALS3, ERO1L, GPAA1, OSBPL2, CLEC5A, SLC15A3

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE6269_E_COLI_VS_STREP_PNEUMO_INF_PBMC_DN	160	22	86.3	YIPF3, CEACAM8, SSR1, CD24, RNASE3, SLC2A5, CAMP, OLR1, TM9SF1, TACSTD2, CYP4F3, ZNF318, PRTN3, GSR, EVA1B, ALOX5, ABHD5, ERO1L, PICALM, GPAA1, LTF, CLEC5A
GSE7460_TCONV_VS_TREG_LN_DN	193	22	88.6	F2R, AHR, SOAT1, KLF6, MTMR3, ABCA5, GPR15, NFKBIZ, CCR6, TGFBR1, SOCS5, CTLA4, GLRX, CCR8, IGF1R, PHTF2, PEAK1, SESN1, CEP290, TNFRSF9
GSE7460_CTRL_VS_TGFB_TREATED_ACT_CD8_TCELL_UP	199	15	92.5	TNFRSF11B, SIRT3, TRPV2, ZNF512, TMEM121, HEXIM2, RIPK3, GNPNAT1, SERPINA5, COX6A2, DPH2, MTM1, DCDC2, ADK, OIP5
GSE7764_IL15_TREATED_VS_CTRL_NK_CELL_24H_DN	198	11	94.4	EPAS1, YAP1, KLHL29, KANSL1L, G0S2, AKAP5, CEP164, ITGA4, COX6A2, TSKS, NPY1R
GSE7764_IL15_NK_CELL_24H_VS_SPLENOCYTE_UP	198	13	93.4	GATSL3, ENTPD7, DDX56, WDR18, RGS16, SIRT3, SPTLC1, GART, KRTCAP2, GRWD1, SLCO4A1, TBC1D7, EGR2
GSE7764_IL15_NK_CELL_24H_VS_SPLENOCYTE_DN	198	20	89.9	ACTN1, HIPK1, TMEM131, IRF9, ZCCHC6, ARHGAP30, UVSSA, YPEL3, RSU1, TPCN1, ITGA4, EVA1B, APOBEC3B, PICALM, TSPAN32, CD72, STK38, NINL, ARHGEF12, TG
GSE7852_TREG_VS_TCONV_FAT_UP	198	16	91.9	IL18, PRKAB2, RGS16, ZC2HC1A, NFKBIZ, PQLC1, EHD4, KDSR, SORBS1, GLRX, BCL3, SLC52A3, MTMR10, PLEC, HERPUD1, SESN1
GSE7852_LN_VS_FAT_TREG_DN	195	28	85.6	IDI1, TBX15, GRK5, ITGA1, RAB20, ELL2, MYO3A, MFSD6, GLRX, HOXD9, SLC52A3, RAP1GAP2, PLEC, PRKAR2B, IRG1, SELM, RGS5, RNF125, RHOB, EYA2, CDH17, ATP2A2, CCR4, SMAD3, CSTB, AHNAK, BRD2, SLC41A2
GSE7852_THYMUS_VS_FAT_TREG_DN	197	27	86.3	PARD3, PRKAB2, NR1D2, FAM177A1, ZDHHC2, GOT1, YAF2, CTNND1, PQLC1, CLDN12, KLF16, IMPACT, ELL2, APOBEC3B, KDSR, SORBS1, EPB41L4A, ZNF703, RAP1GAP2, MRS2, SPACA1, PLEC, ADRB2, CCR3, HERPUD1, FOXL1, IRG1
GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_1MONTH_POST_DX_UP	200	19	90.5	RPN1, WDR18, STAU1, JUN, IDS, NDUFV1, ACTR2, FAM46C, PRIM2, GSR, LSM14A, DNAJB14, LYPLA1, ENSA, ZNF551, GPAA1, SNTB2, ATP5B, NUS1P3
GSE9006_1MONTH_VS_4MONTH_POST_TYPE_1_DIABETES_DX_PBMC_DN	193	16	91.7	ENTPD7, RGS16, CYP2B6, EPHA3, CD58, GART, GUCY2D, TFCP2, FGF22, RHD, FAP, TPT1P8, GSTO1, MBTPS2, CD72, C4BPA
GSE9650_NAIVE_VS_EXHAUSTED_CD8_TCELL_DN	196	22	88.8	YAP1, HTR2C, F2R, ENTPD7, FAM213A, RGS16, TJP1, TERF1, GNPTAB, TLN1, ATP5J2, TBX15, XCR1, HINFP, OSBPL9, CTLA4, SCN7A, EPHB6, SYPL1, ACADVL, EFHD2, IMMT

Gene set name	Gene set size	Core pathway size	Percent reduction	Core pathway member
GSE9650_NAIVE_VS_MEMORY_CD8_TCELL_UP	197	12	93.9	ACTN1, CLK4, DDX6, FARSB, EML5, ULK1, KMT2A, IRF9, IFIT1B, EGR2, TSPAN32, SLC44A1
GSE9650_EFFECTOR_VS_MEMORY_CD8_TCELL_UP	195	25	87.2	TTR, CAPNS1, FHL2, DBI, KIAA0101, USO1, GSTT1, PRIM2, GDAP2, ITGA4, LGALS3, DHRS1, TSPO, E2F8, GLRX, PPIB, DAPK2, BUB1, GZMA, TACC3, IDH3A, TNFRSF9, CKS1B, STAB1, MRPS17
GSE9650_EXHAUSTED_VS_MEMORY_CD8_TCELL_DN	198	26	86.9	GABARAPL2, RPN1, FARSB, SLC25A51, AP3M1, KLF6, SPTLC1, YIPF1, EIF2S1, ADD1, BLVRA, HMP19, SWI5, PIGU, FCGR2B, TTC7B, STK38, ADRB2, TAF11, C11orf31, SGK1, API5, TDRP, IL6R, PIK3CD, ANXA6
GSE9988_ANTI_TREM1_VS_LOW_LPS_MONOCYTE_UP	192	15	92.2	PHF13, NPM1, TNFSF14, MGAT1, ACSL3, ZNF318, SLCO4A1, TGFBR1, LGALS3, TBC1D7, TRIB1, PHF23, LIMK1, PPAP2B, SGK1
GSE9988_ANTI_TREM1_VS_ANTI_TREM1_AND_LPS_MONOCYTE_DN	182	15	91.8	IL18, LINC00936, SDC4, CXCL8, PLEK, KLF6, DYRK3, MFSD2A, NIPA1, GIMAP5, G0S2, IL10, NFKBIZ, TP53BP2, CSRNP1
GSE9988_LPS_VS_CTRL_TREATED_MONOCYTE_UP	182	25	86.3	IL18, LINC00936, CXCL8, FAM177A1, PLEK, DYRK3, MFSD2A, G0S2, IL10, DDX5, KRTAP5-8, NFKBIZ, TP53BP2, RAB21, ELL2, ADTRP, CSRNP1, RAP2C, TXN, TWISTNB, IL1B, BTG3, EGR1, STAT5A, IRG1
GSE9988_LOW_LPS_VS_CTRL_TREATED_MONOCYTE_UP	184	34	81.5	IL18, LINC00936, OTUD1, SDC4, CXCL8, PLEK, EGR3, DYRK3, JUN, MFSD2A, G0S2, IL10, DDX5, NFKBIZ, TP53BP2, RAB21, TRIP10, ELL2, ADTRP, CSRNP1, TXN, NBN, TWISTNB, IL1B, BTG3, EGR1, STAT5A, IRG1, ARHGEF2, B3GNT2, PTGS2, IL1A, DENND5A, MESDC1
GSE9988_LOW_LPS_VS_VEHICLE_TREATED_MONOCYTE_UP	183	15	91.8	IL18, LINC00936, OTUD1, SDC4, CXCL8, PLEK, STAT4, DYRK3, JUN, MFSD2A, G0S2, IL10, NFKBIZ, TP53BP2, ADTRP
GSE9988_ANTI_TREM1_AND_LPS_VS_VEHICLE_TREATED_MONOCYTES_UP	180	23	87.2	IL18, LINC00936, SNAPC1, CXCL8, PLEK, DYRK3, JUN, MFSD2A, ACSL3, IL10, DOLK, MUCL1, NFKBIZ, TP53BP2, SLCO4A1, CKS2, ADTRP, CSRNP1, DCSTAMP, TBC1D7, CCNA1, TXN, IL1B

Chapter 6

Discussion

Early and accurate diagnosis of diseases is essential for appropriate treatment of patients. Complex diseases result from collective action of multiple genetic and non-genetic factors. The technology of DNA microarray analysis provides massive information on transcription activities of all genes simultaneously (Gu et al, 2002). The genetic variations and regulations that influence predisposition and risk for wide range of complex conditions and contribute to complex disease can be evaluated by leveraging a combination of methods available for high throughput data including gene expression analysis. Many statistical methods have been developed to tackle challenges inherent to high throughput data. Our proposed work addresses an important methodological gap in the analysis of data measured by DNA microarray technology by analyzing the outcomes as continuous measurements, incorporating correlations across gene expressions in a gene set, and identifying core genes within a set.

Our gene set reduction method is an extension of GSA self-contained method from binary to continuous phenotypes. We developed the LCT-GSR based on two computationally efficient and powerful methods, SAM and LCT on the ground of self-contained hypothesis. By using self-contained methods we acknowledge that genes are not independent and consider the coordination and network among genes specially those that share biological pathways.

An important limitation of the self-contained approaches is that only a few genes, even one gene, can drive the association between the gene set and the phenotype. In such cases, post-hoc

analysis can be useful to extract significant subset associated with the phenotype. LCT-GSR is a simple analytical tool to reduce gene sets that have been found associated with the phenotype to smaller core sets, by gradually exploring the association of remaining genes as a set with the phenotype. The analyst can choose multiple cut-offs as a stopping rule, moving from more conservative to more liberal values allowing for a flexible reduction process. Scientists can focus on biological interpretation of the reduced sets instead of the whole sets.

We selected the LCT approach among the other GSA methods to identify significant gene sets. The LCT method efficiently incorporates correlations among the genes in a set into the test statistic while the other methods do not have this feature. Incorporating the covariance matrix into the test statistic and using permutation test results in better power (Dinu et al., 2013). The covariance matrix is singular when genes in a set are larger than the sample size and this is a common situation in microarray studies. Shrinkage covariance matrix estimator can deal with this problem but the computational cost of this approach is high. Orthogonal transformation of the gene expression is used to make this approach computationally efficient. As a result, the eigenvalue decomposition of the shrinkage covariance matrix is performed only once for the real gene expression data and there is no need to estimated it for each permuted datum.

6.1 Applications to real microarray data

Our method identified pathways and genes that were previously identified to be associated with the tumor volume as well as new markers that need to be further validated. For example, *Malic Enzyme 3*, a gene known to have an important role in cancer cell proliferation (Zheng FJ, et al., 2012), appears most frequently in the 4 core subsets (p-value<0.01, FDR=0.42). The elevated activity of *Transketolase* (p-value=0.02, FDR=0.60) facilitates tumors' accelerated proliferation (Phan et al., 2014). In particular the thiamine-dependent enzyme *Transketolase* is

essential for cancer cells to synthesize large amounts of nucleic acids needed for rapid cellular growth (Zastre et al., 2013). We were able to identify important genes that were not identified by SAM analysis. *Pyruvate Kinase, Muscle* has significant role in tumor volume reduction. This gene extracted from two significant gene sets while the result from SAM analysis showed marginally significant association (p-value=0.06, FDR=0.88). We found far many more important genes that their role in prostate cancer progression needs to be further investigated.

We identified many important genes from the significant gene sets associated with variation in birth weight. Understanding biological function of these genes provides useful information on underlying mechanism of birth weight and their links to other diseases.

Leptin (LEP) is identified to be associated with birth weight in both gene set databases (p-value=0.003, FDR=0.02). *Leptin* encodes a protein, which acts through the leptin receptor that is secreted by white adipocytes, and which plays a major role in the regulation of body weight. This protein is involved in the regulation of immune and inflammatory responses, angiogenesis and wound healing. Mutations in this gene and/or its regulatory regions cause severe obesity, and morbid obesity with hypogonadism. This gene has also been linked to type 2 diabetes mellitus development (Genecards).

Early growth response 3 (EGR3) is another core gene (p-value=0.001, FDR=0.01) that plays a role in a wide variety of processes including muscle development, lymphocyte development, endothelial cell growth and migration, and neuronal development (Genecards).

6.2 Strengths

The main strength of our gene set reduction approach is integration of the biological information in the construction of the pathways. Identifying the core subsets of significant gene

sets for a continuous phenotype has many advantages. It will improve extracting biological information efficiently from extremely noisy microarray data by interpreting only differentially expressed core sets. There are situations in which genes show no or weak signals at an individual gene analysis, but coordinating with other genes within a pathway they show very strong signals. For example, *Par-3 Family Cell Polarity Regulator* with the SAM p-value 1.0 in the prostate cancer data was identified in the core subset associated with the tumor volume. The method is powerful in detecting biomarkers of complex diseases because it considers biological networks between genes.

Reducing significant gene sets to smaller sets can reduce costs of disease diagnosis and treatment by focusing on smaller number of genes in screening massive databases for association with a continuous phenotype. Examination of redundant genes' expression levels increases unnecessary costs without a significant improvement in clinical decisions. Reduction to the most predictive genes is crucial in advancing our understanding of issues such as disease prevention, faster and more efficient diagnosis, intervention strategies and tailored treatment. Reduction to the most predictive genes can lead to a change of platform from high-dimensional microarray technology to alternative methods, such as real time polymerase chain reaction (PCR) assays that are cheaper and faster. This alternative method is easily applicable to routine clinical setting for diagnosis purposes (West et al., 2006; Pittman et al., 2004; Ein-Doret al., 2006).

The methodological approach to gene set reduction for continuous phenotypes can be applied to a wide range of common situations in which dichotomizing the continuous phenotype is neither easy nor meaningful. The variable may not be informative about the disease mechanism after categorization based on arbitrary or less meaningful cut-off values. Researchers will be able to identify biologically meaningful genes associated with continuous phenotypes of interest by

screening massive databases, and provide additional insights into disease progression, improved treatment strategies, and personalized medicine. These findings may detect novel biological mechanisms and will help formulate new hypotheses opening avenues for future research directions. A better understanding will provide insights into new approaches to screening and preventive interventions and possible targets for drug therapy. We hypothesized roles of gene expression variability and gene expression correlations with each other in the development of outcomes or diseases.

We were able to reduce the significant gene sets by 80% to 90% in the prostate cancer and CANDLE studies. These genes need to be further investigated by experts to comprehend underlying mechanism of prostate cancer prognosis and predicted biomarkers contributing to low birth weight.

6.3 Limitations

While we evaluated the performance of our method by applying it to two real microarray data sets, we were unable to examine its performance through simulation studies due to the complexity of data structure and correlations among them. The methodological development of our method is based on the SAM-GSR method which showed powerful performance in a simulation study (Dinu et al., 2008). Therefore, we are confident that the method is powerful in detecting set of core genes with biological networks for continuous phenotype. This is also supported by biological links to prostate cancer and birth weight.

Our method is based on a linear model, LCT, which is powerful but has its limitations. The LCT tests only linear associations between sets and a continuous phenotype. To check the linearity assumption, exploratory data analysis needs to be done. On the other hand, a small

number of samples can be a limitation to check for non-linearity. The LCT method can be extended to non-linear model if we can collect a large number of samples which in real situations will not be practical. We used the logarithmic transformation of the gene expression data and phenotypes to provide more support to linearity assumption.

6.4 Conclusions and Public Health implications

Complex diseases result from combined effect of multiple genetic and non-genetic factors. Identifying disease biomarkers helps scientists to advance understanding of the biological mechanism of a complex disease or traits through a pathway approach. We note that there is currently no consensus about the best statistical method to examine microarray gene expression data. Our proposed method in combination with biological validation of our findings can yield novel approaches to extract evidence. Knowledge generated from this research can be directly translated into practical clinical and public health applications. Identification of important genetic markers provides insights into efficient screening and preventive strategies, and opens avenues for cost effective personalized medicine. The R code for executing the LCT-GSR will be freely available to facilitate gene expression data analysis for various studies.

6.5 Future directions

In biomedical research, it is common to measure multiple outcomes per individual such as metabolite outcomes, or several protein measurements such as Phosphatase and tensin homolog (PTEN), PSA, Stathmin and Gleason score in prostate cancer studies.

PTEN is one of the most commonly mutated tumor suppressor genes in human prostate cancer. It controls a number of cellular processes, including survival, growth, proliferation, metabolism, migration, and cellular architecture (Ruscetti & Wu, 2013). Patients with prostate cancer who had PTEN mutation had also a significantly greater Gleason score, poorer prognosis,

and higher rate of metastasis. However, this mutation cannot predict the prognosis and the Gleason score is a more precise factor (Pourmand et al., 2007).

Prostate-specific antigen (PSA) is a protein made by prostate gland cells. The amount of PSA in the blood can be measured by a simple blood test. A PSA test may detect early prostate cancer in men who do not have symptoms.

Stathmin is the member of microtubule-destabilizing proteins that regulate the dynamics of microtubule polymerization and depolymerization. Stathmin is expressed at high levels in a variety of human cancers including prostate and provides an attractive molecule to target in cancer therapies that disrupt the mitotic apparatus. It may provide an effective approach for the treatment of prostate cancer (Mistry et al., 2005).

While evaluating the association of gene set expression measurements with each phenotype independently gives scientists insight into prostate cancer progression, evaluation of all these phenotypes together may broaden our understanding of prostate cancer prognosis and provide additional insight to personalized treatments.

One approach to evaluate the association of gene sets with outcome of interest characterized by multiple variables is to analyze each outcome independently. Using this approach, we ignore correlations between outcomes. The next step is to extend our method to multivariate continuous outcomes exhibiting correlations in order to take into account the correlations between outcomes as well as correlations between genes.

6.6 Software Packages

We used R software, version 3.0.3 under Windows 7 for executing the codes for the LCT and LCT-GSR. Free R code for performing LCT for continuous phenotype is available at http://www.ualberta.ca/_yyasui/homepage.html. SAS 9.3 is used to generate 0/1 matrix data using gene expression data and lists of gene sets from C2 and C7 catalog as well as stem cell signatures.

References

Adkins R. M., Tylavsky F. A., and Krushkal J. Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight. Chemistry and Biodiversity. 2012, 9(5):888-899.

Affymetrix, GeneChip Expression Analysis Technical Manual, 2000.

Apostolidou S., Abu-Amero S., O'Donoghue K., Frost J., Olafsdottir O., Chavele K.M., Whittaker J.C., Loughna P., Stanier P., Moore G.E. Elevated placental expression of the imprinted PHLDA2 gene is associated with low birth weight. J. Mol. Med. 2007, 85:379–387.

Ashley D.B., Diekmann J.E., Molenaar K.R. Risk Assessment and Allocation for Highway Construction Management; Federal Highway Administration, US Department of Transportation: Washington, DC, USA, 2006.

Baldi P., Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics. 2001,17:509–519.

Ball D.J., Watt J. Further thoughts on the utility of risk matrices. Risk Anal. 2013, 33, 2068–2078.

Barker DJ. Mother, babies, and health in later life. London, Churchill Livingstone, 1998, p. 13.

Barry WT., Nobel AB., Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005, 21(9):1943–9.

Basso O., Wilcox AJ., Weinberg CR. Birth weight and mortality: causality or confounding? Am J Epidemiol. 2006 Aug 15,164(4):303-11.

Bassols J; Prats-Puig A; Vázquez-Ruíz M; García-González MM; Martínez-Pascual M; Avellí P; Martínez-Martínez R; Fàbrega R; Colomer-Virosta C; Soriano-Rodríguez P; Díaz M; de Zegher F; Ibánez L; López-Bermejo A. Placental FTO expression relates to fetal growth.Int J Obes (Lond). 2010, 34(9):1365-70.

Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995, 57, 289–300.

Bill-Axelson A., Holmberg L., Ruutu M., Häggman M., et al. Radical prostatectomy versus watchful waiting in early prostate cancer. N Engl J Med. 2005, 352:1977-84.

Bill-Axelson A., Holmberg L., Filen F., Ruutu M., et al. For the Scandinavian Prostate Cancer Group Study Number 4: Radical Prostatectomy Versus Watchful Waiting in Localized Prostate Cancer: the Scandinavian Prostate Cancer Group-4 Randomized Trial. J Natl Cancer Inst. 2008, 100:1144-1154.

Chen JJ. Lee T et al. Significance analysis of groups of genes in expression profiling studies. Bioinformatics. 2007, 23: 2104-2112.

Chu G., Narasimham B., Tibshirani R., Tusher V. SAM "Significance Analysis of Microarrays", Users guide and technical document. Stanford University, 2002.

Cole DC., Mondloch MV., Hogg-Johnson S. Early Claimant Cohort Prognostic Modelling Group Listening to injured workers: how recovery expectations predict outcomes—a prospective study. CMAJ. 2002;166:749–754.

Collins J.W, David R.J, Handler A, et al. Very Low Birthweight in African American Infants: The Role of Maternal Exposure to Interpersonal Racial Discrimination, Am J Public Health. 2004 December; 94(12): 2132–2138.

Cox L.A. What's wrong with risk matrices? Risk Anal. 2008, 28, 497–512.

Delongchamp R, Lee T, Velasco C. A method for computing the overall statistical significance of a treatment effect among a group of genes. BMC Bioinformatics. 2006, 7 (Suppl. 2):S11.

DeRisi J., Penland L., Brown PO., et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat. Genet. 1996,14:457–460.

Dinu I., Potter JD., Mueller T., et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinformatics. 2007, 8:242.

Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulsky KS, Halloran PF, Yasui Y. Gene Set Analysis and Reduction. Briefings in Bioinformatics. 2008, 10(1): 24-34.

Dinu I., Wang X., Kelemen LE., Vatanpour S., Pyne S.Linear combination test for gene set analysis of a continuous phenotype. BMC Bioinformatics. 2013 Jul 1, 14:212.

Drinking Water Safety Plan Training Course. Alberta Environment and Sustainable Resource Development (AESRD). 2013. Available online: http://environment.gov.ab.ca/info/

library/8691.pdf (accessed on 8 June 2015).

Eckhardt F., Lewin J., Cortese R., Rakyan VK., et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006 Dec;38(12):1378-85.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1, 30(1):207-10.

Efron B., Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007, 1(1):107–29.

Ein-Dor L, Kela I, Getz G, et al. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics. 2005, 21(2):171-8.

Ein-Dor L., Yuk O., Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome of cancer. Proc Natl Acad Sci USA. 2006, 103(15):5923–28.

Feinberg AP., Irizarry RA., Fradin D., Aryee MJ., et al. Personalized epigenomic signatures that are stable over time and covary with body mass index. Sci Transl Med. 2010 Sep 15, 2(49):49ra67.

GeneCards: The human gene database. Available online: http://www.genecards.org.

Gibson G., Muse S. A Primer of Genomic Science, Sinhauer, 2001.

Gillman MW., Barker D., Bier D., Cagampang F., et al. Meeting report on the 3rd International Congress on Developmental Origins of Health and Disease (DOHaD). Pediatr Res. 2007 May, 61(5 Pt 1):625-9.

Goeman JJ., van de Geer SA., de Kort F., et al. A global test for groups of genes: testing association with clinical outcome. Bioinformatics. 2004, 20(1):93–9.

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

Gu CC, et al. Role of gene expression microarray analysis in finding complex disease genes. Genet. Epidemiol. 2002, 23:37–56.

Guide to Corporate Risk Profiles. Treasury Board Secretariat of Canada Website. 2013. Available online: http://www.tbs-sct.gc.ca/tbs-sct/rm-gr/guides/gcrp-gepropr-eng.asp?format=print (accessed on 8 June 2015).

Hrudey S.E. Limits to science for assessing and managing environmental health risks. Sci. Truth Justice. 2000, 2000, 127–150.

Hrudey S.E., Hrudey E.J. Safe Drinking Water—Lessons from Recent Outbreaks in Affluent Nations; IWA Publishing: London, UK, 2004.

Holt J., Leach A.W., Schrader G., Petter F., MacLeod A., Van Der Gaag D.J., Baker R.H., Mumford J.D. Eliciting and combining decision criteria using a limited palette of utility

functions and uncertainty distributions: Illustrated by application of pest risk analysis. Risk Anal. 2014, 34, 4–16.

Hubbard D., Evans D. Problems with scoring methods and ordinal scales in risk assessment. IBM J. Res. Dev. 2010, 54, 2:1–2:10.

Humphrey PA., Vollmer RT. Intraglandular tumor extent and prognosis in prostatic carcinoma: application of a grid method to prostatectomy specimens. Hum Pathol. 1990 Aug, 21(8):799-804.

ISO. Risk Management: Principles and Guidelines; ISO 31000: 2009; International Organization for Standardization (ISO): Geneva, Switzerland, 2009.

Johnson, R. A. and Wichern, D. W. Applied Multivariate Statistical Analysis, Prentice Hall, 2002.

Kanehisa M. and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000, 28, 27-30.

Kaplan S., Garrick B.J. On the quantitative definition of risk. Risk Anal. 1981, 1, 11–27.

Koike H., Ito K., Takezawa Y., Oyama T., et al. Insulin-like growth factor binding protein-6 inhibits prostate cancer cell proliferation: implication for anticancer effect of diethylstilbestrol in hormone refractory prostate cancer.Br J Cancer. 2005 Apr 25, 92(8):1538-44.

Koutsaki M., Sifakis S., Zaravinos A., Koutroulakis D., Koukoura O., Spandidos DA. Decreased placental expression of hPGH, IGF-I and IGFBP-1 in pregnancies complicated by fetal growth restriction. Growth Horm. IGF Res. 2011, 21, 31.

Lachmann A, Xu H, Krishnan J, Berger SI, et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010 Oct 1, 26(19):2438-44.

Levine E.S. Improving risk matrices: The advantages of logarithmically scaled axes. J. Risk Res. 2012, 15, 209–222.

Liberzon A., Subramanian A., Pinchback R., Thorvaldsdóttir H., et al. Molecular signature database (MSigDB) 3.0. Bioinformatics. 2011, 27 (12), 1739–1740.

Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014 Jul, 15(4):504-18.

Männik J., Vaas P., Rull K., Teesalu P., Rebane T., Laan M. Differential expression profile of growth hormone/chorionic somatomammotropin genes in placenta of small- and large-for-gestational-age newborns. Clin. Endocrinol. Metab. 2010, 95, 2433.

Mansmann U., Meister R. Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach. Methods Inf Med. 2005, 44:449–53.

McMinn J, Wei M, Sadovsky Y, Thaker HM, Tycko B., Imprinting of PEG1/MEST isoform 2 in human placenta. Placenta 2006, 27, 540.

McTernan CL., Draper N., Nicholson H., Chalder SM., et al. Reduced Placental 11-Hydroxysteroid Dehydrogenase Type 2 mRNA Levels in Human Pregnancies Complicated by Intrauterine Growth Restriction: An Analysis of Possible Mechanisms .J. Clin. Endocrinol. Metab. 2001, 86, 4979.

Mericq V., Medina P., Kakarieka E., Márquez L., Johnson MC., Iñiguez G. Differences in expression and activity of 11beta-hydroxysteroid dehydrogenase type 1 and 2 in human placentas of term pregnancies according to birth weight and gender. In˜ iguez, Eur. J. Endocrinol. 2009, 161, 419.

Mistry SJ., Bank A., Atweh GF. Targeting stathmin in prostate cancer. Mol Cancer Ther. December 2005 4, 1821.

Murthi P., Doherty VL., Said JM., Donath S., Brennecke SP., Kalionis B., Homeobox gene ESX1L expression is decreased in human pre-term idiopathic fetal growth restriction. Mol. Hum. Reprod. 2006, 12, 763.

Nam, D. and Kim, S.Y. Gene-set approach for expression pattern analysis. Brief Bioinformatics. 2008, 9, 189–197.

National Health and Medical Research Council (NHMRC). Australian Drinking Water Guidelines. 2013. Available online: http://www.nhmrc.gov.au/guidelines/publications/eh52 (accessed on 8 June 2015).

Newton MA., Quintana FA., den Boon JA. Random set methods identify distinct aspects of the enrichment signal in gene-set analysis. Ann Appl Stat. 2007, 1(1):85–106.

Nishimura D. BioCarta. Biotech Software & Internet Report. June 2001, 2(3): 117-120.

Novershtern N., Subramanian A., Lawton LN., Mak RH., et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011 Jan 21;144(2):296-309.

NPSA/NHS. A Risk Matrix for Risk Managers; National Patient Safety Administration, National Health Service (NPSA/NHS): London, UK, 2008.

Pickering A., Cowley S. Risk matrices: Implied accuracy and false assumptions. J. Health Safe Res. Pract. 2010, 2, 9–16.

Pittman J, Huang E, Dressman H, etal. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci USA. 2004, 101(22):8431–6.

Phan LM., Yeung SJ., Lee MH. Cancer metabolic reprogramming: importance, main features, and potentials for precise targeted anti-cancer therapies. Cancer Biol Med. 2014 Mar, 11(1): 1–19.

Pourmand G1, Ziaee AA, Abedi AR, Mehrsai A, Alavi HA, Ahmadi A, Saadati HR. Role of PTEN gene in progression of prostate cancer. Urol J. 2007 Spring, 4(2):95-100.

Public Health Agency of Canada. Transfusion Transmitted Diseases/Infections. 2007.

Available online: http://www.bridgelin.ca/bridgelin2/files/2576.pdf (accessed on 8 June 2015).

Prostate Cancer Canada, Available online: www.prostatecancer.ca (accessed on 15 December 2015)

Quintella M.C., Addas-Carvalho M., Da Silva M.G.C. Evaluation of the risk analysis technique in blood bank production processes. Chem. Eng. Trans. 2008, 13, 271–278.

Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., et al. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol. 2004, 2, e405.

Renn O. Concepts of Risk. In Social Theories of Risk; Krimsky, S., Golding, D., Eds.;

Praeger Publishing: Westport, CT, USA, 1992; pp. 52–79.

Rizak S., Hrudey S.E. Misinterpretation of drinking water quality monitoring data with implications for risk management. Environ. Sci. Technol. 2006, 40, 5244–5250.

Ruscetti, M., Hong, W., 2013. PTEN and Prostate Cancer, In D.J. Tindall's Book "Prostate Cancer: Biochemistry, Molecular Biology, and Genetics, Protein Reviews", (Chapter 4): 87–137.

Sboner A., Demichelis F., Calza S., Pawitan Y., et al. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. BMC Medical Genomics. 2010, 3:8.

Schafer, J. and K. Strimmer . A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statist. Appl. Genet. Mol. Biol. 2005, 4(32).

Schena M, et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl Acad. Sci. USA. 1996, 93:10614–10619.

Sheikh S., Satoskar P., Bhartiya D., Expression of insulin-like growth factor-I and placental growth hormone mRNA in placentae: a comparison between normal and intrauterine growth retardation pregnancies. Mol. Hum. Reprod. 2001, 7, 287.

Simon R., Korn E., McShane L., Radmacher M., et al. Design and analysis of DNA microarray investigations, Springer, 2003.

Song F., Smith JF., Kimura MT., Morrow AD., et al. Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A. 2005 Mar 1, 102(9):3336-41.

Storey, J.D., Tibshirani, R. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA. 2003, 100, 9440–9445.

Struwe E., Berzl GM., Schild RL., Beckmann MW., Dörr HG., Rascher W., Dötsch J. Simultaneously reduced gene expression of cortisol-activating and cortisol-inactivating enzymes in placentas of small-for-gestational-age neonates. Am. J. Obstet. Gynecol. 2007, 197, 43.

Struwe E., Berzl GM., Schild RL., Dötsch J., Gene expression of placental hormones regulating energy balance in small for gestational age neonates. Eur. J. Obstet. Gynecol. Reprod. Biol. 2009, 142, 38.

Subramanian A., Tamayo P., Mootha V. K., Mukherjee S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005,102 ,15545-50.

The Human Protein Atlas. Available online: http://www.proteinatlas.org.

Thomas JG., Olson JM., Tapscott SJ., Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 2001 Jul, 11(7):1227-36.

Tusher VG., Tibshirani R., Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl Acad. Sci. USA. 2001, 98:5116–5121.

Tzschoppe A., Struwe E., Blessing H., Fahlbusch F., et al. Placental 11beta-HSD2 gene expression at birth is inversely correlated with growth velocity in the first year of life after intrauterine growth restriction. Pediatr. Res. 2009, 65, 647.

Van Vliet G., Liu S., Kramer M. Decreasing Sex Difference in Birth Weight. Volume 20(4), July 2009, p 622.

West M, Ginsburg GS, Huang AT, et al. Embracing the complexity of genomic data for personalized medicine. Genome Res. 2006,16(5):559–66.

WHO. Rapid Risk Assessment of Acute Public Health Events; World Health Organization: Geneva, Switerland, 2012.

WHO. Chapter 4. In Guidelines for Drinking-Water Quality; World Health Organization: Geneva, Switzerland, 2011.

Wieland B., Dhollander S., Salman M., Koenen F. Qualitative risk assessment in a data-scarce environment: A model to assess the impact of control measures on spread of African Swine Fever. Prev. Vet. Med. 2011, 99, 4–14.

Zastre JA., Sweet RL., Hanberry BS., Ye S. Linking vitamin B1 with cancer cell metabolism. Cancer & Metabolism. 2013, 1:16

Zheng FJ., Ye HB., Wu MS., Lian YF., et al. Repressing malic enzyme 1 redirects glucose metabolism, unbalances the redox state, and attenuates migratory and invasive abilities in nasopharyngeal carcinoma cell lines. Chin J Cancer. 2012, 31(11): 519-531.

Appendix

Table A. Gene sets in stem cell signatures associated with birth weight phenotype based on the LCT analysis

Gene set name	Gene set size	p-value
IPA_affects differentiation of embryonic stem cells	41	0
StemCell_Kasper06_30genes_16880536-table1	30	0.001
DMAP_MEGA_UP	46	0.001
DMAP_MONO1_DN	47	0.001
DMAP_PRE_BCELL2_UP	44	0.001
DMAP_PRE_BCELL3_DN	44	0.001
StemCell_Lim08_50genes_18510698-Table1	47	0.002
Ben-Porath_MYC_TARGETS_WITH_EBOX	226	0.002
DB_ESR1-15608294	88	0.002
StemCell_Kocer08_87genes_18667080-TableS6	71	0.003
StemCell_Shim04_25genes_15246160-table6	22	0.003
StemCell_Fruehauf06_110genes_16863911-table1	97	0.003
DMAP_ERY_UP	45	0.003
DMAP_GM_EARLY_DN	42	0.003
DMAP_PRE_BCELL_UP	39	0.003
DMAP_BCELL_DN	44	0.003
DMAP_TCELLA6_DN	45	0.003
StemCell_Tondreau08_52genes_18405367-Table2b	41	0.004
DMAP_BCELLA2_UP	49	0.005
DMAP_TCELLA6_UP	44	0.005
IPA_affects differentiation of stem cells	72	0.006
DMAP_ERY4_DN	47	0.007
IPA_decreases differentiation of stem cells	18	0.007
StemCell_Colombo09_111genes_19123479-TableS1	92	0.008
StemCell_Lim08_25genes_18510698-Table2	25	0.008
DMAP_ERY_DN	46	0.008
DMAP_GM_EARLY_UP	40	0.008
DMAP_HSC1_DN	48	0.008

Gene set name	Gene set size	p-value
DMAP_HSC3_UP	48	0.008
DB_PPARG-19300518	194	0.008
StemCell_Bhattacharya05_2843genes_16207381-Table1Sa	312	0.01
DMAP_MONO2_DN	40	0.01
DMAP_TCELLA2_DN	47	0.01

Table B. Gene sets in C7 catalog associated with birth weight phenotype based on the LCT analysis

Gene set name	Gene set size	LCT p- value
GSE12845_NAIVE_VS_PRE_GC_TONSIL_BCELL_DN	197	0
GSE14308_INDUCED_VS_NATURAL_TREG_DN	197	0
GSE1448_CTRL_VS_ANTI_VBETA5_DP_THYMOCYTE_UP	196	0
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IFNAB_CD8_TCELL_DN	199	0
GSE17974_0H_VS_4H_IN_VITRO_ACT_CD4_TCELL_UP	182	0
GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_DN	195	0
GSE20366_EX_VIVO_VS_HOMEOSTATIC_CONVERSION_TREG_UP	197	0
GSE20366_EX_VIVO_VS_DEC205_CONVERSION_NAIVE_CD4_TCELL_UP	194	0
GSE20366_CD103_POS_VS_NEG_TREG_KLRG1NEG_UP	195	0
GSE22886_IGM_MEMORY_BCELL_VS_BM_PLASMA_CELL_DN	192	0
GSE22886_NEUTROPHIL_VS_DC_DN	200	0
GSE29618_BCELL_VS_MDC_UP	183	0
GSE30962_ACUTE_VS_CHRONIC_LCMV_PRIMARY_INF_CD8_TCELL_UP	194	0
GSE34205_HEALTHY_VS_RSV_INF_INFANT_PBMC_DN	200	0
GSE360_DC_VS_MAC_T_GONDII_DN	195	0
GSE3982_BASOPHIL_VS_EFF_MEMORY_CD4_TCELL_UP	196	0
GSE7460_TCONV_VS_TREG_LN_DN	193	0
GSE9988_LOW_LPS_VS_CTRL_TREATED_MONOCYTE_UP	184	0
GSE10239_NAIVE_VS_MEMORY_CD8_TCELL_UP	199	0.001
GSE10325_CD4_TCELL_VS_LUPUS_CD4_TCELL_UP	189	0.001
GSE10325_CD4_TCELL_VS_LUPUS_CD4_TCELL_DN	198	0.001
GSE11057_NAIVE_CD4_VS_PBMC_CD4_TCELL_DN	189	0.001
GSE13485_DAY1_VS_DAY21_YF17D_VACCINE_PBMC_DN	190	0.001
GSE1460_DP_THYMOCYTE_VS_NAIVE_CD4_TCELL_ADULT_BLOOD_UP	197	0.001
GSE15659_CD45RA_NEG_CD4_TCELL_VS_RESTING_TREG_UP	186	0.001
GSE17721_CTRL_VS_POLYIC_24H_BMDM_UP	200	0.001
GSE17721_POLYIC_VS_GARDIQUIMOD_24H_BMDM_UP	197	0.001
GSE20366_EX_VIVO_VS_HOMEOSTATIC_CONVERSION_NAIVE_CD4_TCELL_UP	196	0.001
GSE24081_CONTROLLER_VS_PROGRESSOR_HIV_SPECIFIC_CD8_TCELL_DN	190	0.001
GSE24634_NAIVE_CD4_TCELL_VS_DAY3_IL4_CONV_TREG_DN	198	0.001
GSE29618_BCELL_VS_MONOCYTE_DAY7_FLU_VACCINE_UP	185	0.001
GSE31082_DN_VS_DP_THYMOCYTE_DN	198	0.001
GSE36476_CTRL_VS_TSST_ACT_72H_MEMORY_CD4_TCELL_OLD_UP	195	0.001
GSE37416_0H_VS_24H_F_TULARENSIS_LVS_NEUTROPHIL_DN	196	0.001
GSE3982_EOSINOPHIL_VS_MAC_UP	192	0.001

Gene set name	Gene set size	LCT p- value
GSE6269_HEALTHY_VS_STREP_AUREUS_INF_PBMC_DN	167	0.001
KAECH_DAY15_EFF_VS_MEMORY_CD8_TCELL_UP	192	0.002
GSE10094_LCMV_VS_LISTERIA_IND_EFF_CD4_TCELL_UP	196	0.002
GSE12845_IGD_POS_BLOOD_VS_PRE_GC_TONSIL_BCELL_DN	199	0.002
GSE17580_TREG_VS_TEFF_S_MANSONI_INF_UP	196	0.002
GSE17721_CTRL_VS_PAM3CSK4_0.5H_BMDM_DN	195	0.002
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_4H_CD4_TCELL_UP	186	0.002
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_72H_CD4_TCELL_DN	197	0.002
GSE17974_IL4_AND_ANTI_IL12_VS_UNTREATED_12H_ACT_CD4_TCELL_UP	187	0.002
GSE17974_1.5H_VS_72H_IL4_AND_ANTI_IL12_ACT_CD4_TCELL_DN	194	0.002
GSE22886_NAIVE_TCELL_VS_NKCELL_DN	197	0.002
GSE22886_NAIVE_CD8_TCELL_VS_NKCELL_DN	196	0.002
GSE26669_CTRL_VS_COSTIM_BLOCK_MLR_CD4_TCELL_DN	195	0.002
GSE27786_LSK_VS_BCELL_UP	197	0.002
GSE2826_WT_VS_XID_BCELL_DN	198	0.002
GSE29618_BCELL_VS_MDC_DAY7_FLU_VACCINE_UP	182	0.002
GSE29618_MONOCYTE_VS_MDC_DAY7_FLU_VACCINE_UP	200	0.002
GSE32423_CTRL_VS_IL4_MEMORY_CD8_TCELL_UP	196	0.002
GSE360_T_GONDII_VS_B_MALAYI_HIGH_DOSE_DC_DN	198	0.002
GSE360_HIGH_VS_LOW_DOSE_B_MALAYI_DC_DN	194	0.002
GSE36476_CTRL_VS_TSST_ACT_16H_MEMORY_CD4_TCELL_OLD_UP	196	0.002
GSE3982_MAST_CELL_VS_TH1_UP	198	0.002
GSE3982_MAC_VS_BASOPHIL_DN	195	0.002
GSE3982_MAC_VS_EFF_MEMORY_CD4_TCELL_UP	198	0.002
GSE7852_TREG_VS_TCONV_FAT_UP	198	0.002
GSE11864_UNTREATED_VS_CSF1_PAM3CYS_IN_MAC_DN	185	0.003
GSE15659_NAIVE_VS_PTPRC_NEG_CD4_TCELL_DN	193	0.003
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	199	0.003
GSE17721_POLYIC_VS_PAM3CSK4_4H_BMDM_DN	190	0.003
GSE17721_LPS_VS_PAM3CSK4_12H_BMDM_DN	195	0.003
GSE22886_NAIVE_CD8_TCELL_VS_MEMORY_TCELL_DN	198	0.003
GSE22886_IGG_IGA_MEMORY_BCELL_VS_BM_PLASMA_CELL_DN	189	0.003
GSE22886_NAIVE_TCELL_VS_DC_DN	200	0.003
GSE2706_R848_VS_R848_AND_LPS_8H_STIM_DC_UP	178	0.003
GSE3982_EOSINOPHIL_VS_NKCELL_DN	197	0.003
GSE39820_CTRL_VS_TGFBETA3_IL6_CD4_TCELL_DN	197	0.003
GSE6269_HEALTHY_VS_E_COLI_INF_PBMC_DN	166	0.003

Gene set name	Gene set size	LCT p- value
GSE6269_E_COLI_VS_STREP_AUREUS_INF_PBMC_DN	172	0.003
GSE9650_NAIVE_VS_EXHAUSTED_CD8_TCELL_DN	196	0.003
GSE9988_LOW_LPS_VS_VEHICLE_TREATED_MONOCYTE_UP	183	0.003
GOLDRATH_EFF_VS_MEMORY_CD8_TCELL_UP	197	0.004
GSE15659_CD45RA_NEG_CD4_TCELL_VS_ACTIVATED_TREG_DN	194	0.004
GSE17721_CTRL_VS_POLYIC_6H_BMDM_UP	195	0.004
GSE17721_0.5H_VS_24H_POLYIC_BMDM_DN	197	0.004
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_1H_CD4_TCELL_UP	178	0.004
GSE22886_IGA_VS_IGM_MEMORY_BCELL_DN	196	0.004
GSE22886_DAY1_VS_DAY7_MONOCYTE_IN_CULTURE_DN	198	0.004
GSE26928_CENTR_MEMORY_VS_CXCR5_POS_CD4_TCELL_DN	180	0.004
GSE2826_WT_VS_BTK_KO_BCELL_DN	199	0.004
GSE29618_BCELL_VS_MONOCYTE_UP	179	0.004
GSE360_CTRL_VS_L_MAJOR_MAC_DN	195	0.004
GSE360_LOW_DOSE_B_MALAYI_VS_M_TUBERCULOSIS_DC_UP	199	0.004
GSE360_L_DONOVANI_VS_B_MALAYI_HIGH_DOSE_MAC_UP	196	0.004
GSE37416_CTRL_VS_12H_F_TULARENSIS_LVS_NEUTROPHIL_DN	196	0.004
GSE7460_CTRL_VS_TGFB_TREATED_ACT_CD8_TCELL_UP	199	0.004
GSE9006_1MONTH_VS_4MONTH_POST_TYPE_1_DIABETES_DX_PBMC_DN	193	0.004
GSE10239_NAIVE_VS_KLRG1INT_EFF_CD8_TCELL_DN	197	0.005
GSE11864_UNTREATED_VS_CSF1_IN_MAC_UP	191	0.005
GSE17721_CTRL_VS_GARDIQUIMOD_12H_BMDM_UP	198	0.005
GSE17721_LPS_VS_POLYIC_24H_BMDM_UP	195	0.005
GSE17721_PAM3CSK4_VS_CPG_1H_BMDM_DN	196	0.005
GSE17721_PAM3CSK4_VS_CPG_4H_BMDM_UP	196	0.005
GSE17721_LPS_VS_GARDIQUIMOD_24H_BMDM_DN	196	0.005
GSE22886_CD8_TCELL_VS_BCELL_NAIVE_UP	197	0.005
GSE22886_UNSTIM_VS_IL15_STIM_NKCELL_DN	198	0.005
GSE26669_CTRL_VS_COSTIM_BLOCK_MLR_CD8_TCELL_DN	199	0.005
GSE27786_CD4_TCELL_VS_NKTCELL_DN	199	0.005
GSE31082_DP_VS_CD4_SP_THYMOCYTE_DN	193	0.005
GSE3337_4H_VS_16H_IFNG_IN_CD8POS_DC_UP	196	0.005
GSE3982_DC_VS_EFF_MEMORY_CD4_TCELL_UP	199	0.005
GSE39820_CTRL_VS_IL1B_IL6_CD4_TCELL_UP	197	0.005
GSE6269_FLU_VS_STREP_PNEUMO_INF_PBMC_DN	173	0.005
GSE9988_LPS_VS_CTRL_TREATED_MONOCYTE_UP	182	0.005
GSE9988_ANTI_TREM1_AND_LPS_VS_VEHICLE_TREATED_MONOCYTES_UP	180	0.005

Gene set name	Gene set size	LCT p- value
GSE10239_NAIVE_VS_KLRG1HIGH_EFF_CD8_TCELL_UP	195	0.006
GSE11057_PBMC_VS_MEM_CD4_TCELL_UP	189	0.006
GSE13738_TCR_VS_BYSTANDER_ACTIVATED_CD4_TCELL_DN	182	0.006
GSE1460_INTRATHYMIC_T_PROGENITOR_VS_THYMIC_STROMAL_CELL_UP	197	0.006
GSE17721_CTRL_VS_PAM3CSK4_8H_BMDM_UP	199	0.006
GSE17721_PAM3CSK4_VS_GADIQUIMOD_4H_BMDM_UP	197	0.006
GSE17721_0.5H_VS_24H_GARDIQUIMOD_BMDM_DN	196	0.006
GSE17974_IL4_AND_ANTI_IL12_VS_UNTREATED_48H_ACT_CD4_TCELL_UP	186	0.006
GSE25087_TREG_VS_TCONV_ADULT_DN	185	0.006
GSE29618_PRE_VS_DAY7_POST_LAIV_FLU_VACCINE_MONOCYTE_UP	194	0.006
GSE30083_SP1_VS_SP3_THYMOCYTE_DN	197	0.006
GSE30083_SP1_VS_SP4_THYMOCYTE_DN	196	0.006
GSE30962_PRIMARY_VS_SECONDARY_ACUTE_LCMV_INF_CD8_TCELL_DN	196	0.006
GSE34205_RSV_VS_FLU_INF_INFANT_PBMC_UP	177	0.006
GSE360_L_MAJOR_VS_T_GONDII_MAC_UP	192	0.006
GSE36392_EOSINOPHIL_VS_MAC_IL25_TREATED_LUNG_DN	196	0.006
GSE3982_EOSINOPHIL_VS_NEUTROPHIL_UP	195	0.006
GSE3982_MAST_CELL_VS_MAC_DN	192	0.006
GSE3982_MAST_CELL_VS_TH2_DN	196	0.006
GSE7764_IL15_NK_CELL_24H_VS_SPLENOCYTE_DN	198	0.006
GSE9650_NAIVE_VS_MEMORY_CD8_TCELL_UP	197	0.006
GSE9650_EXHAUSTED_VS_MEMORY_CD8_TCELL_DN	198	0.006
GSE13306_TREG_VS_TCONV_SPLEEN_DN	196	0.007
GSE14308_TH2_VS_INDUCED_TREG_UP	194	0.007
GSE1448_ANTI_VALPHA2_VS_VBETA5_DP_THYMOCYTE_UP	196	0.007
GSE1460_DP_THYMOCYTE_VS_THYMIC_STROMAL_CELL_DN	197	0.007
GSE16522_MEMORY_VS_NAIVE_CD8_TCELL_DN	195	0.007
GSE17721_CTRL_VS_POLYIC_1H_BMDM_UP	197	0.007
GSE17721_CTRL_VS_GARDIQUIMOD_0.5H_BMDM_UP	198	0.007
GSE17721_LPS_VS_CPG_1H_BMDM_UP	198	0.007
GSE17721_4_VS_24H_GARDIQUIMOD_BMDM_UP	198	0.007
GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_UP	185	0.007
GSE20715_0H_VS_6H_OZONE_TLR4_KO_LUNG_DN	199	0.007
GSE22045_TREG_VS_TCONV_UP	179	0.007
GSE22886_NAIVE_VS_IGG_IGA_MEMORY_BCELL_DN	192	0.007
GSE22886_NAIVE_CD4_TCELL_VS_DC_DN	198	0.007
GSE26669_CD4_VS_CD8_TCELL_IN_MLR_COSTIM_BLOCK_DN	196	0.007

Gene set name	Gene set size	LCT p- value
GSE27786_NKTCELL_VS_ERYTHROBLAST_UP	199	0.007
GSE29614_CTRL_VS_DAY7_TIV_FLU_VACCINE_PBMC_UP	170	0.007
GSE34205_HEALTHY_VS_FLU_INF_INFANT_PBMC_DN	199	0.007
GSE3982_MAST_CELL_VS_BASOPHIL_DN	193	0.007
GSE7764_IL15_NK_CELL_24H_VS_SPLENOCYTE_UP	198	0.007
GSE7852_LN_VS_FAT_TREG_DN	195	0.007
GSE9988_ANTI_TREM1_VS_LOW_LPS_MONOCYTE_UP	192	0.007
KAECH_NAIVE_VS_DAY8_EFF_CD8_TCELL_UP	198	0.008
GSE11864_CSF1_IFNG_VS_CSF1_IFNG_PAM3CYS_IN_MAC_DN	184	0.008
GSE14000_4H_VS_16H_LPS_DC_TRANSLATED_RNA_DN	194	0.008
GSE17721_CPG_VS_GARDIQUIMOD_16H_BMDM_UP	198	0.008
GSE17721_0.5H_VS_4H_LPS_BMDM_UP	199	0.008
GSE24634_TREG_VS_TCONV_POST_DAY3_IL4_CONVERSION_DN	199	0.008
GSE24634_TEFF_VS_TCONV_DAY7_IN_CULTURE_UP	195	0.008
GSE27786_LSK_VS_ERYTHROBLAST_UP	198	0.008
GSE27786_LIN_NEG_VS_BCELL_UP	197	0.008
GSE29618_MONOCYTE_VS_MDC_UP	200	0.008
GSE30083_SP3_VS_SP4_THYMOCYTE_DN	193	0.008
GSE339_CD4POS_VS_CD8POS_DC_UP	194	0.008
GSE360_DC_VS_MAC_M_TUBERCULOSIS_DN	195	0.008
GSE360_HIGH_DOSE_B_MALAYI_VS_M_TUBERCULOSIS_MAC_DN	195	0.008
GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_UP	188	0.008
GSE5960_TH1_VS_ANERGIC_TH1_UP	198	0.008
GSE6269_E_COLI_VS_STREP_PNEUMO_INF_PBMC_DN	160	0.008
GSE7764_IL15_TREATED_VS_CTRL_NK_CELL_24H_DN	198	0.008
GSE7852_THYMUS_VS_FAT_TREG_DN	197	0.008
GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_1MONTH_POST_DX_UP	200	0.008
GSE9988_ANTI_TREM1_VS_ANTI_TREM1_AND_LPS_MONOCYTE_DN	182	0.008
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_BCELL_UP	195	0.009
GSE15659_NAIVE_CD4_TCELL_VS_ACTIVATED_TREG_DN	195	0.009
GSE15659_RESTING_TREG_VS_NONSUPPRESSIVE_TCELL_DN	193	0.009
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_IFNAB_CD8_TCELL_UP	197	0.009
GSE15930_NAIVE_VS_72H_IN_VITRO_STIM_TRICHOSTATINA_CD8_TCELL_DN	198	0.009
GSE17721_CTRL_VS_CPG_1H_BMDM_DN	199	0.009
GSE17721_PAM3CSK4_VS_GADIQUIMOD_6H_BMDM_DN	198	0.009
GSE17721_LPS_VS_CPG_4H_BMDM_UP	199	0.009
GSE17721_0.5H_VS_12H_PAM3CSK4_BMDM_UP	199	0.009

Gene set name	Gene set size	LCT p- value
GSE17721_0.5H_VS_8H_PAM3CSK4_BMDM_UP	198	0.009
GSE17974_0H_VS_0.5H_IN_VITRO_ACT_CD4_TCELL_UP	176	0.009
GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_18H_UP	192	0.009
GSE22886_IGM_MEMORY_BCELL_VS_BM_PLASMA_CELL_UP	197	0.009
GSE29618_BCELL_VS_PDC_UP	186	0.009
GSE30083_SP2_VS_SP3_THYMOCYTE_DN	195	0.009
GSE360_T_GONDII_VS_B_MALAYI_HIGH_DOSE_MAC_UP	195	0.009
GSE37416_CTRL_VS_3H_F_TULARENSIS_LVS_NEUTROPHIL_UP	184	0.009
GSE37416_CTRL_VS_24H_F_TULARENSIS_LVS_NEUTROPHIL_UP	187	0.009
GSE3982_MAC_VS_TH2_DN	197	0.009
GSE9650_EFFECTOR_VS_MEMORY_CD8_TCELL_UP	195	0.009
KAECH_NAIVE_VS_DAY8_EFF_CD8_TCELL_DN	194	0.01
GSE12845_IGD_NEG_BLOOD_VS_NAIVE_TONSIL_BCELL_UP	195	0.01
GSE13411_NAIVE_BCELL_VS_PLASMA_CELL_UP	193	0.01
GSE13484_UNSTIM_VS_3H_YF17D_VACCINE_STIM_PBMC_DN	193	0.01
GSE13484_12H_UNSTIM_VS_YF17D_VACCINE_STIM_PBMC_UP	197	0.01
GSE13484_12H_VS_3H_YF17D_VACCINE_STIM_PBMC_UP	194	0.01
GSE13485_CTRL_VS_DAY7_YF17D_VACCINE_PBMC_UP	172	0.01
GSE15750_WT_VS_TRAF6KO_DAY10_EFF_CD8_TCELL_UP	198	0.01
GSE16522_ANTI_CD3CD28_STIM_VS_UNSTIM_NAIVE_CD8_TCELL_DN	199	0.01
GSE17974_0H_VS_4H_IN_VITRO_ACT_CD4_TCELL_DN	192	0.01
GSE20366_EX_VIVO_VS_DEC205_CONVERSION_UP	197	0.01
GSE22886_DAY0_VS_DAY7_MONOCYTE_IN_CULTURE_DN	200	0.01
GSE24634_NAIVE_CD4_TCELL_VS_DAY10_IL4_CONV_TREG_DN	199	0.01
GSE29618_BCELL_VS_MONOCYTE_DN	200	0.01
GSE32423_IL7_VS_IL4_MEMORY_CD8_TCELL_UP	197	0.01
GSE360_DC_VS_MAC_B_MALAYI_LOW_DOSE_DN	200	0.01

Table C. Frequency of the genes within core pathway of stem cells signatures

Gene Name	Frequency	SAM p-value	SAM FDR
RNF2	2	0.000	0.020
HSPA1B	2	0.000	0.020
CTSB	2	0.000	0.007
MCM2	2	0.000	0.020
ANGPT1	2	0.000	0.007
MS4A3	2	0.000	0.007
GP5	2	0.001	0.029
DMP1	2	0.001	0.011
PLEK	3	0.001	0.011
KLF6	4	0.001	0.011
EGR3	4	0.001	0.011
CD58	2	0.001	0.029
JUN	4	0.002	0.014
IDS	2	0.002	0.029
ZNF124	2	0.002	0.029
DBI	4	0.002	0.041
CHIC2	2	0.002	0.041
GIMAP5	2	0.003	0.041
TLN1	2	0.003	0.020
BTG1	2	0.003	0.041
KIAA0020	2	0.004	0.041
ZCCHC6	2	0.004	0.029
CD79B	2	0.004	0.029
NANOG	2	0.005	0.029
ASCC2	2	0.005	0.029
SERPINA5	2	0.006	0.059
MPO	2	0.006	0.041
ZNF600	1	0.009	0.059
MFSD6	2	0.009	0.059
LOC55338	3	0.011	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
654056	2	0.011	0.077
PREP	2	0.012	0.077
AGT	2	0.013	0.059
ARAP3	2	0.013	0.059
TIMP3	2	0.013	0.059
CALD1	3	0.013	0.059
PHF20L1	2	0.014	0.059
GNG11	2	0.014	0.059
SIX3	2	0.016	0.059
PRKAR2B	2	0.017	0.077
POLH	3	0.018	0.077
ZFP36L2	2	0.019	0.077
DNAJA1	2	0.019	0.097
DNAJC6	2	0.020	0.097
SSX1	2	0.021	0.097
ARHGEF17	2	0.021	0.097
CSPP1	2	0.021	0.120
TFAP2A	2	0.022	0.077
LAMB4	2	0.031	0.120

Table D. Frequency of the genes within core pathway of C7 catalog

Gene Name	Frequency	SAM p-value	SAM FDR
LGALS3	17	0.006	0.041
G0S2	17	0.003	0.020
EPAS1	16	0.000	0.000
IDS	15	0.002	0.029
CXCL8	15	0.000	0.007
CD79B	14	0.004	0.029
ITGA4	14	0.006	0.041
SYPL1	14	0.009	0.059
EHD4	14	0.007	0.059
CCR6	13	0.006	0.041
IL18	13	0.000	0.000
PLEK	13	0.001	0.011
PEA15	13	0.005	0.041
APP	12	0.000	0.007
CD72	12	0.015	0.077
CTSB	12	0.000	0.007
EGR1	12	0.016	0.077
GLRX	12	0.010	0.077
S100A8	12	0.009	0.041
NFKBIZ	12	0.004	0.029
FAS	11	0.007	0.041
BTG1	11	0.003	0.041
DBI	11	0.002	0.041
GZMA	11	0.017	0.097
IDI1	11	0.000	0.020
IL10	11	0.003	0.020
MCM5	11	0.001	0.029
SDC4	11	0.000	0.007
PDLIM1	11	0.000	0.007
TM9SF1	11	0.002	0.020
ELL2	11	0.007	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
RAP1GAP2	11	0.014	0.077
PHTF2	11	0.013	0.059
ACTN1	10	0.000	0.000
AHR	10	0.000	0.000
EIF4B	10	0.001	0.011
H1F0	10	0.006	0.059
JUN	10	0.002	0.014
RNASE2	10	0.003	0.029
EVI5	10	0.001	0.011
NR1D2	10	0.000	0.020
MTHFD2	10	0.008	0.059
FAM46C	10	0.003	0.029
SPC25	10	0.001	0.029
ZMIZ2	10	0.001	0.029
BLVRA	9	0.007	0.059
CASP1	9	0.008	0.041
KLF6	9	0.001	0.011
CTLA4	9	0.008	0.059
PREP	9	0.012	0.077
RPA3	9	0.000	0.020
SNTB2	9	0.013	0.077
TXN	9	0.010	0.077
VEGFB	9	0.007	0.059
TNFSF14	9	0.001	0.007
KIAA0101	9	0.002	0.020
TRIB1	9	0.012	0.059
SMC2	9	0.005	0.059
CREB3L2	9	0.007	0.059
ATP6V1C1	8	0.001	0.029
GRK5	8	0.003	0.041
GSR	8	0.005	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
IGFBP7	8	0.000	0.000
LAIR1	8	0.004	0.029
MCM2	8	0.000	0.020
SERPINB9	8	0.002	0.041
PSMD12	8	0.002	0.020
STAT4	8	0.001	0.029
TP53BP2	8	0.005	0.041
DYRK3	8	0.001	0.029
GSTO1	8	0.010	0.077
FEZ2	8	0.002	0.029
TACC3	8	0.018	0.097
TNFRSF13B	8	0.002	0.020
LDLRAP1	8	0.011	0.077
RSRP1	8	0.001	0.029
ENTPD7	8	0.000	0.020
EML5	8	0.000	0.000
CD24	8	0.001	0.007
ALOX5	7	0.008	0.041
BLVRB	7	0.001	0.014
C4BPA	7	0.015	0.097
CD58	7	0.001	0.029
DDX6	7	0.000	0.000
EGR3	7	0.001	0.011
FCGR2B	7	0.014	0.077
GPR18	7	0.002	0.029
HLA-DMA	7	0.015	0.059
SP110	7	0.005	0.029
JARID2	7	0.011	0.077
MPO	7	0.006	0.041
NBN	7	0.010	0.059
NDUFB3	7	0.002	0.020

Gene Name	Frequency	SAM p-value	SAM FDR
PRIM2	7	0.003	0.041
SGK1	7	0.020	0.077
TGFBR1	7	0.006	0.059
MAP4K4	7	0.006	0.059
GMFG	7	0.001	0.011
APOBEC3B	7	0.007	0.041
SOCS5	7	0.007	0.059
TBC1D5	7	0.007	0.041
ZNF318	7	0.003	0.041
PDCD4	7	0.002	0.014
HIPK2	7	0.001	0.014
ERO1L	7	0.009	0.059
IMPACT	7	0.007	0.059
NAGK	7	0.002	0.041
GRWD1	7	0.002	0.041
ADTRP	7	0.008	0.041
PHF13	7	0.000	0.020
LINC00936	7	0.000	0.000
ZFP36L2	6	0.019	0.077
CD1C	6	0.002	0.014
CD2	6	0.004	0.041
DHPS	6	0.001	0.029
DSCAM	6	0.006	0.059
EGR2	6	0.009	0.059
ENO2	6	0.015	0.077
F2R	6	0.000	0.000
ACSL3	6	0.002	0.041
FCN1	6	0.001	0.011
HSPA8	6	0.006	0.041
IL1B	6	0.013	0.059
LGALS3BP	6	0.014	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
LIPA	6	0.011	0.059
LTF	6	0.013	0.059
TACSTD2	6	0.002	0.020
KMT2A	6	0.003	0.020
MTM1	6	0.014	0.077
TRIM37	6	0.004	0.029
UBL3	6	0.009	0.059
PRKAB2	6	0.000	0.020
RGS16	6	0.001	0.011
RPL5	6	0.005	0.041
RPN1	6	0.000	0.020
MSMO1	6	0.006	0.059
SCN8A	6	0.001	0.011
SDHC	6	0.003	0.041
SNAPC1	6	0.000	0.007
STXBP1	6	0.001	0.007
LAMTOR3	6	0.011	0.077
BANF1	6	0.001	0.029
MBD4	6	0.002	0.020
ATP6V1F	6	0.012	0.077
ESPL1	6	0.003	0.041
HERPUD1	6	0.019	0.097
ARPC2	6	0.011	0.059
IRF9	6	0.004	0.041
TUBGCP3	6	0.004	0.029
TIMM17A	6	0.013	0.077
HTATIP2	6	0.007	0.041
IMMT	6	0.012	0.077
STK38	6	0.016	0.097
RAB21	6	0.005	0.059
SLC44A1	6	0.010	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
CLEC5A	6	0.016	0.077
SPATS2L	6	0.011	0.077
SESN1	6	0.020	0.097
SAP30BP	6	0.011	0.059
ZC2HC1A	6	0.003	0.029
TRPV2	6	0.002	0.029
HMP19	6	0.010	0.077
MPP6	6	0.008	0.059
SELT	6	0.001	0.029
MFSD6	6	0.009	0.059
GIMAP5	6	0.003	0.041
CCDC47	6	0.014	0.077
SLAMF7	6	0.008	0.059
CSRNP1	6	0.008	0.041
E2F8	6	0.010	0.059
EDEM3	6	0.004	0.041
YPEL3	6	0.005	0.041
OSBPL9	6	0.005	0.059
DHRS1	6	0.007	0.059
FBXL14	6	0.006	0.041
42071	6	0.005	0.059
ADD1	5	0.007	0.041
ALDH9A1	5	0.001	0.029
BCL3	5	0.010	0.059
BRCA1	5	0.006	0.059
BST2	5	0.009	0.059
TSPO	5	0.007	0.041
CAPNS1	5	0.002	0.029
CAPZB	5	0.003	0.020
CDC6	5	0.002	0.020
CEACAM8	5	0.000	0.007

Gene Name	Frequency	SAM p-value	SAM FDR
CKS2	5	0.005	0.059
CR2	5	0.011	0.059
DDX5	5	0.004	0.029
S1PR1	5	0.033	0.148
FDFT1	5	0.010	0.059
GABRB1	5	0.007	0.041
HBZ	5	0.001	0.029
IGF1R	5	0.012	0.059
JAK2	5	0.000	0.020
CYP4F3	5	0.003	0.020
SH2D1A	5	0.022	0.120
MGMT	5	0.000	0.011
NAP1L1	5	0.001	0.011
NUCB2	5	0.002	0.020
OLR1	5	0.002	0.014
OPA1	5	0.011	0.059
SERPINE1	5	0.008	0.059
PLAGL2	5	0.004	0.041
PLEC	5	0.016	0.077
POLR2I	5	0.008	0.059
DNAJC3	5	0.014	0.059
RELA	5	0.010	0.059
RFX3	5	0.009	0.041
RNASE3	5	0.001	0.007
RPE	5	0.001	0.029
RPL11	5	0.005	0.029
RSU1	5	0.005	0.029
RYK	5	0.018	0.097
SLC2A5	5	0.001	0.011
SPI1	5	0.011	0.059
SUOX	5	0.007	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
YWHAE	5	0.001	0.029
ZNF43	5	0.007	0.059
PICALM	5	0.009	0.059
PIP4K2B	5	0.000	0.000
PPAP2B	5	0.016	0.077
GPAA1	5	0.012	0.077
MTMR3	5	0.001	0.029
PGLYRP1	5	0.023	0.077
CBFA2T2	5	0.004	0.041
REPS2	5	0.006	0.041
MED20	5	0.016	0.097
IQSEC1	5	0.014	0.077
MBNL2	5	0.007	0.059
MPZL2	5	0.002	0.041
NET1	5	0.012	0.077
CEPT1	5	0.001	0.029
YAP1	5	0.000	0.000
SPTLC1	5	0.001	0.011
LILRB1	5	0.000	0.000
BTG3	5	0.016	0.097
FCHO1	5	0.011	0.077
KIAA0922	5	0.016	0.097
FBXW11	5	0.005	0.029
RNF19A	5	0.000	0.020
HBP1	5	0.000	0.020
AP3M1	5	0.001	0.029
VPS41	5	0.006	0.041
DSE	5	0.018	0.097
STOML2	5	0.001	0.029
ABHD5	5	0.008	0.059
ZDHHC2	5	0.001	0.011

Gene Name	Frequency	SAM p-value	SAM FDR
MRPL51	5	0.007	0.059
OSER1	5	0.012	0.077
ACSL5	5	0.001	0.011
YIPF1	5	0.002	0.041
SMOX	5	0.004	0.041
SUSD4	5	0.011	0.059
CDV3	5	0.004	0.029
MAP7D1	5	0.004	0.041
CLK4	5	0.000	0.014
ABRACL	5	0.016	0.097
MRPL17	5	0.017	0.097
MCUR1	5	0.008	0.059
GGCT	5	0.004	0.041
ALG8	5	0.006	0.059
EFHD2	5	0.012	0.059
CHPF	5	0.005	0.041
CEP97	5	0.002	0.029
QSER1	5	0.004	0.041
PEAK1	5	0.017	0.097
FAM213A	5	0.001	0.011
MCEE	5	0.001	0.029
MFSD2A	5	0.002	0.014
CDKN2AIPNL	5	0.002	0.014
ST7-AS1	5	0.008	0.041
SLC52A3	5	0.012	0.059
OTUD1	5	0.000	0.007
FRYL	5	0.001	0.029
ACADVL	4	0.010	0.077
ADCY8	4	0.001	0.011
ADK	4	0.016	0.077
ADRB2	4	0.017	0.097

Gene Name	Frequency	SAM p-value	SAM FDR
ALDH2	4	0.015	0.059
AMPD1	4	0.008	0.041
ARF4	4	0.017	0.077
ATP5B	4	0.015	0.097
CD22	4	0.032	0.148
CD81	4	0.038	0.097
CSF2RA	4	0.001	0.007
CTNND1	4	0.005	0.029
DNMT3B	4	0.002	0.020
DOCK3	4	0.000	0.000
ELK4	4	0.004	0.041
GART	4	0.002	0.029
HOXD9	4	0.011	0.059
TNFRSF9	4	0.022	0.120
RPSA	4	0.017	0.097
ABLIM1	4	0.035	0.148
MEOX1	4	0.013	0.059
MGAT1	4	0.001	0.029
MX1	4	0.015	0.059
MYH10	4	0.014	0.077
NCAM1	4	0.006	0.059
NDUFV1	4	0.002	0.041
NDUFS6	4	0.013	0.059
PDK4	4	0.006	0.059
PFKFB1	4	0.004	0.029
PSMA1	4	0.004	0.041
RARRES3	4	0.001	0.029
RHD	4	0.006	0.059
RPS23	4	0.001	0.011
SRP68	4	0.009	0.059
HSPA13	4	0.005	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
TIMP1	4	0.021	0.097
TIMP3	4	0.013	0.059
TJP1	4	0.002	0.029
TOP3A	4	0.001	0.011
ZNF124	4	0.002	0.029
SLC30A1	4	0.017	0.097
SLC10A3	4	0.006	0.059
CBX4	4	0.009	0.041
DEGS1	4	0.015	0.059
AKR1C3	4	0.001	0.029
TIMELESS	4	0.004	0.041
EXO1	4	0.012	0.059
STK17A	4	0.022	0.120
TRIP10	4	0.006	0.041
RAB9A	4	0.011	0.077
GRHPR	4	0.004	0.041
ROCK2	4	0.000	0.000
ATP5J2	4	0.003	0.041
DLGAP5	4	0.015	0.077
KIAA0020	4	0.004	0.041
FARSB	4	0.000	0.000
TSPAN32	4	0.010	0.077
MICU1	4	0.002	0.041
GNLY	4	0.001	0.029
POLD3	4	0.023	0.120
ZNF273	4	0.003	0.041
RUNDC3A	4	0.003	0.041
SMPDL3A	4	0.037	0.148
LILRB3	4	0.000	0.000
GABARAPL2	4	1.000	0.596
STAB1	4	0.026	0.097

Gene Name	Frequency	SAM p-value	SAM FDR
SYT11	4	0.018	0.077
ARHGEF12	4	0.019	0.077
SEC61G	4	0.011	0.077
MORC3	4	0.011	0.077
ZMYND8	4	0.015	0.097
METTL7A	4	0.026	0.120
POT1	4	0.001	0.029
SYF2	4	0.008	0.059
DHRS7B	4	0.004	0.041
HINFP	4	0.004	0.041
ARFGAP3	4	0.004	0.041
CHIC2	4	0.002	0.041
TUBGCP4	4	0.005	0.029
CRCP	4	0.000	0.000
SLCO4A1	4	0.005	0.029
MRPS18B	4	0.017	0.097
MRPL18	4	0.011	0.077
THYN1	4	0.010	0.059
PYCARD	4	0.011	0.077
PARVB	4	0.019	0.077
PSAT1	4	0.004	0.041
SLC40A1	4	0.000	0.007
ZNRD1	4	0.005	0.059
TBC1D7	4	0.009	0.059
SLC15A3	4	0.017	0.077
TDP2	4	0.016	0.097
CAB39	4	0.003	0.041
GDAP2	4	0.004	0.041
ZDHHC4	4	0.008	0.059
STEAP3	4	0.004	0.041
SLC29A3	4	0.003	0.020

Gene Name	Frequency	SAM p-value	SAM FDR
ZDHHC7	4	0.027	0.120
FLVCR2	4	0.009	0.059
RAB20	4	0.007	0.041
C1orf112	4	0.010	0.077
PLSCR3	4	0.003	0.041
CORO1B	4	0.011	0.077
THAP11	4	0.003	0.041
GNPNAT1	4	0.005	0.059
KXD1	4	0.005	0.059
PHF23	4	0.013	0.077
PLEKHF1	4	0.008	0.059
LILRA6	4	0.005	0.059
ZXDC	4	0.005	0.059
ZCCHC6	4	0.004	0.029
PLBD1	4	0.033	0.097
DNAJB14	4	0.007	0.059
ZNF703	4	0.011	0.059
PQLC1	4	0.006	0.041
C19orf12	4	0.003	0.041
DIRC2	4	0.008	0.059
HIST1H2BK	4	0.002	0.029
PNPT1	4	0.000	0.000
GADD45GIP1	4	0.004	0.041
SLC25A26	4	0.005	0.059
ANTXR2	4	0.006	0.041
PSTK	4	0.006	0.059
NIPA1	4	0.002	0.029
GAB3	4	0.019	0.097
FAM134C	4	0.011	0.077
MPZL3	4	0.006	0.059
KRTCAP2	4	0.002	0.029

Gene Name	Frequency	SAM p-value	SAM FDR
HIPK1	4	0.001	0.029
BRWD3	4	0.002	0.020
ARHGAP30	4	0.004	0.041
DDX51	4	0.001	0.007
FRRS1	4	0.008	0.059
IRG1	4	0.020	0.077
ABCF1	3	0.002	0.041
AFM	3	0.001	0.014
AMELX	3	0.016	0.077
RHOB	3	0.021	0.077
RHOG	3	0.003	0.041
ATP2B4	3	0.019	0.097
BPHL	3	0.000	0.020
BUB1	3	0.014	0.077
CBLB	3	0.031	0.148
CDK5	3	0.019	0.077
CKS1B	3	0.025	0.120
CCR8	3	0.011	0.077
COL8A2	3	0.006	0.059
COX6A2	3	0.006	0.041
CRY2	3	0.000	0.020
CSNK1D	3	0.013	0.077
DSP	3	0.001	0.011
EIF2S1	3	0.007	0.059
ENSA	3	0.010	0.077
ETF1	3	0.014	0.077
ACSL4	3	0.007	0.059
PTK2B	3	0.006	0.059
FCGRT	3	0.028	0.120
FKTN	3	0.000	0.007
FHIT	3	0.020	0.077

Gene Name	Frequency	SAM p-value	SAM FDR
FOLR2	3	0.012	0.077
KDSR	3	0.007	0.059
GSTP1	3	0.001	0.020
HSD17B10	3	0.028	0.097
HAS3	3	0.007	0.041
HMBS	3	0.006	0.059
HNRNPK	3	0.001	0.011
IDH3A	3	0.022	0.120
IL6R	3	0.023	0.120
INPP4A	3	0.021	0.097
KLRC3	3	0.011	0.077
LIMK1	3	0.015	0.059
MTIF2	3	0.007	0.059
MYB	3	0.034	0.097
NDUFA4	3	0.010	0.059
NKG7	3	0.004	0.041
CNOT3	3	0.012	0.059
TNFRSF11B	3	0.001	0.011
OSBP	3	0.004	0.029
PIK3CD	3	0.024	0.120
PMS2P5	3	0.011	0.077
PPP2R2B	3	0.002	0.029
PPP2R3A	3	0.006	0.041
MAPK13	3	0.006	0.059
PRTN3	3	0.005	0.041
PSMB7	3	0.011	0.059
RNF2	3	0.000	0.020
RPE65	3	0.015	0.059
SCN7A	3	0.009	0.059
CCL17	3	0.006	0.041
SLC7A1	3	0.011	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
SOAT1	3	0.001	0.029
SSR1	3	0.000	0.007
SSR4	3	0.020	0.097
STAT5A	3	0.019	0.097
STAU1	3	0.001	0.029
SURF1	3	0.005	0.059
CNTN2	3	0.007	0.041
TERF1	3	0.002	0.029
TERF2	3	0.007	0.059
TFCP2	3	0.004	0.041
THBS2	3	0.015	0.097
TPD52	3	0.042	0.120
UBE2G1	3	0.004	0.041
UBE2I	3	0.009	0.059
DENR	3	0.008	0.041
SKAP1	3	0.022	0.120
AP1G2	3	0.017	0.097
AP1M1	3	0.010	0.059
KRT75	3	0.002	0.020
RABEP1	3	0.017	0.097
ZMYM6	3	0.002	0.029
HOMER1	3	0.036	0.148
EIF5B	3	0.004	0.041
DEPDC5	3	0.006	0.059
DCLRE1A	3	0.007	0.041
THRAP3	3	0.010	0.077
TSFM	3	0.018	0.097
RBM5	3	0.027	0.120
NDC80	3	0.006	0.059
DDX17	3	0.014	0.077
NPC2	3	0.033	0.148

Gene Name	Frequency	SAM p-value	SAM FDR
SORBS1	3	0.009	0.059
BRD8	3	0.002	0.029
SLC27A2	3	0.012	0.059
RIPK3	3	0.005	0.041
SUPT16H	3	0.011	0.077
TCF25	3	0.000	0.020
MAST1	3	0.000	0.007
ZNF292	3	0.010	0.077
ZC3H13	3	0.010	0.059
FAM120A	3	0.008	0.059
KIAA1033	3	0.014	0.077
SIRT3	3	0.001	0.029
LINC01565	3	0.002	0.020
CBX5	3	0.005	0.059
TMEM131	3	0.003	0.020
ORC6	3	0.001	0.011
PITPNB	3	0.003	0.020
YIPF3	3	0.000	0.014
LSM14A	3	0.006	0.041
TRAF3IP1	3	0.030	0.097
ZNF337	3	0.007	0.059
HEYL	3	0.010	0.077
TIMM10B	3	0.006	0.059
RBMXL2	3	0.002	0.020
P2RY10	3	0.042	0.148
MRPL42	3	0.017	0.097
C1GALT1C1	3	0.034	0.148
METTL5	3	0.005	0.059
SNX15	3	0.011	0.077
LRP12	3	0.005	0.029
COPS7A	3	0.016	0.097

Gene Name	Frequency	SAM p-value	SAM FDR
ASCC1	3	0.016	0.097
RPS27L	3	0.004	0.041
ZNF706	3	0.002	0.029
EMC4	3	0.001	0.011
RNF181	3	0.018	0.097
42065	3	0.003	0.041
MBTPS2	3	0.012	0.077
TMEM14C	3	0.012	0.077
UFM1	3	0.007	0.041
MYOZ2	3	0.005	0.029
TPCN1	3	0.005	0.059
CPSF2	3	0.006	0.059
TMCO1	3	0.010	0.077
SGTB	3	0.001	0.020
OTUD4	3	0.016	0.097
MTMR10	3	0.013	0.077
FOCAD	3	0.012	0.059
RHBDL2	3	0.001	0.011
EVA1B	3	0.006	0.041
SIRPG	3	0.016	0.077
DENND4C	3	0.018	0.077
LMBRD1	3	0.006	0.059
WDR45B	3	0.010	0.059
GJC2	3	0.014	0.077
WDR18	3	0.001	0.029
TSKS	3	0.006	0.041
NPAS3	3	0.002	0.029
ARAP3	3	0.013	0.059
RMND5B	3	0.003	0.041
C3orf52	3	0.003	0.020
COLGALT1	3	0.003	0.041

Gene Name	Frequency	SAM p-value	SAM FDR
SHCBP1	3	0.001	0.011
FAM124B	3	0.004	0.041
TMEM156	3	0.004	0.029
ADAMTS20	3	0.012	0.059
ZNF606	3	0.003	0.041
SPSB1	3	0.015	0.059
SLC44A4	3	0.000	0.000
NDFIP1	3	0.013	0.077
APOL3	3	0.014	0.059
DCSTAMP	3	0.008	0.041
FRMD8	3	0.001	0.029
KLF16	3	0.006	0.041
BCO2	3	0.002	0.014
ZNF644	3	0.001	0.029
ZCCHC9	3	0.008	0.041
YIPF4	3	0.000	0.007
HOOK3	3	0.026	0.120
CYSTM1	3	0.012	0.077
ANKRD44	3	0.009	0.041
SLC25A51	3	0.001	0.011
ORMDL1	3	0.001	0.029
KLHL29	3	0.001	0.029
GPR146	3	0.008	0.041
NAA30	3	0.002	0.041
LEO1	3	0.003	0.041
HEXIM2	3	0.003	0.041
TAF8	3	0.007	0.059
TTC36	3	0.005	0.029
HSCB	3	0.011	0.059
TTL	3	0.018	0.097
KANSL1L	3	0.002	0.029

Gene Name	Frequency	SAM p-value	SAM FDR
RNF145	3	0.003	0.041
RDH10	3	0.011	0.059
C8orf37	3	0.000	0.007
HIGD2A	3	0.001	0.007
TWISTNB	3	0.011	0.077
ADGRD1	3	0.005	0.059
ZNF326	3	0.014	0.059
SWI5	3	0.010	0.059
HACD4	3	0.001	0.014
ALDH3A1	2	0.009	0.059
ANGPT1	2	0.000	0.007
ANXA6	2	0.025	0.120
ATP1A2	2	0.001	0.011
ATP2A2	2	0.027	0.120
ATP5O	2	0.004	0.041
BARD1	2	0.039	0.120
BTC	2	0.002	0.020
CAMP	2	0.001	0.011
CASP2	2	0.000	0.020
KRIT1	2	0.006	0.041
MS4A3	2	0.000	0.007
CD47	2	0.064	0.148
CEACAM7	2	0.005	0.041
COL6A3	2	0.018	0.077
COX15	2	0.003	0.041
CREB1	2	0.000	0.007
CSNK2B	2	0.002	0.029
CSTB	2	0.036	0.148
CTGF	2	0.009	0.059
CYP2B6	2	0.001	0.029
CYP27B1	2	0.010	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
CD55	2	0.026	0.120
DBT	2	0.015	0.059
DCC	2	0.013	0.059
DNASE1L2	2	0.013	0.059
DPH2	2	0.013	0.077
DPYS	2	0.020	0.077
DVL3	2	0.016	0.097
EIF4EBP2	2	0.011	0.077
ELAVL3	2	0.008	0.041
EPHB6	2	0.009	0.059
ERF	2	0.016	0.077
FANCG	2	0.002	0.041
FHL2	2	0.002	0.029
FOXL1	2	0.019	0.077
FPR2	2	0.007	0.059
GARS	2	0.018	0.097
GJA3	2	0.005	0.029
GNB3	2	0.001	0.014
GNG11	2	0.014	0.059
GOT1	2	0.002	0.041
GPR15	2	0.002	0.014
GRIK3	2	0.001	0.014
GSPT1	2	0.018	0.077
GSTA3	2	0.006	0.041
GSTT1	2	0.003	0.041
GTF2A2	2	0.037	0.148
HBD	2	0.013	0.059
CFHR2	2	0.002	0.014
HSPA1B	2	0.000	0.020
IDO1	2	0.005	0.029
INHBC	2	0.000	0.007

Gene Name	Frequency	SAM p-value	SAM FDR
INPP1	2	0.037	0.097
ITGA1	2	0.007	0.041
ITGA2B	2	0.011	0.059
KCNJ9	2	0.004	0.029
SMAD3	2	0.035	0.148
MARS	2	0.003	0.041
MFAP1	2	0.026	0.120
MLN	2	0.013	0.059
MT1H	2	0.025	0.077
MUC1	2	0.005	0.041
MYC	2	0.056	0.120
NDUFA1	2	0.020	0.097
NEDD8	2	0.020	0.097
NKX6-1	2	0.007	0.041
NPM1	2	0.000	0.007
NPY1R	2	0.006	0.041
OMP	2	0.001	0.011
P2RX5	2	0.066	0.148
SERPINA5	2	0.006	0.059
PDE9A	2	0.002	0.041
PDK1	2	0.036	0.148
PDYN	2	0.008	0.059
PFDN5	2	0.010	0.077
PFN2	2	0.004	0.029
PMS2P1	2	0.003	0.041
POU2AF1	2	0.065	0.205
PPID	2	0.010	0.077
EIF2AK2	2	0.022	0.077
TMPRSS15	2	0.000	0.014
PSMD7	2	0.020	0.097
PTGIR	2	0.008	0.041

Gene Name	Frequency	SAM p-value	SAM FDR
PTGS2	2	0.029	0.097
RAB6A	2	0.001	0.011
RAD21	2	0.010	0.077
RAD51	2	0.018	0.077
RHEB	2	0.035	0.148
RPL9	2	0.001	0.014
RPL39	2	0.014	0.077
RPS3A	2	0.017	0.077
SFTPC	2	0.001	0.011
SHB	2	0.029	0.097
SMARCA4	2	0.002	0.041
SMARCD2	2	0.010	0.059
SNAPC3	2	0.002	0.029
SPN	2	0.021	0.077
ST14	2	0.020	0.077
ADAM17	2	0.004	0.029
TBCE	2	0.042	0.148
TBX5	2	0.012	0.059
TBX15	2	0.003	0.020
TCEB3	2	0.004	0.029
TG	2	0.020	0.077
GPR137B	2	0.032	0.148
TMPO	2	0.025	0.120
TPM3	2	0.014	0.077
USP1	2	0.007	0.059
VCAM1	2	0.002	0.014
BEST1	2	0.027	0.097
XK	2	0.005	0.029
ZNF180	2	0.029	0.097
ZFAND5	2	0.022	0.120
SLMAP	2	0.017	0.077

Gene Name	Frequency	SAM p-value	SAM FDR
KAT6A	2	0.036	0.148
MFAP5	2	0.002	0.014
ULK1	2	0.000	0.020
CUL4B	2	0.012	0.077
RGS5	2	0.020	0.077
ITGA10	2	0.009	0.041
RNASET2	2	0.019	0.077
B4GALT3	2	0.001	0.029
CCNA1	2	0.009	0.059
P4HA2	2	0.012	0.059
CLDN12	2	0.006	0.059
PIGQ	2	0.016	0.097
SLC16A6	2	0.028	0.097
ARHGEF2	2	0.020	0.077
KIF3B	2	0.003	0.041
CRIPT	2	0.000	0.020
EI24	2	0.004	0.041
TTI1	2	0.019	0.097
KIAA0391	2	0.023	0.120
NOS1AP	2	0.011	0.059
EIF4A3	2	0.017	0.097
DNAJC6	2	0.020	0.097
ZNF623	2	0.015	0.059
MELK	2	0.021	0.097
EPM2AIP1	2	0.017	0.097
CEP350	2	0.000	0.020
TOMM70A	2	0.005	0.041
SEC16A	2	0.003	0.041
CASP8AP2	2	0.012	0.059
IL18BP	2	0.013	0.059
ACTR2	2	0.002	0.020

Gene Name	Frequency	SAM p-value	SAM FDR
YAF2	2	0.003	0.041
ANGPTL7	2	0.025	0.120
APBB3	2	0.002	0.029
LYPLA1	2	0.010	0.077
ATP5H	2	0.010	0.077
SEMA3C	2	0.020	0.077
DEAF1	2	0.003	0.041
PDPN	2	0.004	0.041
B3GNT2	2	0.022	0.077
ARPP19	2	0.001	0.007
USP20	2	0.012	0.077
HCST	2	0.017	0.077
GPR75	2	0.005	0.029
AFG3L2	2	0.007	0.059
RAB35	2	0.010	0.077
WWP1	2	0.014	0.077
C10orf10	2	0.003	0.041
ESM1	2	0.005	0.059
RPP14	2	0.022	0.120
DMC1	2	0.007	0.041
NUDT6	2	0.009	0.041
RASSF1	2	0.029	0.120
PROSC	2	0.009	0.059
RPL35	2	0.006	0.041
FKBP9	2	0.013	0.077
RRAS2	2	0.061	0.205
ZNF652	2	0.001	0.011
COBLL1	2	0.054	0.176
DOLK	2	0.003	0.041
SEC31A	2	0.036	0.097
CEP164	2	0.005	0.059

Gene Name	Frequency	SAM p-value	SAM FDR
RAB18	2	0.001	0.029
FBXO21	2	0.021	0.097
ZHX3	2	0.002	0.041
CLUAP1	2	0.024	0.077
MRPS27	2	0.000	0.014
CLASP1	2	0.013	0.077
UBR4	2	0.001	0.029
ATP13A2	2	0.017	0.077
FBXO46	2	0.005	0.041
MLYCD	2	0.027	0.120
ABCA5	2	0.001	0.029
SEC11A	2	0.012	0.059
DAPK2	2	0.011	0.059
RUSC1	2	0.024	0.120
IL17RA	2	0.033	0.148
SAMHD1	2	0.020	0.097
FAM98A	2	0.031	0.120
SEC31B	2	0.014	0.077
KANK2	2	0.006	0.041
SIPA1L1	2	0.035	0.148
SETBP1	2	0.042	0.148
AUTS2	2	0.024	0.120
ANKRD17	2	0.002	0.041
APPL1	2	0.037	0.148
TES	2	0.011	0.077
GMEB2	2	0.002	0.041
CHORDC1	2	0.000	0.007
PPA2	2	0.000	0.007
MMADHC	2	0.001	0.029
R3HCC1L	2	0.005	0.059
NKIRAS1	2	0.000	0.020

Gene Name	Frequency	SAM p-value	SAM FDR
COMMD9	2	0.016	0.097
VPREB3	2	0.019	0.097
CTAG2	2	0.001	0.014
HDGFRP3	2	0.008	0.059
RNF141	2	0.034	0.148
PHF20L1	2	0.014	0.059
RMDN1	2	0.006	0.059
DYNC1LI1	2	0.004	0.041
TUBD1	2	0.006	0.059
ZMYND10	2	0.002	0.020
ZNF589	2	0.009	0.059
AIG1	2	0.007	0.041
ACTL6B	2	0.006	0.041
ETV7	2	0.007	0.041
VTA1	2	0.029	0.097
CUTA	2	0.017	0.097
KLF13	2	0.013	0.077
TRIM34	2	0.016	0.097
TOLLIP	2	0.022	0.077
HEATR5B	2	0.008	0.059
DDX56	2	0.001	0.029
WDR74	2	0.001	0.014
CROT	2	0.015	0.059
AHI1	2	0.003	0.041
RNF125	2	0.021	0.097
OXSM	2	0.022	0.120
PARPBP	2	0.006	0.059
PIH1D1	2	0.002	0.020
VPS37C	2	0.004	0.041
TMEM51	2	0.010	0.059
TBCCD1	2	0.047	0.176

Gene Name	Frequency	SAM p-value	SAM FDR
LOC55338	2	0.011	0.059
GPALPP1	2	0.004	0.041
KIF16B	2	0.017	0.077
TASP1	2	0.001	0.011
ZNF692	2	0.013	0.077
HIF1AN	2	0.003	0.020
ARHGEF40	2	0.000	0.014
FAR2	2	0.003	0.020
N4BP2	2	0.001	0.007
CDK5RAP2	2	0.011	0.077
BDP1	2	0.002	0.029
ECHDC1	2	0.033	0.148
SLC50A1	2	0.037	0.097
ALG1	2	0.007	0.041
METTL3	2	0.009	0.059
TUBB7P	2	0.001	0.029
THAP10	2	0.012	0.059
PDSS2	2	0.009	0.059
MRS2	2	0.014	0.077
PRR12	2	0.004	0.041
USP28	2	0.016	0.097
PHF12	2	0.016	0.077
RNF213	2	0.012	0.059
USP37	2	0.006	0.059
HMHB1	2	0.011	0.077
RAP2C	2	0.009	0.059
HRH4	2	0.002	0.029
NIF3L1	2	0.022	0.120
DMRTB1	2	0.004	0.041
MCCC2	2	0.014	0.077
EPB41L4A	2	0.011	0.077

Gene Name	Frequency	SAM p-value	SAM FDR
CENPK	2	0.010	0.077
DCLRE1C	2	0.017	0.077
ACBD3	2	0.005	0.029
AEN	2	0.017	0.097
RFX7	2	0.009	0.059
PAPOLG	2	0.016	0.097
MRPL14	2	0.000	0.011
MRPL44	2	0.001	0.029
TRAK2	2	0.015	0.059
MIS12	2	0.005	0.041
AHNAK	2	0.037	0.097
SPATA5L1	2	0.016	0.097
PRRG4	2	0.026	0.097
GNPTAB	2	0.002	0.041
ZNF557	2	0.003	0.020
NKAIN1	2	0.007	0.041
C10orf76	2	0.006	0.059
CLIP4	2	0.019	0.097
TXNDC15	2	0.001	0.029
WDR76	2	0.014	0.077
ESRP2	2	0.018	0.097
CCDC33	2	0.001	0.029
NUBPL	2	0.000	0.007
ZNF430	2	0.003	0.020
FER1L4	2	0.004	0.029
ADAM33	2	0.007	0.059
TMEM121	2	0.003	0.029
ITIH5	2	0.000	0.007
SPACA1	2	0.016	0.097
ESPN	2	0.004	0.041
RBM4B	2	0.004	0.041

Gene Name	Frequency	SAM p-value	SAM FDR
TMEM120A	2	0.029	0.120
BRIP1	2	0.008	0.059
FAM160A2	2	0.022	0.120
SPATA22	2	0.009	0.041
RAX2	2	0.004	0.029
FAM104A	2	0.023	0.077
RSPRY1	2	0.001	0.029
TMEM263	2	0.004	0.041
ZNF799	2	0.001	0.029
BOD1	2	0.007	0.059
PCED1B	2	0.013	0.077
WDR20	2	0.009	0.059
FANK1	2	0.012	0.077
IGSF8	2	0.007	0.041
ARHGAP12	2	0.011	0.077
PPP1R14A	2	0.012	0.059
GLCCI1	2	0.008	0.059
CCDC124	2	0.004	0.029
MUCL1	2	0.004	0.041
SPACA7	2	0.015	0.077
LRR1	2	0.002	0.029
METTL23	2	0.008	0.041
UHMK1	2	0.003	0.020
PIGU	2	0.011	0.077
MITD1	2	0.021	0.097
C4orf33	2	0.001	0.029
PACRG	2	0.001	0.029
PIWIL4	2	0.018	0.077
C18orf25	2	0.001	0.029
RNF187	2	0.023	0.120
COMMD7	2	0.002	0.014

Gene Name	Frequency	SAM p-value	SAM FDR
CKAP2L	2	0.008	0.059
NFXL1	2	0.018	0.077
C9orf66	2	0.007	0.059
KRT28	2	0.001	0.011
UBXN2A	2	0.004	0.029
DCP1B	2	0.002	0.014
SLC5A9	2	0.013	0.059
CENPV	2	0.005	0.041
PDE12	2	0.016	0.097
MTIF3	2	0.007	0.059
ZNRF2	2	0.012	0.077
FDCSP	2	0.006	0.041
C11orf31	2	0.019	0.097
FAM177A1	2	0.001	0.011
HERC2P3	2	0.002	0.029
IRGM	2	0.027	0.097
SERINC2	2	0.014	0.059
TEPP	2	0.001	0.029
FAM73A	2	0.001	0.014
C12orf75	2	0.025	0.120
C2orf68	2	0.024	0.077
C4orf3	2	0.004	0.029
GTF2H5	2	0.024	0.120
MTHFD2L	2	0.003	0.041
LOC646870	2	0.001	0.014
GATSL3	2	0.000	0.011
MUC5B	2	0.005	0.041