

5-2013

## INTEGRATIVE BIOMARKER IDENTIFICATION AND CLASSIFICATION USING HIGH THROUGHPUT ASSAYS

Pan Tong

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)



Part of the [Biostatistics Commons](#), [Medicine and Health Sciences Commons](#), [Microarrays Commons](#), [Multivariate Analysis Commons](#), and the [Statistical Models Commons](#)

---

### Recommended Citation

Tong, Pan, "INTEGRATIVE BIOMARKER IDENTIFICATION AND CLASSIFICATION USING HIGH THROUGHPUT ASSAYS" (2013). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 337.  
[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/337](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/337)

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

INTEGRATIVE BIOMARKER IDENTIFICATION AND CLASSIFICATION  
USING HIGH THROUGHPUT ASSAYS

by

Pan Tong, BE

APPROVED:

---

Kevin R. Coombes, Ph.D., Supervisory Professor

---

Lynne V. Abruzzo, M.D., Ph.D.

---

Keith A. Baggerly, Ph.D.

---

Gordon B. Mills, M.D., Ph.D.

---

John N. Weinstein, M.D., Ph.D.

APPROVED:

---

Dean, The University of Texas

Graduate School of Biomedical Sciences at Houston

**INTEGRATIVE BIOMARKER  
IDENTIFICATION AND CLASSIFICATION  
USING HIGH THROUGHPUT ASSAYS**

A

DISSERTATION

Presented to the Faculty of

The University of Texas

Health Science Center at Houston

and

The University of Texas

M. D. Anderson Cancer Center

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Pan Tong, B.E.

Houston, Texas

May, 2013

## Acknowledgments

First and foremost I would like to express my sincere gratitude to my mentor Dr. Kevin R. Coombes for his constant help, patience, support, and encouragement during my Ph.D. training. His knowledge, enthusiasm, creativity and vision for science has set a role model for me. I could not have imagined having a better advisor for my Ph.D. study. I would also like to thank my committee members Drs. Lynne V. Abruzzo, Keith A. Baggerly, Shoudan Liang, Gordon B. Mills and John N. Weinstein for their encouragement, insightful comments and hard questions that make my Ph.D. training rewarding. Besides the advisory committees, I am also thankful to other faculty in the Departments of Biostatistics, Bioinformatics and Computational Biology at MD Anderson including Drs. Xuelin Huang, Yuan Ji, Jeffrey S. Morris, Jing Wang and Li Zhang. I am honored to interact with these faculties and to learn from their diverse expertise.

My sincere thanks also goes to Dr. Victoria P. Knutson for providing academic oversight and suggestions that makes my progress smoother. I am also indebted to Yolanda Vidaurri and Ryan Thompson who provided administrative support during my Ph.D. training. I would like to extend my gratitude to the students in the Division of Quantitative Sciences who have enriched my life tremendously.

Lastly, I would like to thank my family for their unconditional love and care. My parents have taught me to be a good person and supported me in all my pursuits. My sister has been my best friend. My special thanks goes to the newest additions to my family, Xiao, my wife as well as her wonderful family who all have been supportive and caring. I truly thank Xiao for sticking by my side and instilling confidence in front of challenges and uncertainties. There is no words to convey how much I love her.



## Abstract

It is well accepted that tumorigenesis is a multi-step procedure involving aberrant functioning of genes regulating cell proliferation, differentiation, apoptosis, genome stability, angiogenesis and motility. To obtain a full understanding of tumorigenesis, it is necessary to collect information on all aspects of cell activity. Recent advances in high throughput technologies allow biologists to generate massive amounts of data, more than might have been imagined decades ago. These advances have made it possible to launch comprehensive projects such as (TCGA) and (ICGC) which systematically characterize the molecular fingerprints of cancer cells using gene expression, methylation, copy number, microRNA and SNP microarrays as well as next generation sequencing assays interrogating somatic mutation, insertion, deletion, translocation and structural rearrangements. Given the massive amount of data, a major challenge is to integrate information from multiple sources and formulate testable hypotheses.

This thesis focuses on developing methodologies for integrative analyses of genomic assays profiled on the same set of samples. We have developed several novel methods for integrative biomarker identification and cancer classification. We introduce a regression-based approach to identify biomarkers predictive to therapy response or survival by integrating multiple assays including gene expression, methylation and copy number data through penalized regression. To identify key cancer-specific genes accounting for multiple mechanisms of regulation, we have developed the `integIRTy` software that provides robust and reliable inferences about gene alteration by automatically adjusting for sample heterogeneity as well as technical artifacts using Item Response Theory.

To cope with the increasing need for accurate cancer diagnosis and individualized therapy, we have developed a robust and powerful algorithm called

SIBER to systematically identify bimodally expressed genes using next generation RNAseq data. We have shown that prediction models built from these bimodal genes have the same accuracy as models built from all genes. Further, prediction models with dichotomized gene expression measurements based on their bimodal shapes still perform well. The effectiveness of outcome prediction using discretized signals paves the road for more accurate and interpretable cancer classification by integrating signals from multiple sources.

# Contents

<b>Title Page</b>	<b>ii</b>
<b>Acknowledgments</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background	1
1.2 Related work and motivation	3
1.3 Thesis organization and contributions	6
<b>2 Prognostic biomarker identification through data integration</b>	<b>8</b>
2.1 Background	8
2.2 Methods	10
2.2.1 Logistic regression model	10
2.2.2 Cox Proportional Hazards model	12
2.2.3 Hypothesis testing	13
2.3 Results	15

2.3.1	Cancer subtypes and prognosis	15
2.3.2	Predictive power of different assay types	18
2.3.3	Cancer genes and predictive power	23
2.4	Discussion	26
<b>3</b>	<b>Gene alteration identified by the Item Response Model</b>	<b>28</b>
3.1	Background	28
3.2	Methods	31
3.2.1	Parameter Estimation	32
3.2.2	Estimation of Latent Traits from Integrated Data	34
3.2.3	Statistical Significance Assessment	34
3.2.4	Data Dichotomization	35
3.3	Results	36
3.3.1	Alteration Patterns Across Assays	36
3.3.2	Sensitivity Analysis	37
3.3.3	Contribution of Individual Assays to the Integrated Analysis	37
3.3.4	Novel Altered Genes Emerge When Data Is Integrated	39
3.3.5	Comparison to Conventional Methods	40
3.3.6	Item Parameters Characterize Properties of Samples	43
3.3.7	Complementary Information Provided by integIRTy and CNAmet	44
3.4	Discussion	44
3.5	Appendix	46
3.5.1	Dataset Assembly (OV, BRCA and GBM)	46
3.5.2	Data transformation	48
3.5.3	Simulation Study	48
3.5.4	Supplemental Figures	50
3.5.5	Supplemental Tables	64
<b>4</b>	<b>Bimodality identification from RNAseq Data</b>	<b>67</b>

4.1	Background	67
4.2	Methods	70
4.2.1	Mixture Models For RNAseq Count Data	71
4.2.2	Generalized Bimodality Index	73
4.2.3	Adjusting For Library Size and Gene Length Effect	74
4.3	Results	74
4.3.1	Simulation Study	75
4.3.1.1	NB, GP and LN Datasets	75
4.3.1.2	Effect of the Mixture Proportion	77
4.3.1.3	Performance Evaluation Metrics	78
4.3.1.4	Performance Under Correctly Specified Models	78
4.3.1.5	Performance Under Misspecified Models	79
4.3.1.6	Difficulty in Identifying the True Model	80
4.3.1.7	Robustness to Outlier Data	80
4.3.1.8	Comparison to Alternative Approaches	81
4.3.2	Real Data Analysis	82
4.3.2.1	LN Model Fits Best For RNAseq Data	82
4.3.2.2	Bimodal Genes Identified Using RNAseq Data	83
4.4	Discussion	84
4.5	Appendix	87
4.5.1	Derivation of the Generalized Bimodality Index	88
4.5.1.1	Normal Mixture with Unequal Variance	88
4.5.1.2	Negative Binomial Mixture	89
4.5.1.3	Generalized Poisson Mixture	90
4.5.1.4	Zero-Inflation	91
4.5.2	Investigation of Outlier Data	91
4.5.3	Comparison with COPA and PACK	92
4.5.4	Supplemental Figures	93
4.5.5	Supplemental Tables	109

<b>5</b>	<b>Bimodal genes contain most information for predicting outcome</b>	<b>110</b>
5.1	Background	110
5.2	Methods	112
5.2.1	Datasets	112
5.2.2	Identifying Bimodally Expressed Genes	115
5.2.3	Classification Methods	116
5.2.4	Performance Evaluation Metrics	117
5.3	Results	118
5.3.1	MAQC-II Binary-Class Data	118
5.3.2	Tan et al Multi-Class Data	118
5.3.3	Director's Challenge Lung Cancer Data	119
5.4	Discussion	121
<b>6</b>	<b>Conclusions and future research</b>	<b>124</b>
6.1	Conclusions	124
6.2	Future research	126
	<b>References</b>	<b>128</b>
	<b>Vita</b>	<b>143</b>

## List of Figures

2.1	Data assembly for TCGA OV data	16
2.2	Kaplan-Meier survival curves for TCGA ovarian cancer patients	17
2.3	Two-way clustering of 379 solid tumor samples for TCGA OV expression data	18
2.4	BUM plot of logistic regression model	20
2.5	BUM plot of Cox PH model	21
2.6	Example genes predictive of overall survival	24
2.7	Cancer related genes are not enriched with predictive genes	25
3.1	Illustration of Item Characteristic Curve (ICC)	33
3.2	Alteration Pattern Across Assays	35
3.3	Relations between integrated and individual gene lists in OV data	38
3.4	Example genes with discordant calls between conventional methods and the IRT method	41
3.5	Complementary information provided by <code>integIRTy</code> and <code>CNAmet</code>	43
3.6	Sensitivity analyses of latent trait estimates	51
3.7	Naive score is not invariant	52
3.8	Rank boosting from data integration	53
3.9	Original data for CDKN2A	54
3.10	Original data for VEGFC	55
3.11	Original data for STMN1	56
3.12	Comparison of the conventional method and the proposed method	57

3.13	Box-and-whisker plot of item difficulty	58
3.14	Comparison of <code>integIRTy</code> and <code>CNAmet</code>	59
3.15	Comparison of item difficulty estimates and truth in simulation study ( <i>Figure reprinted from Tong, P. et al, Bioinformatics, 2012</i> )	60
3.16	Comparison of item discrimination estimates and truth in simulation study ( <i>Figure reprinted from Tong, P. et al, Bioinformatics, 2012</i> )	60
3.17	Comparison of latent trait estimates and truth in simulation study ( <i>Figure reprinted from Tong, P. et al, Bioinformatics, 2012</i> )	61
3.18	Comparison of 1-way and all possible two-way integration	62
3.19	Comparison of all possible two-way and 3-way integration	63
4.1	Bimodal Index (BI) as a function of the size ( $\pi$ )	76
4.2	Robustness of NB, GP and LN models	77
4.3	RNAseq data best fit by LN model	84
4.4	Example bimodal genes from TCGA BRCA data	85
4.5	Performance under the correctly specified model	94
4.6	ROC curves under true models	95
4.7	Boxplot of fitted BI on simulated bimodal genes by the NB, GP and LN models stratified by component size	96
4.8	Cases where it is impossible to identify the data generating model by BIC	97
4.9	Robustness to heavy tailed distributions	98
4.10	Robustness to extreme values	98
4.11	Effect of heavy tailed distributions on the estimation of BI	99
4.12	Effect of extreme values on the estimation of BI	100
4.13	ROC curves for $BI_{LN}$ , PACK and COPA	101
4.14	PACK finds it difficult to detect bimodal genes with 20%-80% or 30%-70% split	102



4.15	COPA does not work well when bimodal expression is 50%-50% or 10%-90% split	103
4.16	COPA detects different set of bimodal genes as the quantile used for ranking changes	104
4.17	Comparison of robustness to heavy tailed distributions	105
4.18	Comparison of robustness to extreme values	106
4.19	Mean-variance relationship in the BRCA RNAseq data	107
4.20	Construction of curated gene list	108
5.1	Performance comparison of Naive Bayes, Bayesian Network and CART on Director's challenge lung cancer data	120
5.2	The dependency network learnt by Bayesian Network on Director's challenge lung data	121

## List of Tables

2.1	Numbers of significant genes selected at different FDR rates for the logistic regression models.	22
2.2	Numbers of significant genes selected at different FDR rates for the Cox PH models.	23
3.1	Number of patients per dataset ( <i>Table reprinted from Tong, P. et al, Bioinformatics, 2012</i> )	36
3.2	Latent trait and rank for top 20 genes selected by integrated analysis of TCGA OV data	40
4.1	Performance on NB datasets ( <i>Table reprinted from Tong, P. et al, Bioinformatics, 2013</i> )	80
4.2	Performance on GP datasets ( <i>Table reprinted from Tong, P. et al, Bioinformatics, 2013</i> )	81
4.3	Performance on LN datasets ( <i>Table reprinted from Tong, P. et al, Bioinformatics, 2013</i> )	82
5.1	Summary of MAQC-II dataset	114
5.2	Summary of Tan et al dataset	114
5.3	Summary of Director's Challenge Lung Cancer data	114
5.4	Result of MAQC-II dataset	119
5.5	Result of Tan et al dataset	119

## Abbreviations

AIC	Akaike's information criterion
BI	Bimodality Index
BUM	Beta-Uniform Mixture
CN	Copy Number
COPA	Cancer Outlier Profile Analysis
Cox PH	Cox Proportional Hazards
FDR	False Discovery Rate
GP	Generalized Poisson
ICGC	International Cancer Genome Consortium
IRT	Item Response Theory
LN	Lognormal
LRT	Likelihood Ratio Test
NB	Negative Binomial
OV	Ovarian serous cystadenocarcinoma
PACK	Profile Analysis using Clustering and Kurtosis

TCGA

The Cancer Genome Atlas

# Chapter 1

## Introduction

### 1.1 Background

Recent advances in biotechnology allow biologists to generate massive amounts of data, which is more than one could imagine decades ago. For example, it is routine to monitor the whole genome transcription level through various microarray and next generation sequencing platforms. Besides the transcriptome, many other aspects of cell activity are also frequently measured, including mutation, DNA methylation, DNA copy number change, microRNA expression, protein expression, and phosphorylation. Further, new technologies are still being developed that will make bioassays more diverse, powerful and inexpensive.

This leads to a rich body of biological information accessible through various public repositories. According to the update on Bioinformatics Links Directory [[Brazas et al., 2010](#)] and the review by Zhang [[Zhang et al., 2011](#)], there are around 1500 unique publicly available data sources which can be summarized into six categories: (1) sequence database such as GenBank [[Benson et al., 1997](#)], RefSeq [[Pruitt et al., 2009](#)] and CMR (Comprehensive Microbial Resource) [[Peter-](#)

son et al., 2001]; (2) functional genomics database including GEO (Gene Expression Omnibus) [Barrett et al., 2011], ArrayExpress [Parkinson et al., 2011] and FFGED (Filamentous Fungal Gene Expression Database) [Zhang and Townsend, 2010]; (3) protein-protein interaction database such as BIND (Biomolecular Interaction Network Database) [Bader et al., 2003], DIP (Database of Interacting Proteins) [Salwinski et al., 2004], IncAct [Aranda et al., 2010] and MINT (Molecular Interactions Database) [Ceol et al., 2010]; (4) pathway database such as KEGG (Kyoto Encyclopedia of Genes and Genomes) [Kanehisa et al., 2010]; (5) structure database such as CATH (Class Architecture Topology Homology) [Greene et al., 2007] and PDB (Protein Data Bank) [Rose et al., 2011]; (6) annotation database such as GO (Gene Ontology) [Ashburner et al., 2000] and NCBI Taxonomy [Sayers et al., 2011].

Given the technology advancement as well as the rich information provided by public databases, data integration becomes an indispensable component for biomedical research due to at least two reasons: (1) most of the research effort becomes the analysis and interpretation of data rather than data generation because of the high level of automation in data generation. This is especially true for projects involving next generation sequencing technology where approximately four fifths of the effort goes to the integration and analysis of the collected data [Mardis, 2010], and (2) the answers to most biological questions are rarely provided directly by the experimental results. Downstream bioinformatics analysis involving integrating diverse data sources is required.

Many techniques and systems have been exploited for integrating biomedical data. As summarized in [Goble et al., 2008], current approaches for data integration can be roughly grouped into five groups: data warehousing, service-oriented integration, semantic integration, wiki-based integration, and hypothesis-driven integration. Data warehousing aims to provide a “one-stop shop” access to different but related data sources. Usually a pre-defined data model is needed to

extract, clean and formulate data from existing sources. Data warehousing suffers from frequent data updates. In contrast, service-oriented integration leverages the power of web services where individual data sources agree to open their data via web services and thus data integration becomes a communication between computers over the web. Most web pages are created for human reading which are not efficient for a computer to understand. Therefore, semantic integration that uses semantic web standards as a universal medium for data exchange has been proposed. To allow user participation and contribution, the wiki-based integration becomes necessary. Finally, to incorporate domain knowledge, hypothesis-driven integration is needed which explicitly makes assumptions about the data and applies statistical approaches for data integration.

This thesis mainly focuses on hypothesis-driven integration. The primary question we are trying to address is how to extract biological insights from multiple high throughput biological assays profiled on the same set of samples. In particular, we are interested in identifying biomarkers and building accurate classifiers by integrating information from different assay types. The final goal of our analysis would be to formulate testable hypothesis suggesting follow-up studies.

## 1.2 Related work and motivation

It is widely agreed that tumorigenesis is a multi-step procedure that involves aberrant functioning of genes regulating various aspects of cell proliferation, differentiation, apoptosis, genome stability, angiogenesis and motility. To obtain a full picture of cancer, we need to gather information on all aspects of cell activity. The Cancer Genome Atlas [[McLendon et al., 2008](#)] project has taken the initiative to profile more than twenty cancers with almost all existing biological assays including mutation, gene expression, DNA methylation, DNA copy num-

ber (CN), microRNA expression and protein expression. However, data collection is only the first step towards curing cancer. Extracting biological insights from this comprehensive dataset through integrated analysis is a major challenge.

Below we review some of the mostly widely used data integration approaches developed in the last decade. In terms of the adopted procedure for data integration, current approaches can be classified into four categories: step-wise, regression-based, correlation based, and latent variable models [Lahti et al., 2012, Huang et al., 2012]. The last three methods jointly model different assay types and hence are called joint methods according to Huang et al. [2012]. Step-wise methods analyze the individual assay type and then manually combine the results; joint modeling specifies a model, usually in the form of a linear model or latent variable model, to combine evidence from different sources before making inferences.

In addition to the various procedures used, existing data integration methods also differ in their analysis goals. Many current methods focus on the dependency between gene expression and CN. These include the correlation and regression based methods that explicitly search for genes with correlated measurements. For example, Menezes et al. [2009] applied linear mixed models to identify genes whose expression is regulated by CN change. In Peng et al. [2010], the dependence between RNA expression and DNA copy number change is modeled through penalized multiple regression models. There are also methods that identify overlapped genetic alterations. The SODEGIR (Significant Overlap of Differentially Expressed and Genomic Imbalanced Regions) method is designed to infer genomic regions with both differential expression and copy number change [Bicciato et al., 2009]. The CONEXIC (COpy Number and EXpression In Cancer) method aims to identify driving mutations and the affected biological processes [Akavia et al., 2010]. The remMap method searches concomitant gene expression and CN alteration in cancer [Peng et al., 2010]. It is also possible to simulta-



neously integrate gene expression, CN and clinical data. For example canonical correlation analysis has been used to identify associations among gene expression, CN change and clinical outcome [Waaijenborg et al., 2008, Lê Cao et al., 2009, Witten et al., 2009]. The integrated classification problem is still emerging in high throughput data analysis and hence has not been well studied.

Most existing integration methods deal with two assay types such as GE and CN. The CNAmet method advances by simultaneously integrating GE, CN and methylation data [Louhimo and Hautaniemi, 2011]. Still, due to the high degree of heterogeneity in the data, existing approaches are not flexible enough to integrate an arbitrary number of assay types. This motivates us to develop a more general approach for data integration. The first approach we use is through regression where different assay types enter the regression model as covariates. This approach enables us to evaluate how predictive a gene is by combining information from diverse sources. Our second approach is similar in concept to Louhimo and Hautaniemi [2011] where the integration is performed on binary signals derived from the original data. By integrating data from derived binary signals, we gain several benefits: (1) data integration becomes more flexible; (2) the binary signals are easy to interpret and understand; (3) the implementation and inference becomes simpler. Under this framework, we have developed several methods. For example, we have developed the **integIRTy** pipeline which is able to integrate an arbitrary number of assay types [Tong and Coombes, 2012]. We have also developed the **SIBER** algorithm which systematically extracts binary signals from the data [Tong et al., 2013]. We also formally investigate how well the binary signals perform in terms of predicting clinical outcome and established that bimodal genes contain the same information as provided by all genes. The effectiveness of building classifiers from discrete signals will greatly facilitate integrated classification using multiple data sources.

### 1.3 Thesis organization and contributions

This thesis focuses on approaches for the integrative analysis of genomic assays profiled on the same set of samples. We have developed several novel methods to solve problems not addressed by existing approaches.

We begin with integrative biomarker identification in Chapter 2. We first propose a regression framework to integrate multiple assays including gene expression, methylation, and copy number data. We discuss the dependency problem where measurements from different assays are correlated violating the standard regression assumption, and we propose a penalized regression approach to obtain accurate inference. The proposed model is applied to the TCGA (<http://cancergenome.nih.gov/>) ovarian serous cystadenocarcinoma (OV) datasets and identifies a set of genes predictive of treatment response and overall survival. We find that known cancer related genes are not enriched for predictive genes.

We then introduce the `integIRTy` method in Chapter 3 to identify gene alterations from multiple assays using the Item Response Model. This is another way to identify biomarkers through data integration. This method is motivated by the fact that tumor suppressors can be blocked (or oncogenes activated) by different mechanisms in different patients. Hence, simply looking at one assay at a time will miss genes that alter rarely in individual assay but in a consistent manner across assays. After extensive simulation and real data analysis, we find that `integIRTy` is more robust and reliable than conventional methods when applied to a single assay. When applied to multiple assays, `integIRTy` can identify novel genes that cannot be found by looking at individual assays separately. Further, `integIRTy` allows us to explore the global alteration pattern across multiple assays.

Chapters 4 and 5 provide the foundation for integrative classification

using information from multiple sources. We base the integration on discrete signals. Differing from most current approaches which model continuous signals, we find that discrete signals such as bimodal expression and discrete copy number changes are effective and easy for data integration. We start with bimodality which identifies natural binary signals in the cell. Chapter 4 deals with how to identify bimodal genes from RNAseq data. We present the Bimodality Index (BI) approach which generalizes the existing method by Wang et al. [2009] developed for microarray data based on mixture model. The generalized BI proves to be robust, powerful, invariant to shifting and scaling, has no blind spots, and has a sample-size-free interpretation.

Chapter 5 addresses the question: are bimodal genes enough for prediction? This question is important because it is the basis for discrete-scale integration. We approach this problem by assembling an established benchmark dataset, and we compare the classification performance between bimodal-gene-only model and all-gene model. We find there is no significant difference between the two models and conclude that bimodal genes contain all the information needed to predict outcome. These results pave the road for a comprehensive study that performs classification with discrete features extracted from different data sources.

Finally, Chapter 6 concludes the thesis and discusses several future research directions.

## Chapter 2

# Prognostic biomarker identification through data integration

### 2.1 Background

Biomarkers play a crucial role in medicine. The usage of biomarkers enables more accurate diagnosis, prognosis, and more effective treatment. For example, predictive biomarkers give indications of the probable effect of a certain treatment. These include drug-related biomarkers that indicate whether a drug is likely to be effective on a specific patient. Prognostic biomarkers provide information on how a disease may develop. It is thought that genetic biomarkers are the key to personalized medicine [Tevzak et al., 2010]. The great promise of biomarkers has led many organizations and big pharmaceutical companies to invest heavily in biomarker and drug development.

It turns out that biomarker identification is a central component for drug development. With the completion of the Human Genome Project, biomedical research has advanced tremendously in the past ten years. Rather than measure cell activities one at a time, it has now become easy and cheap to monitor

genome-wide events thanks to the evolving high throughput technologies that cover DNA, mRNA, protein and metabolites. It is expected that these biotechnologies will usher in a paradigm shift in genomic medicine where patients can receive personalized treatment tailored to their genetic composition. There is no doubt that data can be collected at an unprecedented pace. However, the challenge becomes how to analyze this data and transform it into knowledge. It becomes a more serious problem when deriving candidate biomarkers since it is almost impossible to follow up every target that is measured. Further, it is one thing to identify therapeutic candidates through these high throughput assays, but it is another to have these biomarkers going through clinical trials and being marketed.

There are two major methods for biomarker identification: filter and wrapper [Inza et al., 2004]. The filter method selects biomarkers by examining the relevance of the features to the outcome. Usually this is in the form of statistical tests (e.g., student’s t-test or F test) or information metric e.g., information gain or mutual information) [Liu and Motoda, 1998]. Feature selection by the filter method is separated from evaluating the prediction model. In comparison, the wrapper method embeds feature selection into the prediction model. Wrappers train a new model for each subset of features and scores the feature subset with the prediction performance. As a result, wrappers can produce a feature set that is tuned to a specific predictive model and usually yield better accuracy in industrial machine learning applications. However, wrappers are computationally intensive and more likely to overfit for high throughput data. Instead, filter methods are mostly adopted in the analysis of high throughput data [Chu et al., 2005, Inza et al., 2004].

Here we propose a regression framework that integrates information across different types of assays for biomarker identification. This method is an example of a filter method that is flexible enough to deal with both binary and survival

outcome. The rest of this chapter is organized as follows. We first introduce the multiple logistic regression and Cox proportional hazard models in section 2.2. We discuss the collinearity issue arising from the correlated measurements as well as methods to solve it. We then present the results of applying this method to identify prognostic markers in TCGA ovarian cancer data in Section 2.3. We finish this chapter with a brief summary and discussion of future research.

## 2.2 Methods

We formulate a multiple regression framework for integrated biomarker identification. The goal is to identify genes predictive of therapy response (complete response/non-complete response) or overall survival time. This is a generalization of the commonly used student's t-test and univariate Cox Proportional Hazards (PH) regression models applied to one assay type.

### 2.2.1 Logistic regression model

In the first model, we investigate the relationship between the binary therapy response and measurements from gene expression, methylation and copy number assays. For individual  $i$  ( $i = 1, 2, \dots, N$ ), the measurements for a particular gene are denoted by  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik}, \dots, x_{iK})$  where  $x_{ik}$  is the measurement from the  $k^{th}$  assay ( $k = 1, 2, \dots, K$ ). Note that  $\beta_0$  is the intercept term. In our case, there are two expression assays, one methylation assay, and one copy number assay, and hence  $K = 4$ . The therapy response is denoted by  $\mathbf{y}_i = (y_1, y_2, \dots, y_N)$  where  $y_i = \{0, 1\}$ .

We apply the following logistic regression model:

$$\log \frac{p_i}{1 - p_i} = \sum_{k=0}^{k=K} \beta_k x_{ik} \quad (2.1)$$

where  $p_i = P(y_i = 1)$  is the probability of achieving complete response for individual  $i$  ( $i = 1, 2, \dots, N$ ).  $\beta_k$  for  $k = 1, 2, \dots, K$  is the regression coefficient for the  $k^{th}$  assay. Note that the gene index is suppressed in this formula.

The expression, methylation, and copy number measurements are expected to be correlated, which violates the independence assumption for multiple regression. Several approaches have been proposed to deal with this issue including LASSO regression [Tibshirani, 1996] and ridge regression [Marquardt, 1970].

Given a set of predictors, it is desirable to identify which set of variables predicts best. Therefore, a model selection procedure is needed. One common practice is to apply a stepwise selection procedure that can be either forward selection, backward elimination, or bidirectional selection. In our implementation, we adopt backward elimination and use (AIC) to select the best model among the candidate models.

The parameters  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$  for the logistic regression model in equation 2.1 can be estimated by maximizing the following log-likelihood function:

$$\ell_1(\beta) = \sum_{i=1}^{i=N} \left\{ y_i \ln \left( \frac{1}{1 + \exp(-\mathbf{x}_i' \beta)} \right) + (1 - y_i) \ln \left( \frac{\exp(-\mathbf{x}_i' \beta)}{1 + \exp(-\mathbf{x}_i' \beta)} \right) \right\} \quad (2.2)$$

The LASSO version of logistic regression is nothing but adding an  $L_1$  penalty on the regression coefficients. This corresponds to maximizing the following constrained log-likelihood equation:

$$\ell_1(\boldsymbol{\beta})_{LASSO} = \ell_1(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{k=K} |\beta_k|$$

where  $\lambda > 0$  is a Lagrangian multiplier. The optimal  $\lambda$  can be obtained through cross-validation.

The Ridge regression model is quite similar. We only need to add an  $L_2$  penalty on the regression coefficients:

$$\ell_1(\boldsymbol{\beta})_{Ridge} = \ell_1(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{k=K} \beta_k^2$$

### 2.2.2 Cox Proportional Hazards model

To investigate the relationship between overall survival time and whole-genome assays, we apply the Cox PH model. For individual  $i$ , we observe  $(t_i, \delta_i)$  where  $t_i$  is the overall survival time and  $\delta_i$  is the censoring indicator ( $\delta_i = 1$  means no censoring;  $\delta_i = 0$  means censoring.)

$$h_i(t|\mathbf{x}_i) = h_0(t) \exp\left(\sum_{k=1}^{k=K} \beta_k x_{ik}\right) \quad (2.3)$$

where  $h_0(t)$  is the baseline hazard function.  $h_0(t)$  can have a parametric form or remain unspecified leading to a semi-parametric model. This model assumes the proportional hazard condition, which means that the hazard ratio between two individuals is independent of time and only time-independent covariates are allowed.

The parameters  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$  for the Cox PH model are usually estimated by maximizing the partial log-likelihood without specifying  $h_0(t)$ :



$$\ell_2(\boldsymbol{\beta}) = \sum_{i=1}^{i=N} \delta_i [\mathbf{x}_i' \boldsymbol{\beta} - \log \{ \sum_{j \in R(i)} \exp(\mathbf{x}_j' \boldsymbol{\beta}) \}] \quad (2.4)$$

where  $R(i)$  is the set of indices for individuals that are at risk at time  $i$ .

To impose LASSO type penalty on the regression coefficient, we modify the partial log-likelihood as:

$$\ell_2(\boldsymbol{\beta})_{LASSO} = \ell_2(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{k=K} |\beta_k|$$

The Ridge regression is quite similar:

$$\ell_2(\boldsymbol{\beta})_{Ridge} = \ell_2(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{k=K} \beta_k^2$$

Both the LASSO and Ridge versions for the logistic regression and Cox PH regression models can be fit through the `glmnet` package in R.

### 2.2.3 Hypothesis testing

Given a fitted model  $M_1$  which may include a single covariate or multiple covariates, we want to test if this model is better than a null model  $M_0$  where only an intercept term is fitted. Under the generalized linear model framework, this can be done through the Analysis of Deviance [[McCullagh and Nelder, 1989](#)]. The deviance of model  $M_1$  is defined as:

$$D(M_1) = -2\{\log L(M_1) - \log L(M_s)\}$$

where  $L(M_1)$  is the likelihood fitted by model  $M_1$  while  $L(M_s)$  is the likelihood from the saturated model  $M_s$ . The saturated model  $M_s$  is the model that fits a parameter for each observation and hence fits the data exactly.

Similarly, the deviance of null model  $M_0$  which only fits an intercept term to the data can be defined as:

$$D(M_0) = -2\{\log L(M_0) - \log L(M_s)\}$$

It then follows that the difference of deviance  $D(y)$  follows a  $\chi^2$  distribution with a degree of freedom equal to the number of extra parameters  $d$  in  $M_1$  compared to  $M_0$ :

$$D(y) = D(M_0) - D(M_1) \sim \chi_d^2$$

Note that it is not needed to evaluate  $\log L(M_s)$  in computing  $D(y)$  since  $\log L(M_s)$  is cancelled out:

$$D(y) = -2\{\log L(M_0) - \log L(M_1)\}$$

Since we are testing each of the genes separately, we have to deal with the multiplicity of simultaneous tests. Many methods have been proposed to account for multiplicity. Usually these methods control different type I error rates such as family-wise error rate, false discovery rate, per-comparison error rate, or per-family error rate [Dudoit et al., 2003]. We adopt the method developed by ? that controls the False Discovery Rate (FDR) . In particular, the set of p values is modeled by a Beta-Uniform Mixture (BUM) model where the uniform component represents non-informative genes while the beta component corresponds to the set of predictive genes.

## 2.3 Results

### 2.3.1 Cancer subtypes and prognosis

We performed an exploratory data analysis to examine if there were obvious subtypes among the samples. We focused on TCGA (<http://cancergenome.nih.gov/>) ovarian serous cystadenocarcinoma (OV) data, as this data contained the most samples at the time of analysis in March 2010.

We started with data assembly. For genewise integration, four assay types are available: Affymetrix U133A expression (BI HT\_HG-U133A), Illumina Infinium 27K methylation arrays (JHU-USC HumanMethylation27), Agilent CN arrays (HMS HG-CGH-244A), and Agilent expression arrays (UNC AgilentG4502A\_07). We obtained genewise summary for CN data by mapping the segments to the human genome. The details are provided in Section 3.5.1. We restricted our attention to the solid tumor samples simultaneously measured by all platforms. There were four types of tissues selectively profiled for the OV data at the time of analysis: solid tumor (coded as 01 by TCGA consortium), normal tissue (11), cell line (20) and normal blood (10). Most of the profiled samples were from the solid tumor tissue. As a proof of concept, we aimed to integrate across all platforms and thus only focused on genes measured in all assays. At the end, we got 207 shared solid tumor samples and 9855 genes for integration (see Figure 2.1).

Measurements ( $\beta$  values) from the methylation data are bounded between 0 and 1. Measurements from the CN data are log2 intensity ratios ( $\log_2 R$ ) between two channels. For easy interpretation, we categorize the methylation data into three groups:  $\beta < 0.25$  (no methylation),  $0.25 \leq \beta < 0.75$  (partial methylation),  $\beta \geq 0.75$  (complete methylation). We also categorize the CN data into three groups:  $\log_2 R < -0.35$  (loss),  $-0.35 \leq \log_2 R < 0.2$  (neutral),  $\log_2 R \geq 0.2$  (gain)

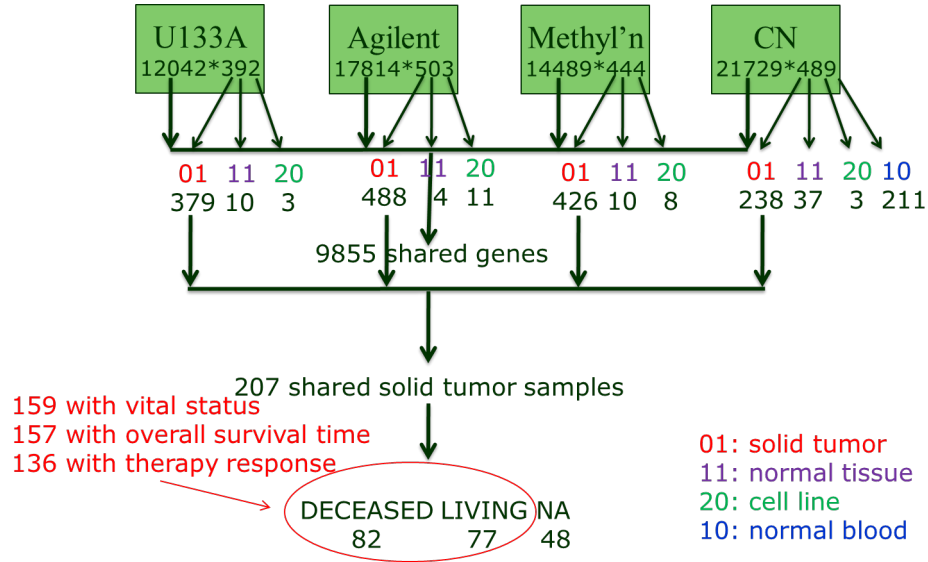


Figure 2.1: Data assembling for TCGA OV data. TCGA level 3 data for gene expression (U133A, Agilent), methylation and copy number (CN) are downloaded from <http://cancergenome.nih.gov/> and assembled by breaking down the tissue types. Clinical information including vital status, overall survival and therapy response is also indicated.

based on exploratory analysis.

We also examined the clinical data. The clinical file contained basic information about the patients including age, sex, ethnicity group, tumor grade, primary therapy response and survival information. The primary therapy response fell into four categories: complete response, partial response, progressive disease and stable disease. We found that the survival for partial response, progressive disease and stable disease is quite similar. Therefore, we grouped the three categories into non-complete response. Figure 2.2 shown that the overall survival differed between complete response and non-complete response. It is therefore important to examine whether there are markers that predict complete response/non-complete response status.

An immediate question to ask is whether there are subtypes among the OV patients. Further, we would like to evaluate if the patients with complete response and non-complete response form natural groups. To answer these questions, we performed two-way clustering using the Affymetrix expression data with all 379 solid tumor samples. We selected genes with the BI method which led to

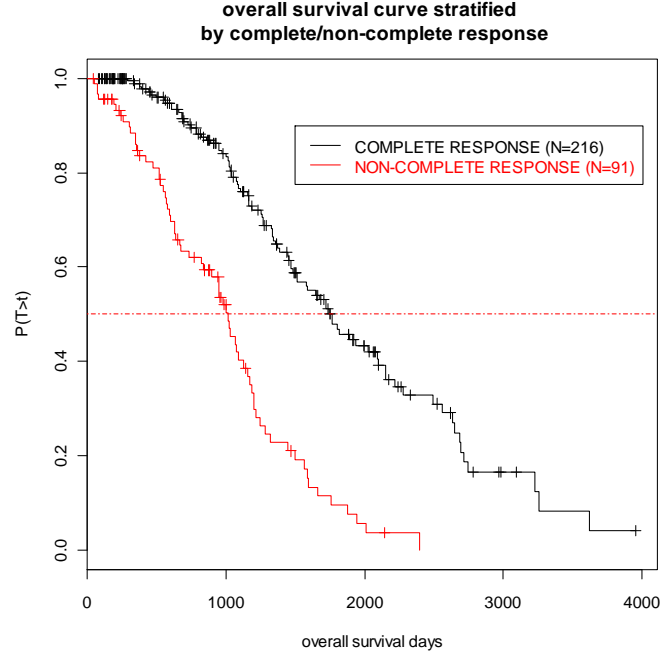


Figure 2.2: Kaplan-Meier survival curves for TCGA ovarian cancer patients. For patients with primary therapy response status, 216 have complete response while 91 have non-complete response. The median survival time is indicated by the red dashed line.

1130 bimodal genes [Wang et al., 2009]. Figure 2.3 shows the dendrogram from two-way clustering. The samples form three groups, while there are at least four groups of genes. We split the samples based on the clustering result and examined if this grouping is associated with therapy response status or overall survival. Unfortunately, this analysis did not find any significant association between the patient clusters and clinical outcome (therapy response or survival time). For the four groups of genes, we queried the DAVID (<http://david.abcc.ncifcrf.gov/>) database and found that the second group indicated in Figure 2.3 was significantly associated with immune response ( $p$  value= $10^{-31}$ ) by GO term enrichment analysis.

The patient clusters using all bimodal genes did not associate with either therapy response or overall survival time. We conjectured that clusters derived from a subset of the genes are correlated with clinical outcome. We therefore split the genes into four groups based on the gene clustering result and repeated our analysis. Unfortunately, the clusters formed by subsets of genes still do not

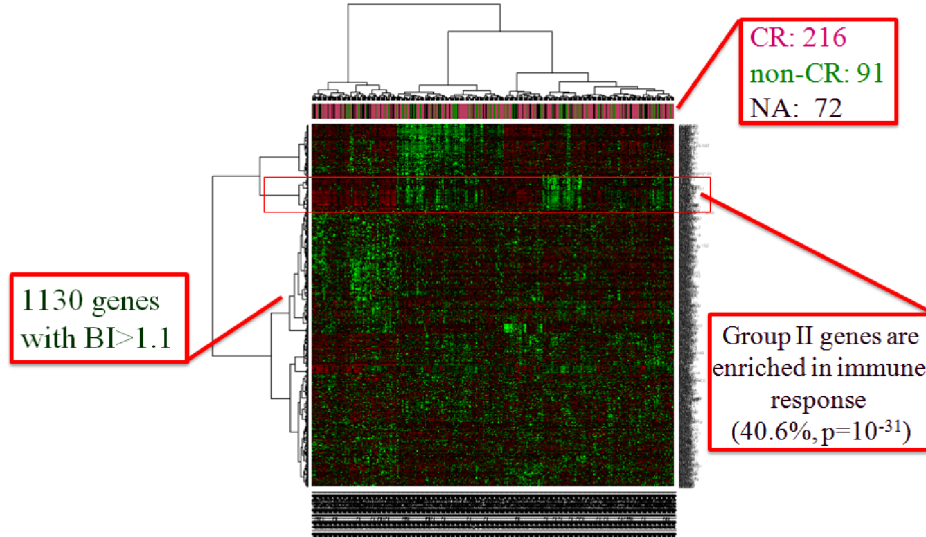


Figure 2.3: Two-way clustering of 379 solid tumor samples for TCGA OV expression data

predict outcome. This suggests that a supervised learning approach is needed to derive predictive markers and signatures.

### 2.3.2 Predictive power of different assay types

Our logistic regression model, as well as the Cox PH model, can identify biomarkers (either expression, methylation or copy number) predictive of clinical outcome. After fitting the regression models, we obtain a p value for each gene. For a direct comparison of the predictive power of different assays as well as the combined data, we first fit univariate regression model only allowing one covariate. We then compared the results to a full model where all measurements enter the model. We found that the LASSO and ridge regression behave similarly. Therefore, we only present results from ridge regression.

Figure 2.4 shows the results from our logistic regression models. In each panel, we show the histogram with a fitted Beta-Uniform Mixture (BUM) model superimposed [Pounds and Morris, 2003]. The blue line indicates the uniform component while the green line indicates the beta component. The uniform component corresponds to no predictive power while the Beta component suggests

different degrees of predictive power depending on its shape as well as the component size. We find that the four assay types have varying predictive power for therapy response. In particular, the Affymetrix expression data (Expr-U133A) has an obvious Beta component while the Agilent expression data (Expr-Agilent) seems to be pure noise. This however does not indicate Affymetrix is more accurate than Agilent because there are other issues beyond the technology itself. We start with the level 3 data which is preprocessed and summarized measurements. Therefore, the difference might come from data preprocessing. Another possible reason is batch effects.

Comparing across assay types, the copy number data seems to be more predictive than either expression or methylation data. The histogram for methylation data is somewhat irregular. The integrated analysis from the multiple regression model improves the predictive power, which is not surprising since more predictors are included. The last panel in Figure 2.4 shows the result after applying model selection. The stepwise model selection procedure distorts the distribution of p values by deliberately searching for the best model satisfying the AIC criterion. This leads to a null model where no predictor is included for most of the genes (p values close to 0, not shown).

Similarly, Figure 2.5 shows the results from our Cox PH regression models. In this case, the two expression assays behave similarly. The copy number data becomes least predictive. The predictive power for methylation data is also marginal. Compared to results in Figure 2.4, the predictive power for each assay seems to have been changed. This is because the therapy response and overall survival are different outcomes despite being correlated.

The histogram of p values obtained from the methylation data looks bizarre in both the logistic regression model and the Cox PH model. In particular, we notice that there is a high bar in the histogram of p values (p values around

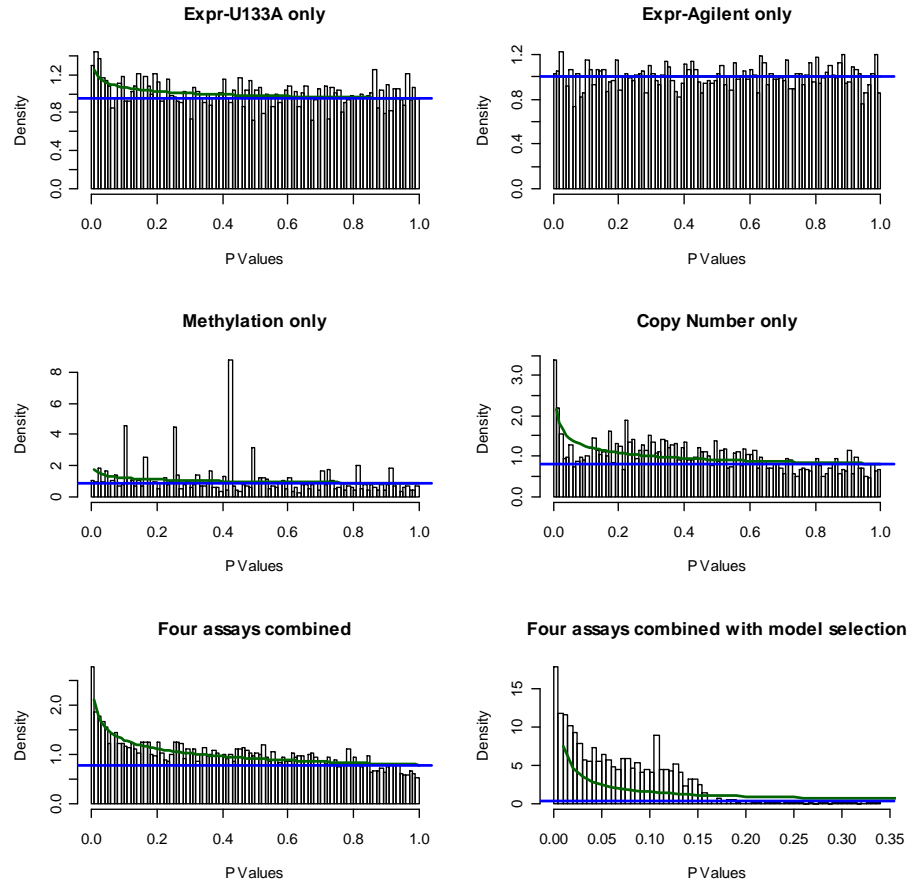


Figure 2.4: BUM plot of logistic regression model. The first four figures show the histograms of p values from the univariate logistic regression models. The fifth figure shows the histogram of p values from our multiple regression models. The last figure shows the result after applying stepwise model selection by AIC. A Beta-Uniform Mixture model (BUM) is fitted for each of the histograms and indicated by the blue and green lines. Note that the BUM model on the last panel is inappropriate due to the embedded multiple testings during model selection.



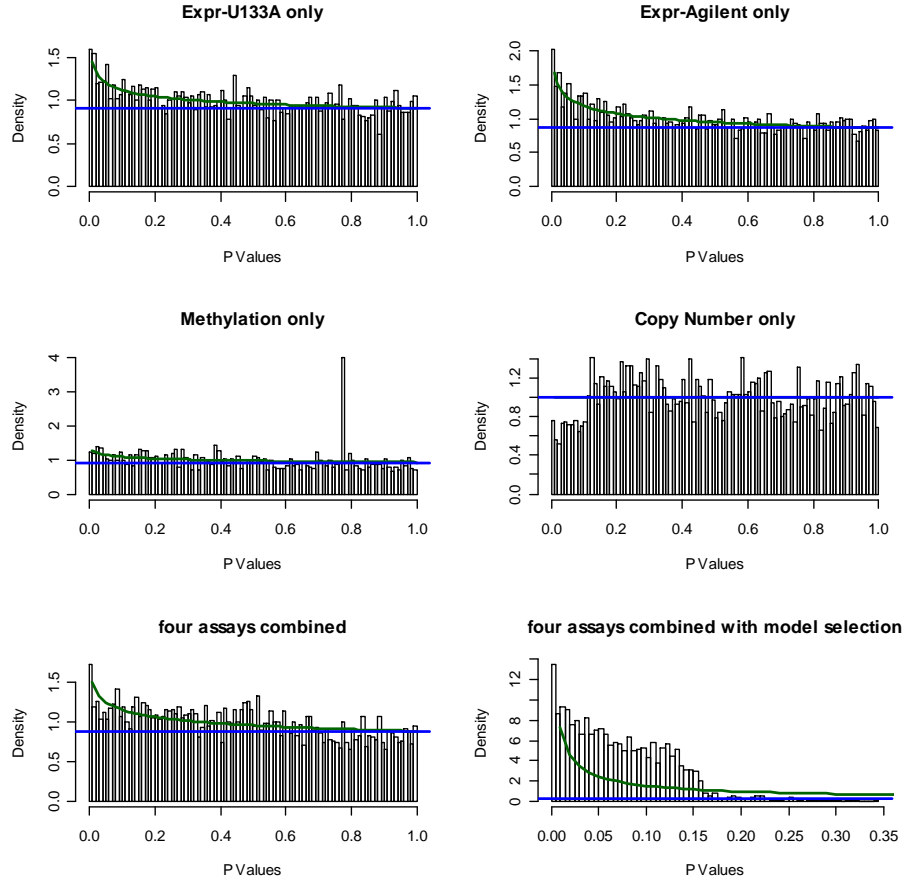


Figure 2.5: BUM plot of Cox PH model. The first four figures show the histogram of p values from univariate Cox PH regression model. The fifth figure shows the histogram of p values from Cox PH model with all covariates. The last figure shows the result after applying stepwise model selection by AIC. A Beta-Uniform Mixture model (BUM) is fitted for each of the histograms and indicated by the blue and green lines. The spike of p values around 0.77~0.78 in methylation data is because around 200 genes have

0.42~0.43 in the logistic regression model and 0.77~0.78 in the Cox PH model). The high bar in the logistic regression model corresponds to 452 genes having exactly the same p value (there are 41 other genes having similar p value in the bar). The reason is that for these 452 genes, 135 patients have been partially methylated and one patient has no methylation. As a result, these genes have the same p value in the logistic regression model. Similarly, the high bar in the histogram of p values from the Cox PH model mostly corresponds to 204 genes. For these 204 genes, 156 patients have no methylation and 1 patient have partial methylation. Note that the sample size differs in the logistic regression model and the Cox PH model since there are missing values in the clinical data.

# selected	U133A	Agilent	Methylation	CN	Combined
FDR=0.1	0	0	0	23	3
FDR=0.2	0	0	1	174	62
FDR=0.3	0	0	6	409	266
FDR=0.4	1	0	40	780	761
FDR=0.5	9	0	203	1336	1683
FDR=0.6	28	0	605	2681	3362
FDR=0.7	137	0	1969	5324	6068
FDR=0.8	667	0	4173	8752	9216
FDR=0.9	3366	0	5592	9214	9216
FDR=1	9216	9216	5592	9214	9216
Bonferroni	0	0	0	0	0

Table 2.1: Numbers of significant genes selected at different FDR rates for the logistic regression models.

The BUM model allows us to control for multiple testing using False Discovery Rate (FDR). Table 2.1 and Table 2.2 list numbers of genes selected at different FDRs using Affymetrix arrays (U133A), Agilent arrays (Agilent), methylation, copy number (CN) and integrated data (combined). The numbers of genes remaining after Bonferroni correction are also indicated. Note that for the integrated data, the FDR estimate is dubious since there is inherent multiple testing during model selection which violates the assumption of BUM model. More strict control of multiple testing for the integrated data could be attained using permutation.

According to Table 2.1, CN data tends to identify more genes related to therapy response. However, even at FDR=0.1, there are only 23 genes identified which indicates there are not many predictive genes for therapy response. Note that at FDR=1, methylation and CN do not select all genes. The reason is that there are some genes with all methylation or CN data falling into only one group for the categorized data. There is no gene identified after Bonferroni correction.

Table 2.2 summarizes the results from the Cox PH models. There is no significant gene identified by CN data using any FDR cutoffs. Expression (both Affymetrix and Agilent) data identifies the most genes significantly associated with overall survival.

# selected	U133A	agilent	methylation	CN	Combined
FDR=0.1	0	0	1	0	0
FDR=0.2	2	4	1	0	0
FDR=0.3	5	26	1	0	2
FDR=0.4	27	120	1	0	32
FDR=0.5	81	402	2	0	119
FDR=0.6	295	1080	11	0	429
FDR=0.7	888	2665	123	0	1455
FDR=0.8	2661	5794	766	0	4351
FDR=0.9	8106	9216	3685	0	9216
FDR=1	9216	9216	5666	9214	9216
Bonferroni	1	0	1	0	0

Table 2.2: Numbers of significant genes selected at different FDR rates for the Cox PH models.

Example genes predictive of overall survival time are shown in Figure 2.6. We select example genes from the top genes from integrated analysis that are also contained in the sequencing list. For gene SLC2A5 (ranked 9th in the integrated analysis using our Cox PH model), increased methylation leads to improved survival. For gene SKP2 (ranked 6th in the integrated analysis using our Cox PH model), increased CN improves survival. However, we should interpret this result with caution since the result is based on the training data. A more rigid evaluation should be performed on independent validation data.

### 2.3.3 Cancer genes and predictive power

We have identified genes predictive of therapy response and overall survival. Beyond this, there are genes that are deemed to be cancer related, for example the oncogene and tumor suppressors. This prior knowledge of “cancer genes” can be used as a reference to examine whether biologically important genes are also predictive of clinical outcome. We extract a cancer gene list from the targeted sequencing candidates. The targeted sequencing candidates consisting of 1326 genes were curated by TCGA consortium to identify important mutations due to their biological importance. As a result, these genes are likely to

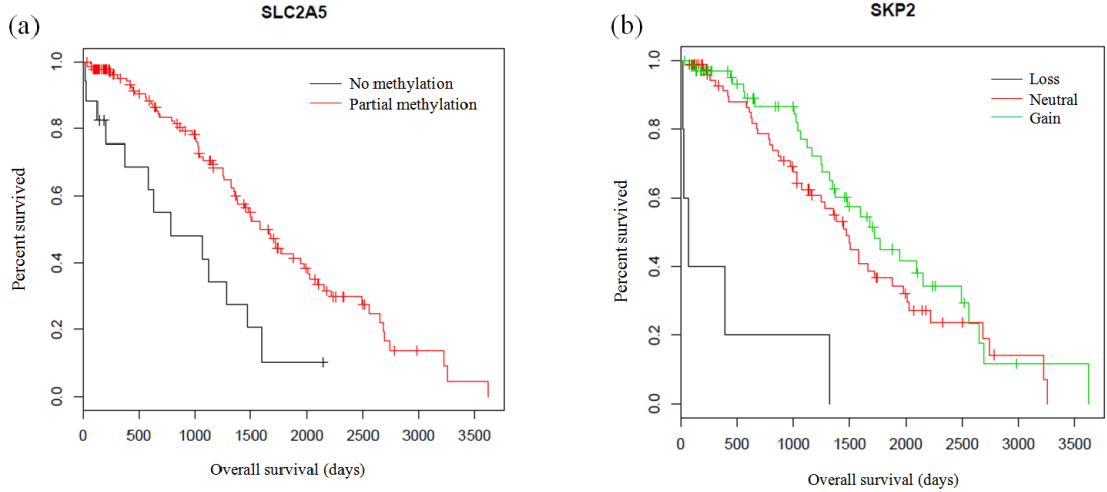


Figure 2.6: Kaplan-Meier curves for example genes predictive to overall survival. (a) The two groups of samples defined by methylation status (no methylation and partial methylation) separate well in terms of Kaplan-Meier survival curve for gene *SLC2A5*. Increased methylation leads to improved survival. (b) The three groups of samples defined by CN status have different survival curves in gene *SKP2*. Increased CN leads to improved survival.

be key drivers for cancer. We construct three gene lists: (1) genes predictive of therapy response (denoted as logistic in Figure 2.7); (2) genes predictive of overall survival (denoted as coxph); (3) the cancer related genes (denoted as sequenced). The sequenced gene list is pre-determined; the logistic and coxph gene lists are defined as genes passing a particular FDR cutoff. To test if the three gene lists are independent, we apply a  $\chi^2$  test to every combination of two gene lists. Figure 2.7 shows the relationship between  $\chi^2$  p value and FDR. We find that genes predictive of therapy response or overall survival rarely overlap with cancer genes. A possible explanation for this would be that the cancer gene list is mostly based on mutation data. Therefore, to get improved enrichment, we should use mutation data rather than expression, methylation or copy number for the association analysis. The genes predictive of therapy response overlap with genes predictive of overall survival at high FDR rate. At low FDR rates, there is only marginal overlap between the two predictive gene lists. In fact, at  $\text{FDR} < 0.3$ , there is so little overlap that the 2-by-2 table degenerates to a 2-by-1 table, and hence no p value is computed.

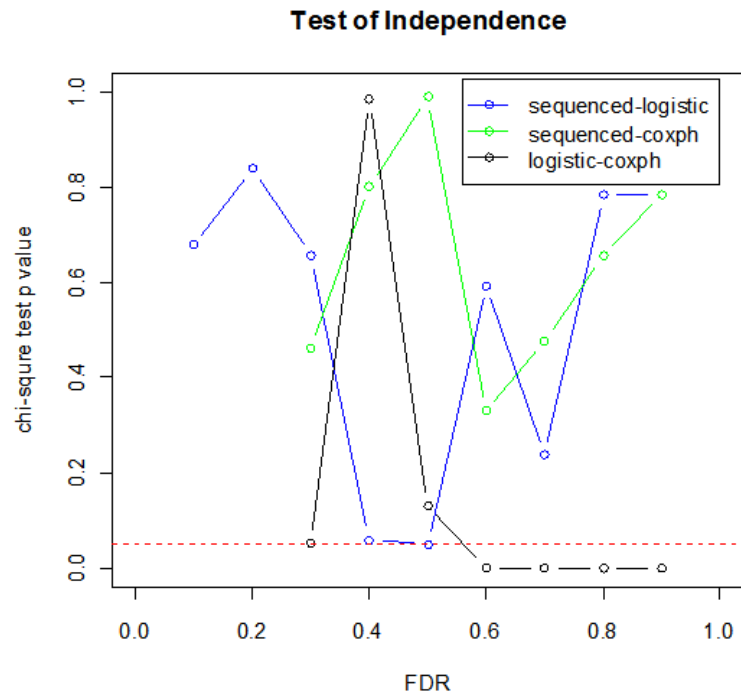


Figure 2.7: Test of independence between predictive gene list and cancer related gene list. The cancer related genes are defined as the genes selected for targeted sequencing. The predictive gene list is defined as genes passing a given FDR. We test if the sequenced gene list is independent of the selected predictive gene list with  $\chi^2$  test. The x-axis shows FDR cutoff while y-axis shows the  $\chi^2$  test p value. Blue line indicates the p value from testing sequenced gene list and predictive gene list to therapy response. The horizontal red line represents the 0.05 significance cutoff.

## 2.4 Discussion

In this chapter we have presented a flexible framework for gene-wise integration using multiple biological assays based on regression. This method allows us to evaluate the predictive power of individual assays as well as the combined data. To pick the set of most predictive measurements and remove measurements that are non-informative, we propose stepwise model selection. An application to the TCGA ovarian cancer data is illustrated. Our analysis shows that the gene expression, methylation and copy number have different power to predict either therapy response status or overall survival. Further, the genes predictive of therapy response also differ from genes predictive of overall survival. We find that the curated cancer gene list downloaded from TCGA data portal is not enriched with predictive genes. The underlying reason might be that the cancer gene list is constructed based on mutation information and hence enrichment is only expected if the association analysis is based on mutation data.

We have applied the Beta-Uniform Mixture model to deal with multiplicity in testing thousands of genes. However, this approach does not work when model selection is applied. The inherent multiple testing in stepwise regression model would deflate the p value that makes the BUM model inappropriate. One possible solution would be permutation based correction as described in [Dudoit et al. \[2003\]](#).

Our approach for integrative biomarker identification is performed at the gene level. Extension work would perform integration at the network level. Network based biomarkers have many attractive features. It has been reported that network based biomarkers can be more predictive than gene based biomarkers [Chuang et al. \[2012\]](#). Besides, network based biomarker can illuminate the relationship between individual genes and hence suggest underlying regulation

mechanisms. Currently, the work for deriving network based biomarkers is limited to one type of assay such as gene expression or protein array. Developing a method to identify network based biomarkers from multiple assay types would be an interesting topic.

## Chapter 3

# Gene alteration identified by the Item Response Model

*(Most of the materials in this chapter have been published online in Bioinformatics, September 2012: Tong, P. et al, “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)*

### 3.1 Background

The previous chapter introduced a regression based method to identify predictive biomarkers utilizing multiple assay types. This is very important for deriving novel therapeutics and ultimately allocating patients into different risk groups based on their genetic fingerprints. Now we shift to another important topic where the goal is to illuminate cancer related alterations. These alterations are not necessarily predictive to clinical outcome directly. Instead, they characterize the irreversible events transforming a normal cell to a cancerous cell. As



a result, such alterations can help us understand the mechanisms driving cancer formation and suggest biologically driven approaches for cancer treatment.

Cancer related alterations can happen in all aspects of the hallmarks of cancer: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming energy metabolism, and evading immune destruction [Hanahan and Weinberg, 2011]. It is realized that identifying such alterations is the first step towards individualized cancer treatment. With the recent explosion of high throughput technologies, researchers have been employing multiple types of assays to interrogate the genetic characteristics of cancers. For example, The Cancer Genome Atlas (TCGA) is aimed at systematically profiling over 20 different cancer types using gene expression, methylation, microRNA, SNP and copy number array as well as next generation sequencing to explore alterations in DNA and RNA [McLendon et al., 2008]. The rich information gathered through the TCGA project poses a great challenge for data integration.

Currently, there are four major methods proposed to integrate gene expression and copy number (CN) data: stepwise, regression-based, correlation-based, and latent variable models [Lahti et al., 2012, Huang et al., 2012]. Most of these methods focus on pairwise correlation between assays and/or outcome. For example, SODEGIR is a stepwise method to identify overlapping genomic regions of differential expression and genomic imbalance [Bicciato et al., 2009]. Linear mixed models have also been used to search for genes whose expression is affected by CN change [Menezes et al., 2009]. Similarly, penalized multiple regression is another method to model the dependence between gene expression and CN change [Peng et al., 2010]. To search for associations among gene expression, CN change and clinical outcome, methods based on canonical correlation analysis have also been developed [Waaijenborg et al., 2008, Lê Cao et al., 2009, Witten et al., 2009]. A comprehensive comparison study of these methods can be found

in Louhimo et al. [2012]. Recently, the CNAmets method has been developed to integrate gene expression, methylation and copy number with a goal of searching for synergistic regulations [Louhimo and Hautaniemi, 2011].

All existing approaches explicitly model the dependence of gene expression, methylation, and copy number. However, methylation and copy number only accounts partially for the expression variation. The correlations between expression and copy number/methylation are sometimes weak or even in the opposite direction since there are many other regulators of gene expression. Therefore, inference relying on the dependence structure might be too restrictive. We proposed a novel method called **integIRTy** (**integration using Item Response Theory**) that is able to infer gene alterations from multiple assays without assuming any correlation structure. **IntegIRTy** is a latent variable approach that automatically adjusts for sample heterogeneity as well as technical artifacts such that alteration scores are put on a common scale for easy comparison.

**IntegIRTy** is based on the Item Response Theory (IRT) developed in psychology. The IRT model has been widely used to construct and score psychological and educational tests such as the SAT and GRE tests [Baker and Kim, 2004]. A test consists of a set of items (or questions) constructed to measure a certain ability. Once the response to the items is obtained, IRT can be used to infer the latent ability score after adjusting for item difficulty. The IRT model enjoys the invariant property where the latent ability estimates are invariant to item difficulty level. This is quite appealing when it is necessary to compare scores from different years or different institutions.

The context of IRT in educational tests resembles the challenge of identifying gene alterations. Given a set of samples simultaneously profiled by different assay types, we treat the genes as examinees and samples measured in different assays as items. The heterogeneity among assays as well as samples can be au-

tomatically adjusted by fitting different item parameters. Due to the invariant property of IRT, the latent ability estimates would be comparable. Further, to integrate multiple assay types, we can just assemble a larger test with different assays supplying the items. This gives us a flexible path for data integration.

We evaluate the proposed method through both simulation study and real data analysis. The simulation shows that both item parameters and latent trait estimates are quite accurate. When applied to integrate real data, `integIRTy` can identify novel alterations that cannot be found when analyzing the assays separately. The new method is also compared to conventional methods such as student's t-test and Wilcoxon rank-sum test when there is only one assay type. We found that `integIRTy` is more robust and reliable than these conventional methods. Comparison with the CNAmets approach shows that the two approaches provide complementary information when data is integrated.

### 3.2 Methods

Item Response Theory (IRT) refers to a family of models that describe the relationship of an examinee's performance on a set of test items to his or her underlying (latent) ability level. In practice, this relationship is modeled by a monotonically increasing function called the Item Characteristic Curve (ICC; see Fig.3.1). For binary responses (i.e., right or wrong), a two-parameter logistic model (2PL) can be specified as follows:

$$P_i(\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}} \quad (3.1)$$

The left hand side is the probability of a correct response to item  $i$  for person  $j$  with ability level  $\theta_j$ . On the right side,  $b_i$  is the *item difficulty* parameter for item  $i$  that determines the position of the ICC in relation to the ability scale. The item difficulty is the ability level required to achieve a 50% chance of a correct response on this item. As  $b_i$  increases, the item becomes harder. The remaining parameter,  $a_i$ , is the *item discrimination* for item  $i$ , which represents how well

the item discriminates among the examinees. It is proportional to the slope of the ICC at ability level  $b_i$ . Higher discrimination means that the item discriminates more clearly among the examinees, and hence is more informative. Fig.3.1a shows an ICC with difficulty  $b_i = 0.5$  and discrimination  $a_i = 1$ . Fig.3.1b shows several ICCs fitted from real data.

As noted above, we apply IRT by treating genes as examinees and patients as items. The main parameter of scientific interest is the latent “ability” of each gene to be altered in cancer samples across all assay types and samples. Patients with many altered genes (low item difficulty) provide less useful information than patients with only a few altered genes (high item difficulty). Groups of patients with similar patterns of altered genes tend to have a high item discrimination and so are weighted more heavily than a patient who has an idiosyncratic set of altered genes (and low item discrimination).

Importantly, the IRT model is expressed at the item level rather than the test level. This feature gives IRT models the so-called *invariant property*. The invariant property implies that (i) item parameters are characteristics of the item, and hence are not dependent upon examinees who take the test; (ii) the ability parameter that characterizes an examinee is not test-dependent, and hence scores from different tests are comparable.

The 2PL model can be augmented by introducing a guessing parameter, which is then called the three-parameter logistic model (3PL). There is also a one-parameter logistic model (1PL), obtained by forcing  $a_i$  to equal 1. This model is also called the Rasch model. Since the three models are nested, one can use a Likelihood Ratio test to select the best model [Neyman and Pearson, 1933]. Alternatively, we can use an information-based criterion such as Akaike’s information criterion (AIC) or Schwarz’s Bayesian information criterion (BIC) to identify the best model.

### 3.2.1 Parameter Estimation

Parameter estimation has received tremendous considerations in the IRT literature. Methods under the maximum likelihood framework include joint maximum likelihood (JMLE), marginal maximum likelihood (MMLE), and conditional maximum likelihood (CML). JMLE has been shown to have many inherent problems, the most serious being that it does not provide consistent estimates [Baker and Kim, 2004]. It also fails to estimate the latent trait when all items are answered correctly or incorrectly. In comparison, both MMLE and CML

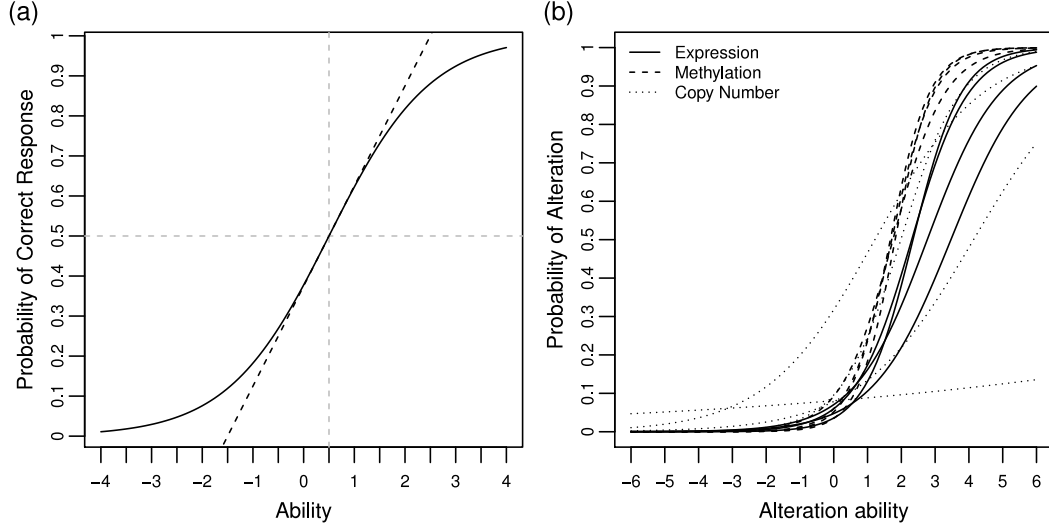


Figure 3.1: Illustration of Item Characteristic Curve (ICC). (a) Exemplar ICC with a difficulty level of 0.5 and discrimination 1. (b) ICCs from real data. The first four OV patient samples for each assay are shown here. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

provide consistent parameter estimates. However, the CML approach is only possible under the Rasch model [Andersen, 1980]. Recently, Bayesian estimation has also been used [Fox, 2010]. We adopt the MMLE procedure that has become the standard method since its introduction, for which an implementation in R is available through the `ltm` package [Rizopoulos, 2006].

According to Baker and Kim, the MMLE can be formulated under the EM framework [Baker and Kim, 2004]. First, the item parameters are estimated by maximizing the observed data likelihood after integrating out the latent traits. For example, the contribution of the  $m$ th examinee to the observed likelihood can be written as:

$$\ell_m = \sum_{i=1}^{i=K} \log \int p(x_{im} = 1 | a_m, b_m; \theta) f(\theta) d\theta \quad (3.2)$$

In Equation 3.2, a prior distribution  $f(\theta)$  needs to be specified for the latent trait (usually a standard normal distribution is used),  $x_{im}$  is the response to item  $i$  for examinee  $m$ , and  $K$  is the total number of items in the test. Since there is no closed-form formula for the observed data likelihood, Gauss-Hermite quadrature is required to evaluate the integral. Second, given current item parameter estimates  $\hat{a}$ ,  $\hat{b}$ , the latent trait is estimated by the posterior mode as:

$$\hat{\theta}_m = \arg \max_{\theta} \sum_{i=1}^{i=K} \log \int p(x_{im} = 1 | \hat{a}, \hat{b}; \theta) f(\theta) d\theta \quad (3.3)$$

### 3.2.2 Estimation of Latent Traits from Integrated Data

Conceptually, estimating the latent trait for integrated data would be the same as estimation from an individual assay. However, when there are many assays to integrate or when there are many items in each assay, parameter estimation could become ill-conditioned. This can happen when there are more unknown parameters (since each item introduces two parameters) than the data can afford to estimate. One way to deal with this problem is to fix the item parameters estimated from the individual assay type when modeling the combined data. That is, the item parameters characterizing the items remain the same when we estimate the integrated latent trait. This approach is valid because the IRT model has the invariant property. Since the item parameters are pre-estimated, we can simply calculate the latent traits for integrated data using the maximum a posteriori estimates [Magis, 2011].

### 3.2.3 Statistical Significance Assessment

In order to identify genes showing statistically significant alteration, we need to derive the null distribution of latent traits. Since there is no existing method for this purpose in the item-response setting, we use a nonparametric test to define empirical  $p$ -values based on permutation. Two alternative strategies can be used to infer the null distribution of latent traits similar to the ‘gene sampling’ and ‘sample label permutation’ methods [Ackermann and Strimmer, 2009].

Gene sampling corresponds to calculating latent traits after permuting the response matrix within samples. In this case, computed  $p$  values measure how different the observed latent trait is from the case where alterations happen randomly on the genes. Note that this method can be used even when normal (control) samples are unavailable. When normal samples are present, the sample label permutation approach can be used. The null latent trait can be computed by following the same procedure as computing the observed latent traits after permuting sample labels. As Ackermann and Strimmer point out, the two approaches can yield quite different results since they test different null hypotheses. Once the empirical  $p$  value is calculated, multiple testing can be adjusted using existing methods [Dudoit et al., 2003]. It should be noted that our integration approach entails no additional price in terms of multiple comparisons compared to analyzing just one dataset.

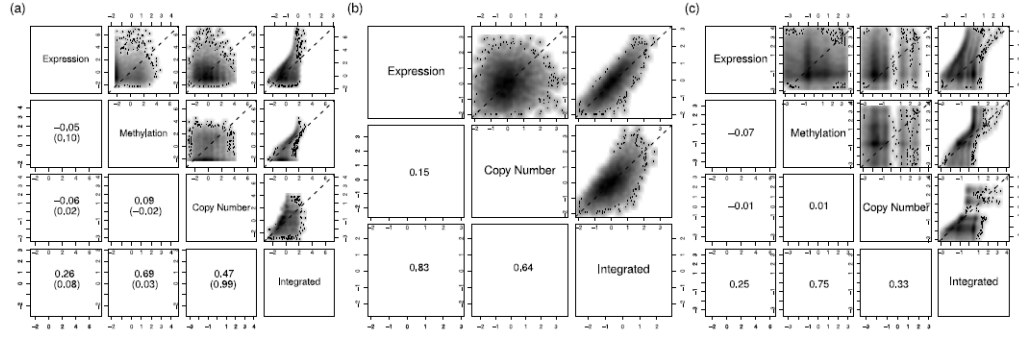


Figure 3.2: Pair-wise smoothed density (darker cloud indicates higher density) of estimated latent trait for alteration (upper panels) and Spearman rank correlations (lower panels) among different assays and integrated data. When normal control samples are available for all assays, we also show the correlations of computed  $p$ -values from conventional methods in bracket. (a) OV dataset. (b) BRCA dataset. (c) GBM dataset. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

### 3.2.4 Data Dichotomization

Proper transformation of the data is needed to fit the IRT model. To do this, one needs to define a metric measuring the alteration magnitude. Then, a specified cutoff is used to dichotomize the data. Many methods can be used to define the alteration magnitude. For two-channel array data which provides the log ratio of intensities from tumor and healthy reference tissue, the log ratio itself can serve as alteration magnitude. When normal tissue is present, a feasible metric for expression data is to use a Z-like statistic that measures the deviation of a tumor sample from normal controls (see Section 3.5.2).

We could dichotomize the methylation data the same way as expression data. However, since methylation measurements (the  $\beta$  value) are bounded between 0 and 1 with an asymmetric distribution, a more biologically relevant method is to first discretize the methylation into three groups: unmethylated ( $\beta < 0.25$ ), partly methylated ( $0.25 \leq \beta \leq 0.75$ ), and highly methylated ( $\beta > 0.75$ ). If the group membership for a tumor sample differs from the normal reference (defined by the normal  $\beta$  mean value), then we code this gene in this sample as altered.

Transformation of CN data is easier. Choosing a fixed threshold (e.g., 0.1, 0.2, ..., 0.7), genes with absolute adjusted log2Ratio larger than this cutoff are converted to 1, and 0 otherwise. We use the adjusted log2Ratio derived by subtracting measurements on matched normal tissue from tumor tissue to exclude germline CN change that is irrelevant to tumorigenesis.

Generally, the choice of cutoff to dichotomize the alteration magnitude should not affect the final result as long as it is sensible. This can be

evaluated through a sensitivity analysis, which we present below by comparing the latent trait estimates obtained from various versions of the data transformed using different cutoffs.

*(This section is an excerpt from Tong, P. et al, “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”, Bioinformatics, 2012)*

### 3.3 Results

We performed a simulation study to show that the model can recover both the item parameters and latent traits in an integration setting (see Section 3.5.3). We then investigated three public datasets (Table 3.1). Details about the samples and preprocessing steps are provided in Section 3.5.1. Both the OV and GBM datasets contain three types of assays interrogating expression (Expr), methylation (Methy) and copy number (CN). The BRCA dataset only contains data for expression and copy number. All three datasets have been examined in detail. However, due to page limitations, we mainly focus on OV data and present other datasets when necessary. The 1PL, 2PL and 3PL models have been fitted on each dataset. Based on BIC, the 2PL model is shown to be the preferred model in all datasets (Supplemental Table 3). Hence, all results presented below are from 2PL model.

Table 3.1: Number of patients per dataset (*Table reprinted from Tong, P. et al, Bioinformatics, 2012*)

	OV		BRCA		GBM	
	Tumor	Normal	Tumor	Normal	Tumor	Normal
Expression	569	8	37	NA	473	10
Methylation	526	10	0	NA	370	6
CN	571	567	37	NA	341	341

#### 3.3.1 Alteration Patterns Across Assays

IntegIRTy allows us to evaluate and compare the alteration pattern across different assays. Fig.3.2 compares the alteration from individual assays and after integration for OV, BRCA and GBM datasets. A



common pattern is that there are not many genes with severe alteration in all assays. We observed little correlation between the assays for either the conventional method or our IRT method. This is in agreement with previous results that showed only a small fraction of the variation in expression was attributable to methylation [Wu et al., 2010] or to CN change [Stranger et al., 2007] in a global sense.

The amount contributed by individual assays to the integrated data differs. For both OV and GBM, the integrated latent trait is primarily influenced by methylation data, followed by copy number, and then expression. In BRCA, expression has a larger influence than copy number. The correlation between latent traits from integrated data and individual datasets is well behaved compared to the conventional method where the integrated  $p$ -value is mostly dominated by the copy number data.

### 3.3.2 Sensitivity Analysis

The latent trait estimates derived using different thresholds to dichotomize the expression data agree well, especially for high latent traits (Supplemental Figure 3.6 a-c). In the low latent trait range, agreement is somewhat worse, mainly due to high standard error (SE) associated with latent trait estimates. The agreement for CN is even better (Supplemental Figure 3.6 d-f). Hence, the proposed method is robust to cutoff choice during data transformation. In comparison, the naive score that simply computes the percentage of “correct responses” (i.e., alterations) varies when using different thresholds (Supplemental Figure 3.7). Although the latent trait estimates are similar using different thresholds, we use relatively stringent thresholds (2.5 for expression and 0.4 for copy number) for further analyses.

### 3.3.3 Contribution of Individual Assays to the Integrated Analysis

We performed a series of analyses to determine how individual assays contribute to the list of genes found by an integrate analysis using `integIRTy`. Fig.3.3 breaks down the lists of “top  $N$ ” genes (for  $N$  from 100 to 1000) from the integrated analysis of the OV dataset to see which genes are on one, two, or all three of the top  $N$  lists from the individual assays. This figure shows that the top  $N$  list for methylation has the best agreement with the integrated list, with expression being second, and copy number having the least agreement.

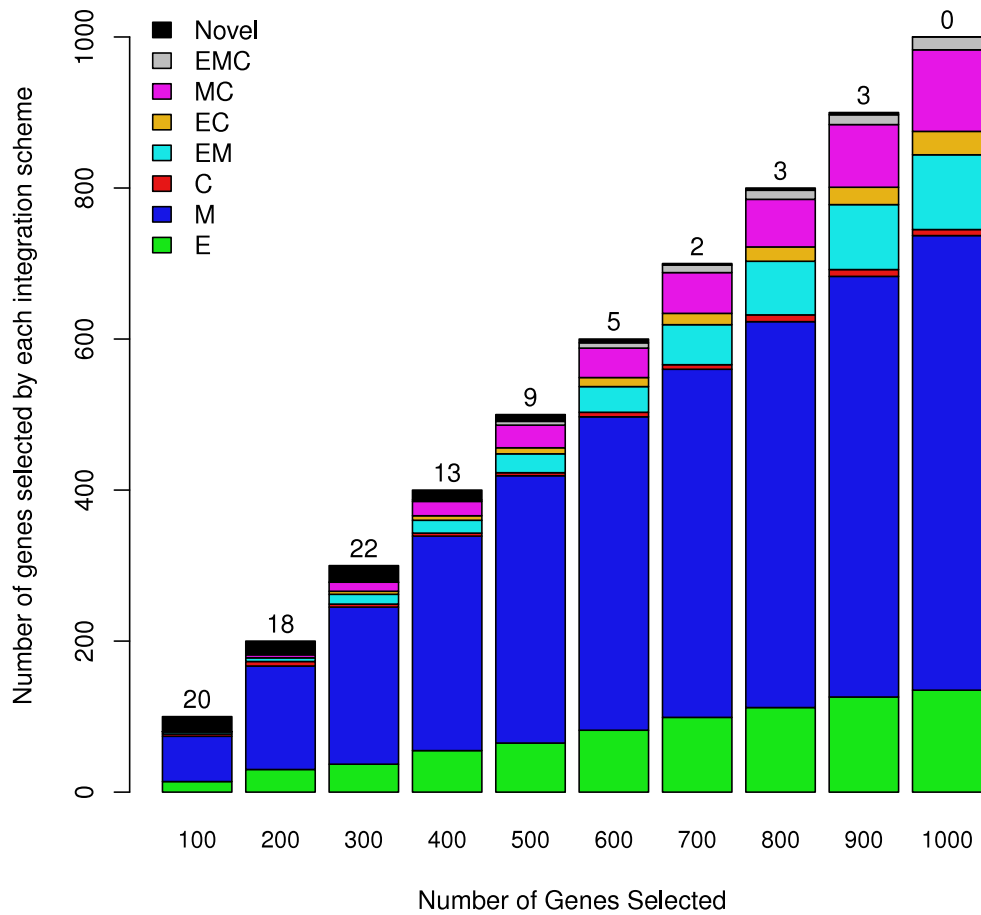


Figure 3.3: Relations between integrated and individual gene lists in OV data. We selected the top (100 to 1000) genes from the integrated analysis and from individual assays (E: expression, M: methylation, C: copy number). Each bar is equivalent to a Venn diagram showing how many of the top genes from the integrated analysis came from one, two (EM = expression and methylation; EC = expression and copy number; MC = methylation and copy number) or all three (EMC) individual assay gene lists. Black regions and numbers at the top of each bar count the number of “novel” genes that only appear on the list from the integrated analysis. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

The relative contributions of the three assays to the integrated list remain consistent as we increase the length of the gene list. Moreover, both the absolute and relative number of “novel” genes decreases as  $N$  increases. We performed similar analyses integrating assays two at a time and constructed similar plots (Supplemental Figure 3.18- Supplemental Figure 3.19); the findings from this sequential integration are consistent with Fig.3.3.

### 3.3.4 Novel Altered Genes Emerge When Data Is Integrated

The latent trait estimated from the integrated data represents the overall propensity of a gene to be altered in at least one assay (expression, methylation or copy number). Although this latent trait is a compromise among the latent traits estimated from individual assays, in many cases the integrated rank (smaller rank for larger latent trait) is smaller than the average rank of the three assays (Supplemental Figure 3.8a, c and e). There are also several occasions (296 genes in OV, 380 genes in BRCA, 811 genes in GBM) where the integrated rank is smaller than any of the ranks from individual assay (Supplemental Figure 3.8b, d and f). The ability to identify altered genes that cannot be identified by individual assays shows the merits of data integration as well as the effectiveness of our method.

Table 3.2 shows the latent trait and rank for the top 20 genes selected using the integrated data for the OV data (similar results for BRCA and GBM data are in Supplementary Tables 1 and 2). The genes CDKN2A, VEGFC and STMN1 are only discovered by the integrated analysis (in top 20) and would not have been discovered using any of the individual assays (out of top 100). To verify that these genes are altered, we examined the original data and their relationships (Supplemental Figure 3.9- 3.11). The original data strongly supports our finding. Interestingly, mRNA up-regulation of CDKN2A is associated with increased methylation. Further, functional annotations show that these genes are linked to cancer. Specifically, CDKN2A is involved in two critical cell cycle regulatory pathways, the p53 pathway and the RB pathway. CDKN2A plays an important role in many human cancers including pancreatic cancer, esophageal and gastric cancers, leukemia, bladder cancer. and cutaneous melanoma. Differential expression of VEGFC is related to the different propensity to lymph node metastasis in thyroid cancers [Hung et al., 2003]. STMN1 is an oncoprotein regulating microtubule dynamics. Defective STMN1 causes constant mitotic spindle assembly and hence unregulated cell growth [Cassimeris, 2002].

Among the top 20 genes discovered by integration, the BRCA datasets

Table 3.2: Latent trait and rank for top 20 genes selected by integrated analysis of TCGA OV data

Genes	Integrated		Expression		Methylation		Copy Number	
	LT*	Rank	LT	Rank	LT	Rank	LT	Rank
TTYH1	3.52	1(E)	5.00	45(E)	3.11	47(M)	0.49	2744
SPARCL1	3.46	2(M)	3.91	108	3.43	18(M)	1.42	1148
SPAG6	3.27	3(E)	6.21	17(E)	2.89	79(M)	-1.30	6973
CRISP2	3.01	4(E)	6.70	8(E)	2.40	225	-0.65	5646
DPT	3.01	5(E)	4.93	49(E)	2.73	115	-1.57	7397
CFD	2.86	6(E)	4.66	63(E)	2.09	433	1.89	616
HNF1B	2.81	7(E)	4.98	47(E)	1.89	607	2.89	133
CDKN2A	2.72	8(I)	3.79	120	2.37	247	0.51	2684
C11orf16	2.71	9(E)	5.83	24(E)	1.96	549	0.49	2752
PDE8B	2.68	10(E)	5.93	22(E)	1.79	718	1.47	1050
RIMBP2	2.68	11(E)	4.44	73(E)	2.23	322	-0.39	4876
PIPOX	2.64	12(M)	1.68	654	4.01	3(M)	2.63	182
VEGFC	2.60	13(I)	3.31	152	2.27	302	1.59	839
AGT	2.56	14(E)	4.33	79(E)	2.18	360	-1.05	6406
CXorf57	2.52	15(E)	4.62	65(E)	1.78	735	1.38	1213
CST6	2.51	16(M)	2.82	224	3.16	40(M)	-2.02	8111
PRAME	2.43	17(E)	4.09	94(E)	1.76	749	1.61	815
CDO1	2.43	18(M)	2.17	396	2.99	61(M)	0.54	2568
FBLN1	2.41	19(C)	2.91	206	2.06	452	3.35	67(C)
STMN1	2.41	20(I)	2.85	216	2.55	168	-0.28	4579

LT\* stands for latent trait. Genes with ranks lower than 100 are coded differently in the rank column (expression: E, methylation: M, copy number: C). Genes identified only by integrated data are coded as I in the integrated rank column.

identify two novel genes (SELENBP1 and EDIL3) while the GBM datasets identify 17 novel genes (Supplemental Table 2). SELENBP1 has been found to mediate the anticancer action of selenium in prostate [Yang and Sytkowski, 1998], lung [Chen et al., 2004], and colon [Kim et al., 2006] cancer. EDIL3 plays an important role in mediating angiogenesis [Aoki et al., 2005]. Functions of the 17 novel genes in GBM data include cell death, hematological system development, cell morphology, nervous system development and cell cycle, according to Ingenuity Pathway Analysis (Ingenuity® Systems, www.ingenuity.com).

### 3.3.5 Comparison to Conventional Methods

Identifying altered genes is essentially a two-group (tumor versus normal) comparison problem. Hence, we can compare our method to conventional methods such as the t-test (for expression and copy number data) or rank test (for methylation data which is bounded between

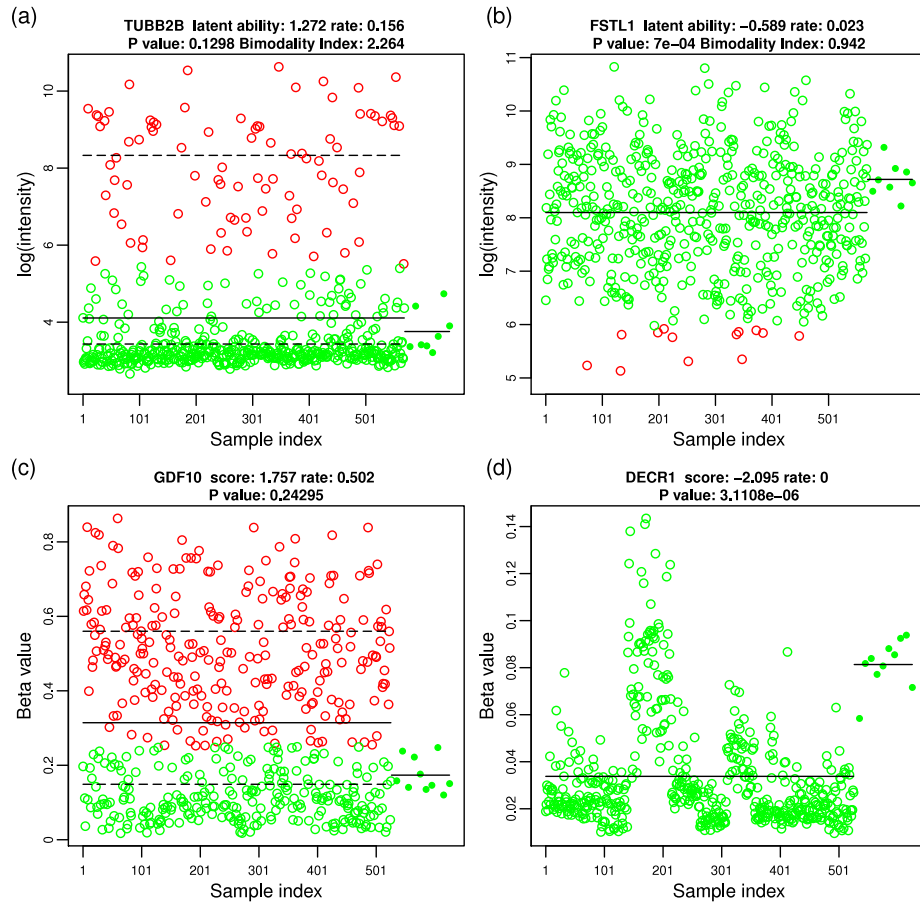


Figure 3.4: Example genes with discordant calls between conventional method and our method. The original measurement is plotted against sample index after sorting by tissue type and batch number. Red circles indicate altered values based on dichotomized data; green circles indicate unaltered values. The expression from normal control samples are indicated by solid green dots. Black solid lines represent tumor and normal mean. Dashed lines denote the component means estimated from 2-component mixture. In the panel titles, we show gene symbol, latent ability, percentage of tumor samples altered (rate), and conventional test  $p$ -value. (a) A typical gene missed by t-test but identified by our method. Bimodality index (BI) shown in the title strongly suggests a sub-group of the tumor samples have a large magnitude of over-expression compared to normal samples, and hence, is likely to be altered. (b) A gene missed by our method but flagged by t-test. This is an example where statistical significance does not imply biological significance. The difference between tumor and normal sample is minor. (c) A typical gene missed by rank test but flagged by our method. Over 50% of the tumor samples have increased methylation which strongly suggests altered methylation. (d) A gene missed by our method but flagged by rank test. The trend of beta value here is mostly due to batch effect, not biological difference. All tumor and normal samples are not methylated ( $\beta < 0.25$ ). Accordingly, our method assigns a very low latent trait estimate. In comparison, the conventional method dictates a strong statistical difference between tumor and normal simply due to batch effect. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

0 and 1)(see Supplemental Figure 3.12). There is no direct test that applies to combined data. Inspired by Fisher’s method for meta-analysis based on  $p$ -values, we use the geometric mean of the  $p$ -values from each assay to represent the “conventional”  $p$ -value from the integrated data. Empirical  $p$ -values for the latent traits were computed based on the permutation test described previously. A gene was assigned a positive call if its  $p$ -value was less than a specific cutoff. Thus, we can divide the genes into four categories: Positive/Negative (positive call by our method and negative call by conventional method), Negative/Positive (negative call by our method and positive call by conventional method), Positive/Positive (both methods give positive calls) and Negative/Negative (both methods give negative calls). Conventional methods model genes separately while our method models all genes and samples simultaneously. Hence, we do not expect a perfect correlation between latent traits and  $p$ -values from conventional methods.

We inspected the original measurements and found that the Positive/Negative genes found by our method are meaningful and very likely to be truly altered genes while Negative/Positive genes missed by our method are actually not severely altered even though they are statistically significant due to increased sample size or batch effect (see Fig.3.4). For example, Positive/Negative genes usually exhibit non-Gaussian expression which cannot be detected by the t-test but can be identified by our method as shown in Fig.3.4a. Compared to normal, this gene is expressed 16 fold higher in more than 15% of the tumor samples which suggests it is likely to be altered. Fig.3.4b shows a typical example of a Negative/Positive gene in expression. Although the increased sample size enables us to compute a significant  $p$ -value, a statistically significant difference doesn’t necessarily mean a biological difference. In this example, there is almost no difference in the mean expression (8.7 vs 8.1) which strongly suggests that our method gives the right decision. In Fig.3.4c, a Positive/Negative gene is shown that obviously exhibits different methylation pattern that the rank test fails to detect. In comparison, Negative/Positive genes are usually not biologically different (i.e. almost all samples have beta value less than 0.25 and hence are unmethylated) but statistically different mainly due to a batch effect (Fig.3.4d). For CN data, there are many negative/positive cases where t-test assigns more than half of the genes a 0  $p$ -value due to large sample size (571 tumor and 567 normal samples). In fact, the difference of mean log2ratio between tumor and normal is biologically negligible for almost all of these genes.

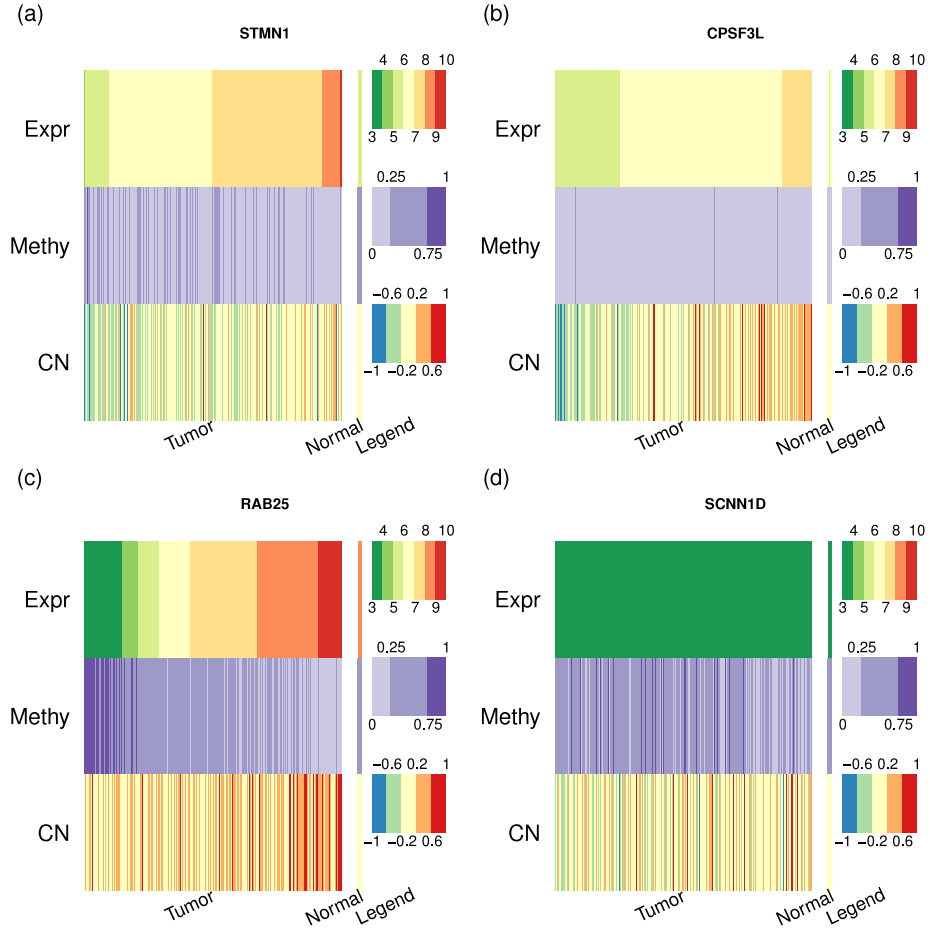


Figure 3.5: Complementary information provided by `integIRTy` and `CNAmets`. (a) Over-expression in tumor where the regulation by methylation and CN is not synergistic. As a result, `CNAmets` fails to detect it. (b) Mild over-expression mainly driven by CN gain. `integIRTy` didn't detect this gene due to the high background CN change. (c) Over-expression in tumor samples driven by hypomethylation and CN gain. Genes like this are easy to be detected by both methods. (d) Expression is turned off in both tumor and normal samples due to hyper-methylation. Since there is little difference between tumor and normal, both methods suggest it is not altered. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

### 3.3.6 Item Parameters Characterize Properties of Samples

While the estimated latent traits characterize the properties of genes, the item parameters can characterize patient samples. Samples with small item difficulty are easier to be altered, and hence contain more alterations. In both OV and GBM data, expression data has the largest (median) difficulty followed by methylation and then CN (Supplemental Figure 3.13). This implies the frequency of CN alteration is higher compared to expression or methylation. The widespread CN change in OV data identified by our model agrees with previous finding [The Cancer Genome Atlas Research Network, 2011]. The median item difficulty for expression and CN in BRCA data is similar, although the difficulty estimates for CN are more variable, suggesting a higher heterogeneity in CN.

### 3.3.7 Complementary Information Provided by `integIRTy` and `CNAmet`

We compared `integIRTy` with `CNAmet`, another method developed to integrate expression, methylation, and CN data [Louhimo and Hautaniemi, 2011]. Although `CNAmet` shares a similar idea by dichotomizing methylation and CN data before integration, it has a different goal. Rather than identify genes altered between tumor and normal samples, `CNAmet` searches for genes whose expression is synergistically regulated by methylation and CN. Conceptually, gene alteration can happen with or without synergistic regulation and vice versa. This is confirmed by real data analysis where both *alteration without synergistic regulation* (Fig.3.5a) and *no alteration yet under synergistic regulation* (Fig.3.5b) are observed. When both methods give concordant calls (Fig.3.5c and Fig.3.5d), `integIRTy` and `CNAmet` provide complementary information that not only tell us whether a gene is altered but the underlying mechanism. Interestingly, roughly half of the altered genes are not under synergistic regulation and half of the genes under synergistic regulation are not altered (Supplemental Figure 3.14).

*(This section is an excerpt from Tong, P. et al, “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”, Bioinformatics, 2012)*

## 3.4 Discussion

Data integration is a critical challenge in integrative biology where multiple assays are simultaneously employed to profile the same set of samples. Most existing approaches for data integration assume a dependency structure among the assays which might not be satisfied in all cases. Here we propose a natural and interpretable framework to integrate heterogeneous high throughput datasets without explicitly assuming any correlation structure.

The proposed method `integIRTy` puts the “propensity to alteration” on a common scale such that meaningful comparison can be attained. This is achieved by automatically adjusting for sample heterogeneity through fitting different item



parameters in an Item Response Model framework. It is found that the integrated latent trait can be used to identify novel genes that alter marginally in individual assay but in a consistent manner across assays. Furthermore, the estimated latent trait together with item parameters characterizing the properties of genes and patient samples can be used to visualize high dimensional dataset intuitively.

Genes identified by our method are more reliable and biologically meaningful than genes found by conventional methods such as student’s t test and rank test. With enough sample size, conventional methods would always declare a significant difference between tumor and normal where the effect size might be biologically negligible. In contrast, genes identified by `integIRTy` are ensured to have biological difference. The reason is that latent traits are computed from the dichotomized data with the biology (alteration status) already built in. Further, `integIRTy` doesn’t make any distributional assumption about the original data. As a result, genes violating the distributional assumption are missed by conventional methods (e.g., t test) but can be still found by our method.

Currently, `integIRTy` proceeds by transforming the data into binary indicators. However, it is possible to work with an ordinal response matrix that might arise from categorizing copy number into loss, neutral and gain. In this case, we can apply the rating scale model [Andrich, 1978], the generalized partial credit model [Muraki, 1992] or the graded response model [Samejima, 1969]. A generalized version of the latent trait model that works on continuous measurements would be quite interesting. It turns out that the natural generalization of the IRT model [Mellenbergh, 1994, Moustaki, 2011] is equivalent to Factor Analysis (FA). FA works well in integrating expression data profiled by different microarray platforms [Wang et al., 2011] where the data is highly homogeneous. To integrate data from multiple assay types, it is necessary to specify a joint distribution leading to a very complicated model that is difficult to track. Therefore, the FA model cannot be easily applied to integrate heterogeneous data from

multiple assays. For continuous bounded data (e.g., data from a methylation array), a generalized IRT model based on beta distributions is available [Noel and Dauvier, 2007]. However, this model does not work on normally distributed data and hence does not apply to our integration setting.

The proposed method can be extended to incorporate clinical variables. This can be achieved by specifying a linear relation between the clinical variable and the latent trait parameter. The latent trait can be thought of as a random effect. In this case, the extended model can be used to examine how the clinical variable affects the population mean of latent traits.

### 3.5 Appendix

*(Excerpts in this Appendix section are from the supplemental materials published online from: Tong, P. et al (2012), Bioinformatics. “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)*

#### 3.5.1 Dataset Assembly (OV, BRCA and GBM)

The latest version of TCGA OV (ovarian carcinoma) and GBM (glioblastoma) data up to December 2011 were downloaded from TCGA data portal at <http://cancergenome.nih.gov/>. For OV data, we selected the Affymetrix U133A array (BI HT\_HG-U133A), the Illumina Infinium HumanMethylation27 BeadChip (JHU-USC HumanMethylation27), and the Agilent CGH array (MSKCC CGH-1X1M\_G4447A) as representative platforms measuring gene expression, methylation, and copy number. For GBM datasets, we assembled the Agilent expression array (UNC AgilentG4502A), Illumina Infinium HumanMethylation27 BeadChip (JHU-USC HumanMethylation27) and Agilent CGH array (HMS HG-CHG-415K, HMS HG-CGH-244A). The TCGA GBM

data didn't contain normal tissue samples for methylation. Instead, we use the methylation data from Wu et al. [2010] which contain six normal human brain tissues from accident victims profiled by NimbleGen tiling array. The gene list identified as methylated in normal tissue is downloaded through the supplement file in their paper. The BRCA data was produced by Pollack et al consisting of copy number and expression array data for 37 breast tumors interrogated by the same 2-channel cDNA microarrays [Pollack et al., 2002]. A global reference sample from a healthy female is used to hybridize each sample. Preprocessed data by the authors were downloaded. Clone ID to gene symbol conversion was done by the Stanford Source website (<http://smd.stanford.edu/cgi-bin/source/sourceSearch>).

The TCGA data portal deposits 4 levels of data. Our analysis is based on the level 3 data that contains the interpreted or segmented data. For expression data, this means the summarized expression for each gene. For methylation, level 3 data contains the  $\beta$  value for each methylation site. Level 3 data for copy number stores segmented regions per sample.

We need summarized measurements for each gene and each sample in our analysis. The Level 3 gene expression data is already in this format and needs no further manipulation. For the copy number, we need to map the segmented regions to individual genes. The RefGene gene coordinates from UCSC annotation database (hg18) are used for this purpose [Fujita et al., 2011]. Genes partially covered by one or more regions need special attention. Our strategy is to use the log2Ratio with the most supporting probes to represent the CN value. We do not use the most extreme measurement since many of these cases contain only 1 or 2 probes which are not robust. Genes on chromosome X and Y are excluded in the final analysis for copy number data. Since there is copy number polymorphism as well as somatic copy number change that might not be related to cancer, we use the adjusted log2Ratio, which is computed by subtracting the  $\log_2\text{Ratio}$  of the paired control sample from the log2Ratio of the solid tumor sample. For methylation, multiple methylation sites for the same gene are summarized by their median to represent the overall gene methylation status. The correspondence information between methylation site and genes is already contained in the level 3 data. (*An excerpt from Tong, P. et al, "integRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory", Bioinformatics, 2012*)

### 3.5.2 Data transformation

Proper transformation of the data is needed to fit the IRT model. Details for dichotomizing methylation and CN data are described previously. Here we give the details for a Z-like metric that dichotomizes expression data. This metric requires normal control samples. The Z-like metric is defined as:

$$Z_{gi} = \frac{x_{gi} - \mu_{g,normal}}{\sqrt{\sigma_{g,tumor}^2 + \frac{\sigma_{g,normal}^2}{N_{normal}}}}$$

Here  $x_{gi}$  is the original (continuous) measurement for gene  $g$  in sample  $i$ ,  $\mu_{g,normal}$  is the normal mean expression for gene  $g$ ,  $\sigma_{g,tumor}$  and  $\sigma_{g,normal}$  are the standard deviations (SDs) for gene  $g$  computed from tumor and normal samples, respectively, and  $N_{normal}$  is the number of normal samples. We observed that  $\sigma_{g,tumor}$  can be inflated when tumor samples exhibit non-Gaussian expression, which would lead to a deflated alteration magnitude. To accommodate this situation, we use the Bimodality Index to flag these genes and replace  $\sigma_{g,tumor}$  with the standard deviation estimated from a two-component normal mixture model [Fraley and Raftery, 2002]. After setting a cutoff for  $|Z_{gi}|$ , say 2.5 or 3, we can convert  $Z_{gi}$  to a binary indicator. A converted 1 means that the expression measurement for gene  $g$  in sample  $i$  differs from the normal tissue, and so that this gene is likely to be altered. (An excerpt from Tong, P. et al, “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”, *Bioinformatics*, 2012)

### 3.5.3 Simulation Study

We simulate data with known latent traits and item parameters and show that the model can recover both item parameter and latent traits. Let  $\theta^E = (\theta_1^E, \theta_2^E, \dots, \theta_G^E)$  denote the vector of latent traits for expression data with  $G$  genes. Similarly, let  $\theta^M = (\theta_1^M, \theta_2^M, \dots, \theta_G^M)$  and  $\theta^C = (\theta_1^C, \theta_2^C, \dots, \theta_G^C)$  denote the latent traits for methylation and CN data, respectively. The latent traits for different genes can be treated as i.i.d. random variables. In particular,

$$\theta_g^E \stackrel{\text{iid}}{\sim} N(0, 1), g = 1, 2, \dots, G$$

$$\theta_g^M \stackrel{\text{iid}}{\sim} N(0, 1), g = 1, 2, \dots, G$$

$$\theta_g^C \stackrel{\text{iid}}{\sim} N(0, 1), g = 1, 2, \dots, G$$

To simulate the dependency among different mechanisms of regulation, for any positive definite 3-by-3 matrix  $\Sigma$ , we can impose the following restrictions:

$$\begin{pmatrix} \theta_g^E \\ \theta_g^M \\ \theta_g^C \end{pmatrix} \sim MVN \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma$$

The item parameters are simulated from a Gamma distribution. For example, for the expression data, we have:

$$b_i^E \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha^E, \beta^E), i = 1, 2, \dots, K$$

$$a_i^E \stackrel{\text{iid}}{\sim} \text{Gamma}(\gamma^E, \delta^E), i = 1, 2, \dots, K$$

In our simulation, we specify  $G=2000$  genes and  $K=200$  samples in each platform. The Gamma distribution for difficulty parameter has mean 3, 2, 1 and variance all as 0.2 for expression, methylation and CN data, respectively. For item discrimination parameters, the Gamma distribution has mean 1.5, 1.2, 1.1 and variance of 0.1 in the three datasets. The covariance matrix  $\Sigma$  is specified as follows to resemble real data where almost no correlation was observed:

$$\Sigma = \begin{pmatrix} 1 & -0.1 & 0.1 \\ -0.1 & 1 & -0.1 \\ 0.1 & -0.1 & 1 \end{pmatrix}$$

After simulating the item parameter and latent traits, the response matrix  $X = (x_{ij})$  for a given platform can be generated as:

$$x_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{with } \text{logit}(p_{ij}) = a_i(\theta_j - b_i)$$

The comparison of estimated item parameter and latent traits to the truth is shown in Supplemental Figure 3.15-3.17. The model correctly recovers both item parameters and latent trait estimates. We see that the deviation of latent trait estimates from truth in the expression data is larger than the other two datasets. This is because the items in expression data have larger difficulty parameters (with mean 3 compared to 2 and 1). As the latent traits have mean 0, it is more difficult to estimate item parameters that deviate more from 0. Also, the deviation of latent trait estimates around mean item difficulty (3 in Expr,

2 in Methy, 1 in CN) is smaller than other regions. Intuitively, when a test is too difficult or too easy, it would provide inferior estimates of ability compared to a test designed with proper difficulty level. (*An excerpt from Tong, P. et al, “integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using item response theory”, Bioinformatics, 2012*)

#### **3.5.4 Supplemental Figures**

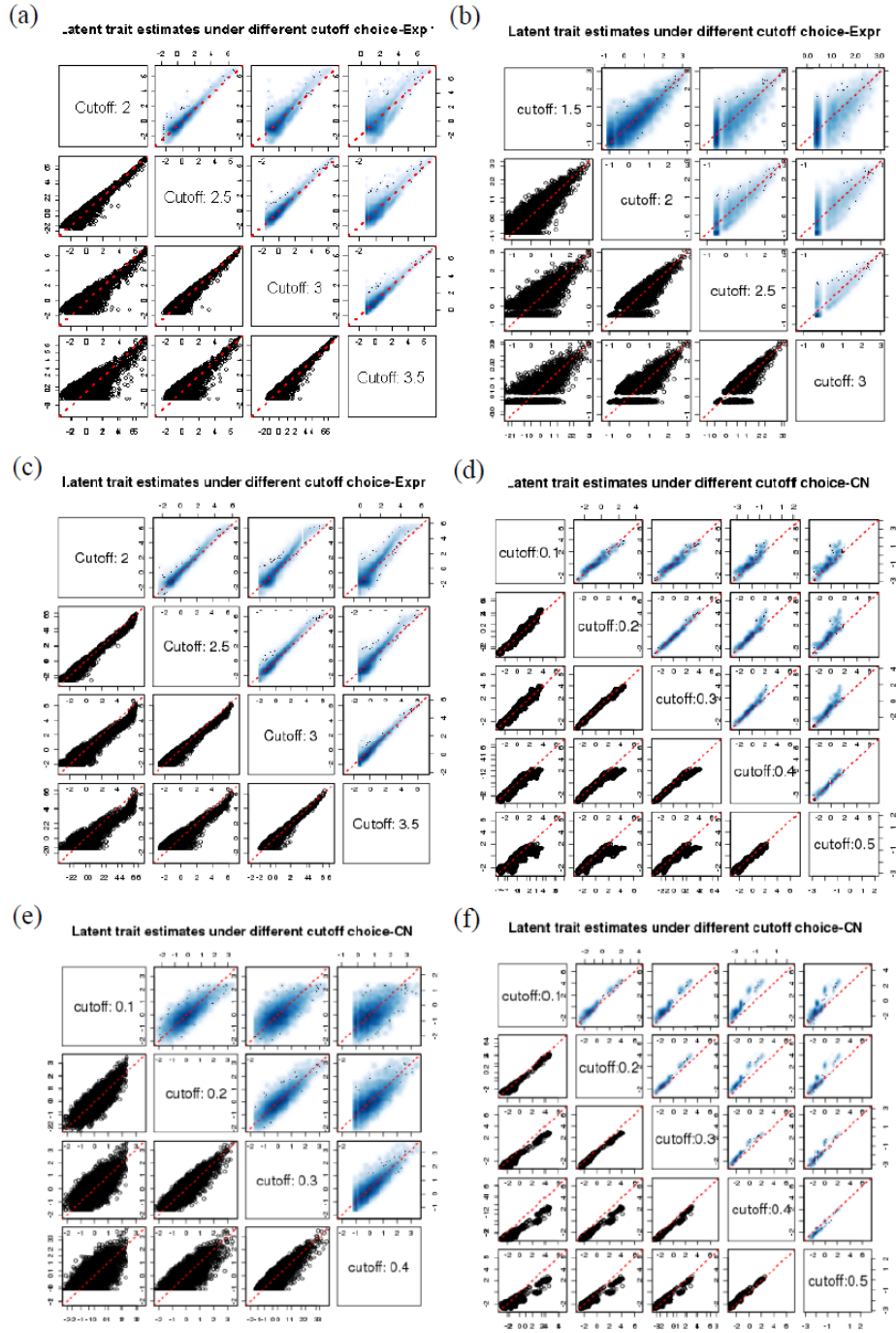


Figure 3.6: Matrix plot of estimated latent traits from OV, BRCA, GBM data under different cutoff choices for data transformation is shown. The upper panel shows smoothed density estimates with blue cloud. The latent trait estimates agreed well, especially in the high ability range. This indicates the proposed method is robust to cutoff choice for data transformation. (a) OV-Expression (b) BRCA-Expression (c) GBM-Expression (d) OV-CN (e) BRCA-CN (f) GBM-CN (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

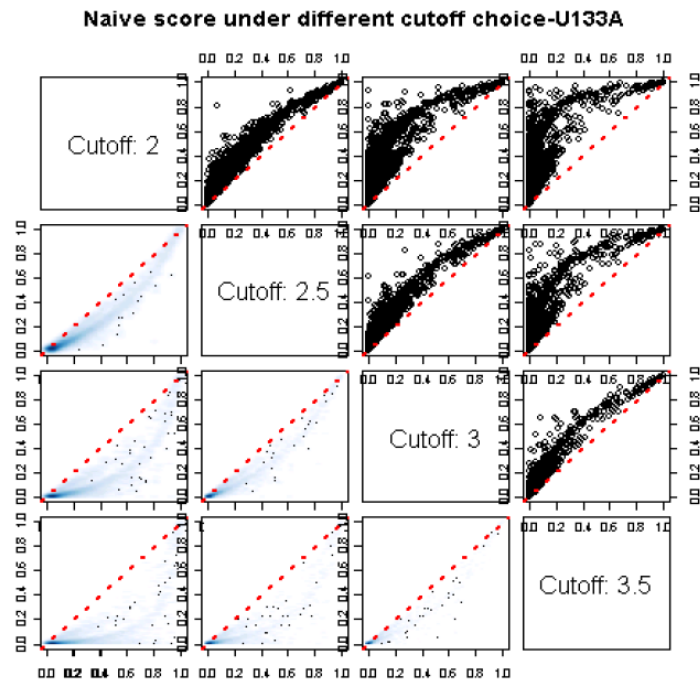


Figure 3.7: Pairwise scatter plot of the naive score obtained by simply computing the percentage of “correct responses” (i.e., alterations) using different thresholds. The OV expression data is shown here. Lower off-diagonal panels show smoothed scatter plot. As the cutoff used for dichotomization changes, the naive score also changes. Therefore, the naive score is sensitive to threshold choice. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)



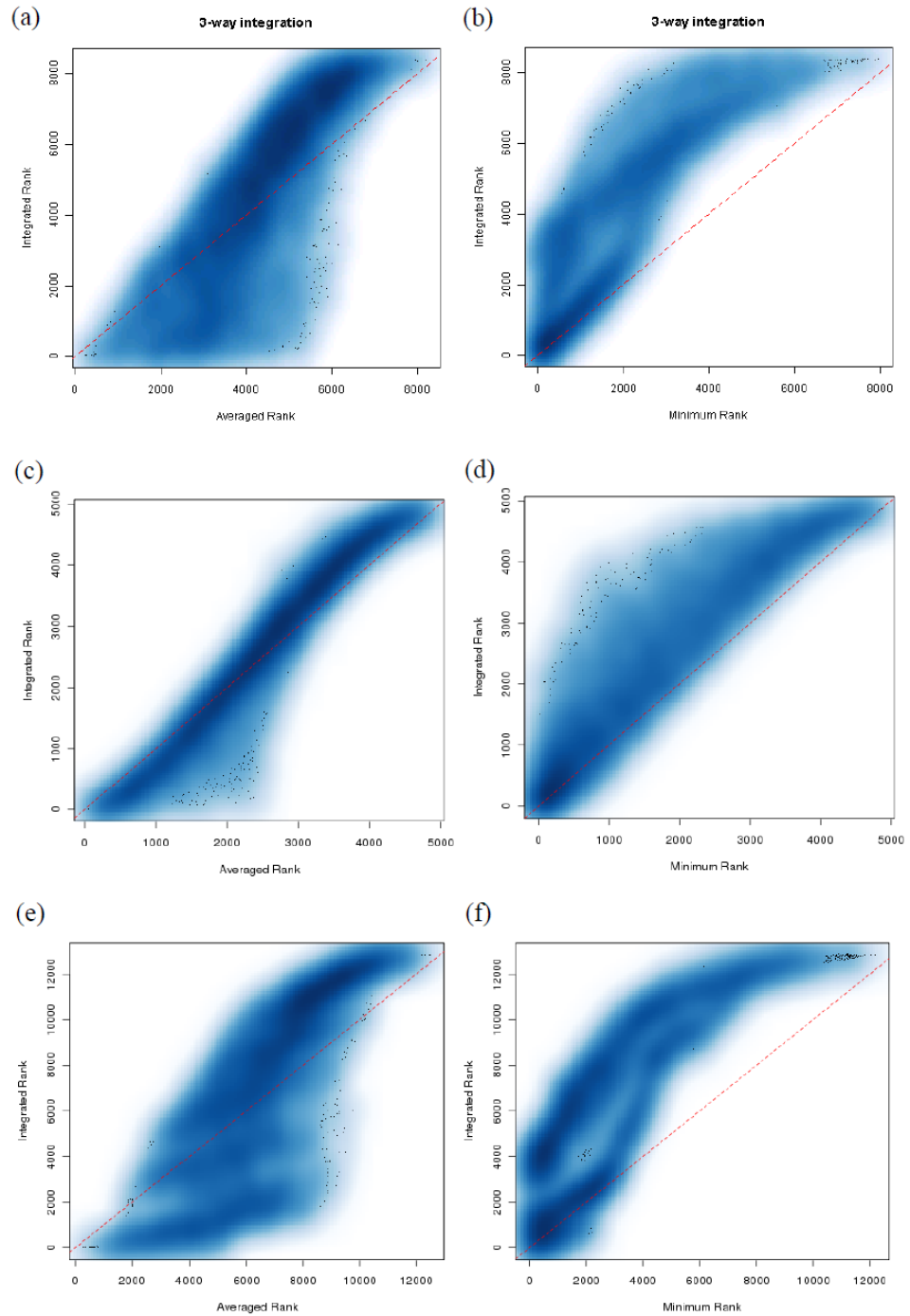


Figure 3.8: Comparison of the rank of integrated latent trait and average or minimum rank from individual platforms. The blue cloud shows the density at each data point. Smaller rank means higher latent trait and hence more severe alteration. (a)~(b) for OV data, (c)~(d) for BRCA data, (e)~(f) for GBM data (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)

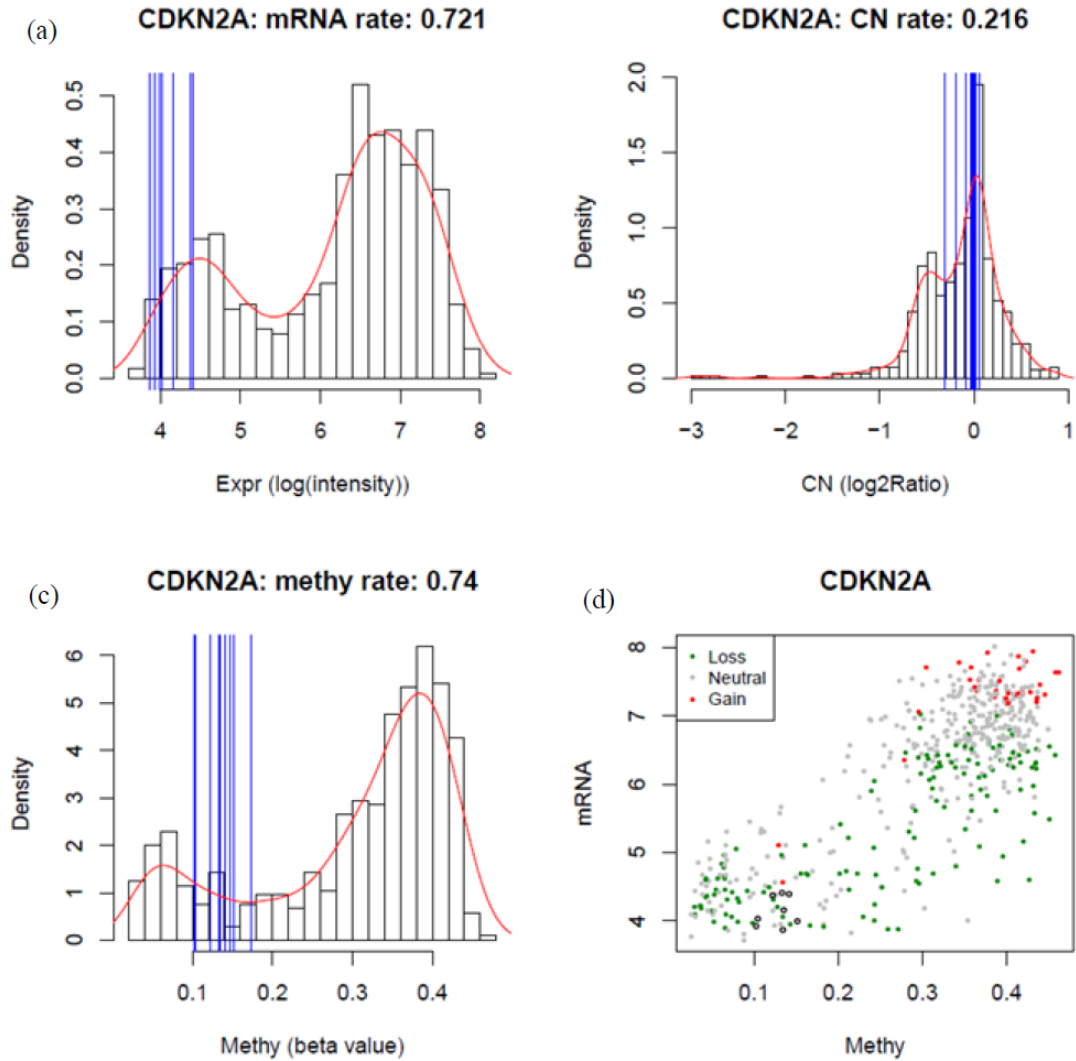


Figure 3.9: Original data for CDKN2A. All platforms suggest CDKN2A is marginally altered while integration makes this gene stand out as top 20. High expression is associated with increased CN and methylation. (a)~(c) are the histograms with estimated kernel density superimposed for Expr, CN and Methy data from tumor samples. The normal samples are indicated by vertical blue lines. (d) shows the scatterplot of Expr and Methy colored by CN (Loss:  $\log_2\text{Ratio} < -0.4$ , Gain:  $\log_2\text{Ratio} > 0.4$ , Neutral: in between). The black circles indicate measurements for available normal samples which only have Expr and Methy data. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

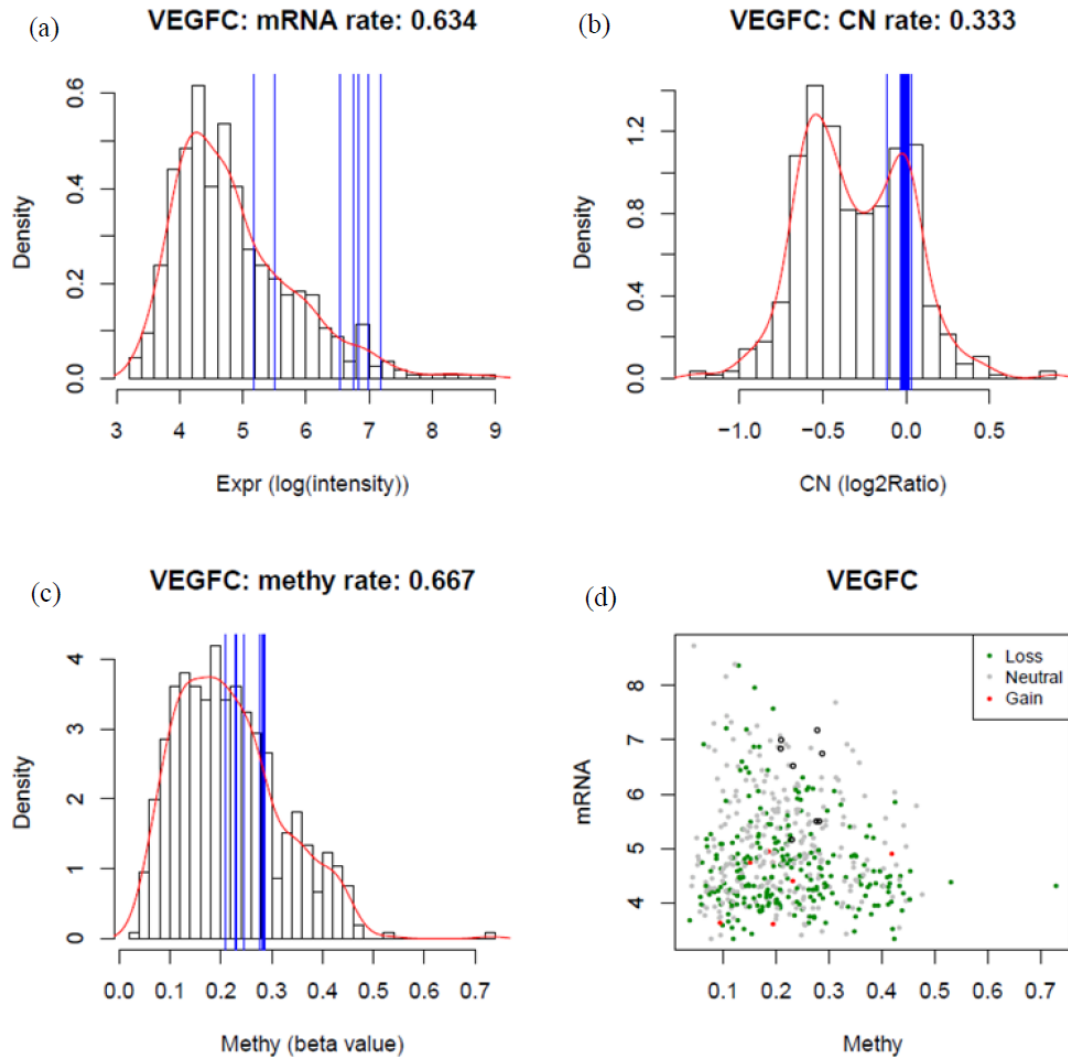


Figure 3.10: Original data for VEGFC. mRNA down-regulation of VEGFC is associated with CN loss and increased methylation. (a)~(c) are the histograms with estimated kernel density superimposed for Expr, CN and Methy data from tumor samples. The normal samples are indicated by vertical blue lines. (d) shows the scatterplot of Expr and Methy colored by CN (Loss:  $\log_2\text{Ratio} < -0.4$ , Gain:  $\log_2\text{Ratio} > 0.4$ , Neutral: in between). The black circles indicate measurements for available normal samples which only have Expr and Methy data. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

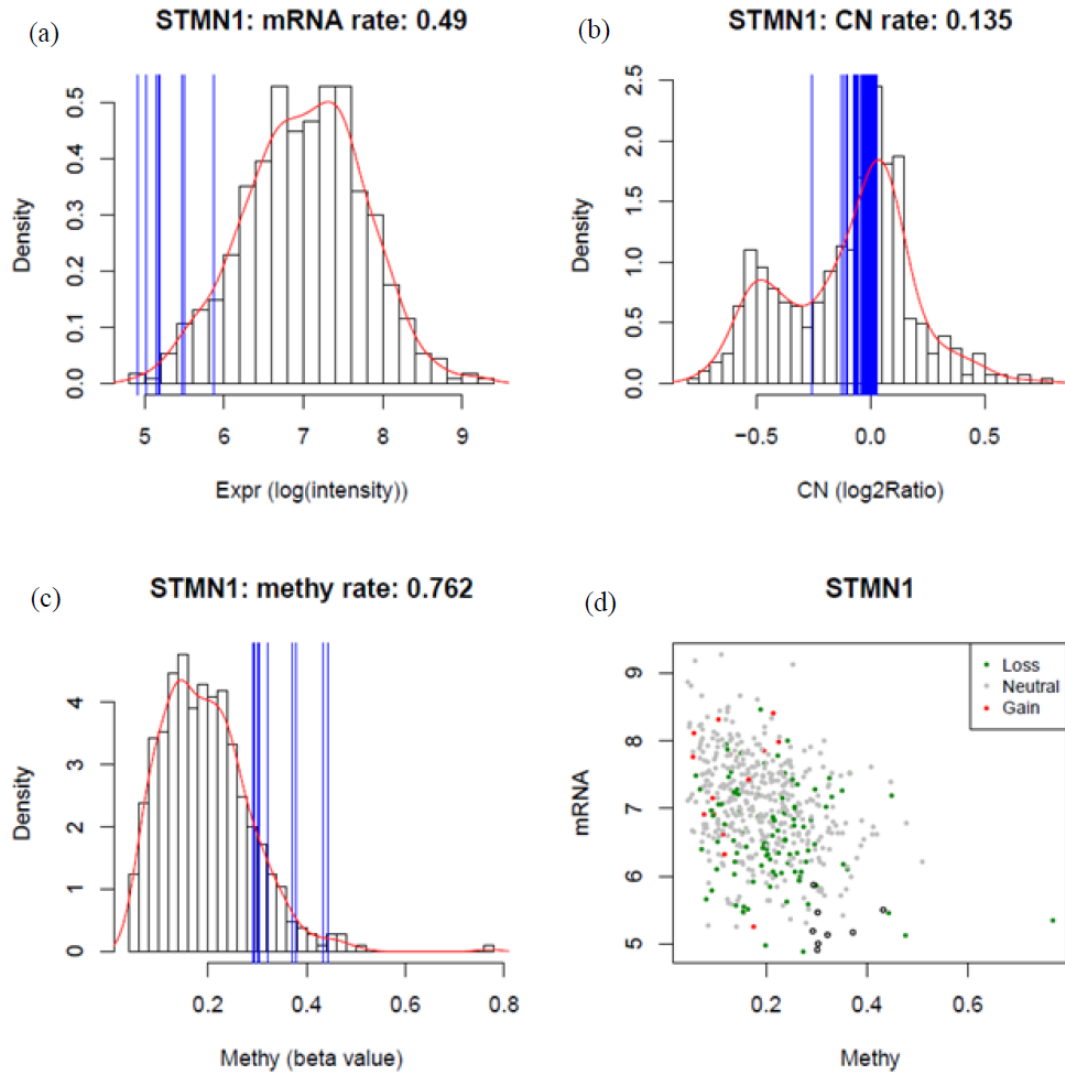


Figure 3.11: Original data for STMN1. Increased expression of STMN1 is associated with decreased methylation and CN gain although there are many samples showing CN loss. (a)~(c) are the histograms with estimated kernel density superimposed for Expr, CN and Methy data from tumor samples. The normal samples are indicated by vertical blue lines. (d) shows the scatterplot of Expr and Methy colored by CN (Loss:  $\log_2\text{Ratio} < -0.4$ , Gain:  $\log_2\text{Ratio} > 0.4$ , Neutral: in between). The black circles indicate measurements for available normal samples which only have Expr and Methy data. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

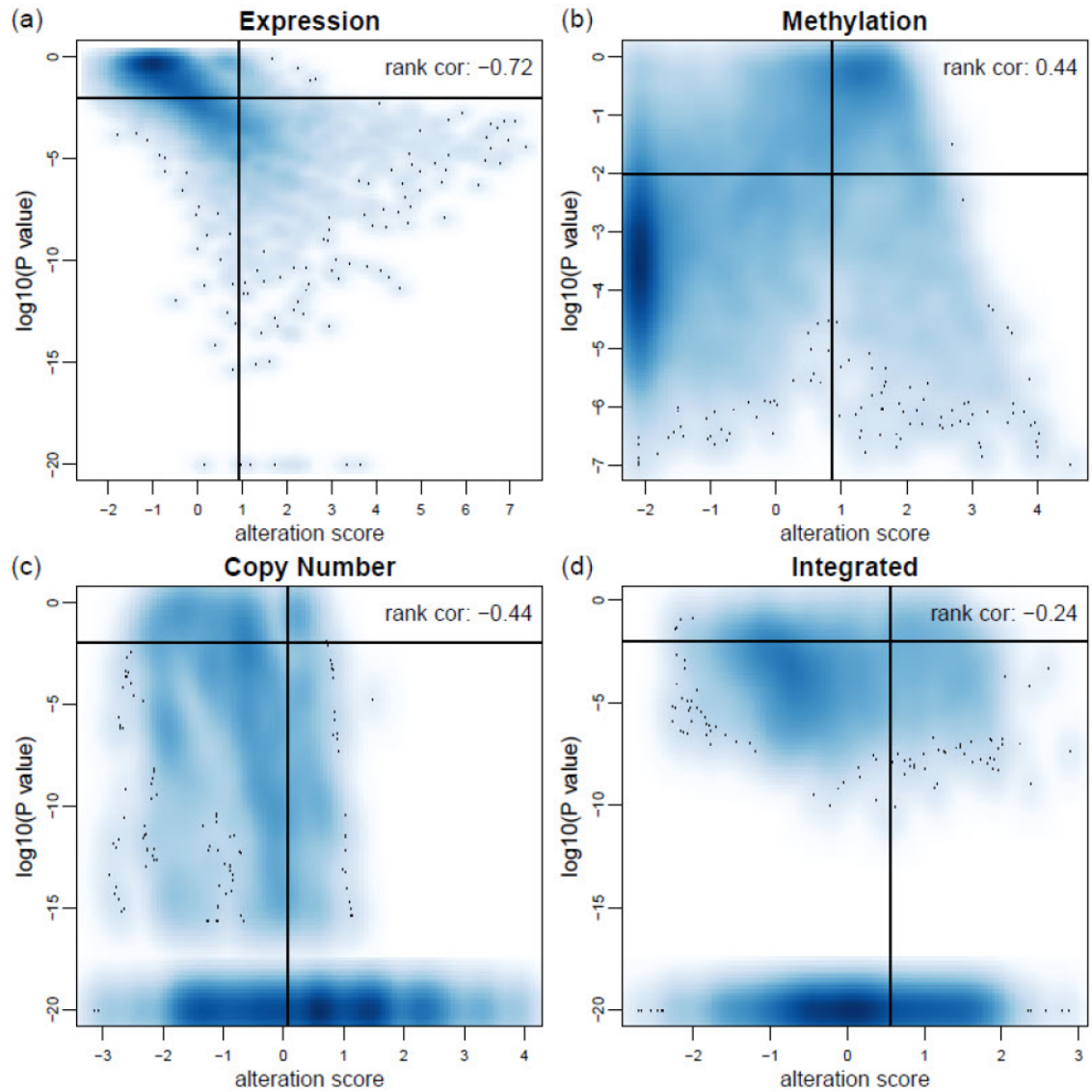


Figure 3.12: Comparison of conventional method and proposed method. The log transformed p-value is compared to latent trait estimates. Smoothed density of the scatterplot is superimposed in each panel. Horizontal solid lines correspond to p-value of 0.01 while vertical solid lines correspond to the 99% quantile of null distribution based on gene sampling (and hence, the empirical p-value of 0.01). Rank correlation between p-value and latent trait is shown on the top of each panel. Examination of the discordant calls (top right and bottom left quadrant) shows our method is more reliable and meaningful than conventional methods (see Fig 4 in the main text). (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

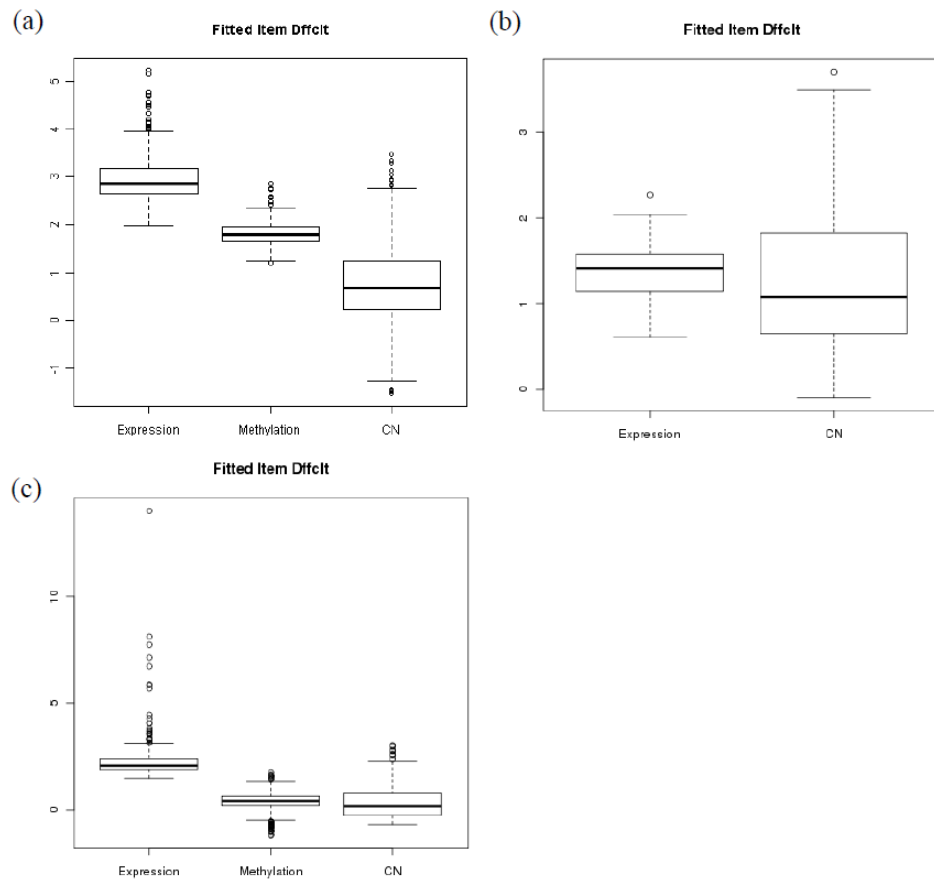


Figure 3.13: Boxplot of item difficulty from different platforms are compared. Samples with small item difficulty contain more alterations. (a) OV data (b) BRCA data (c) GBM data (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

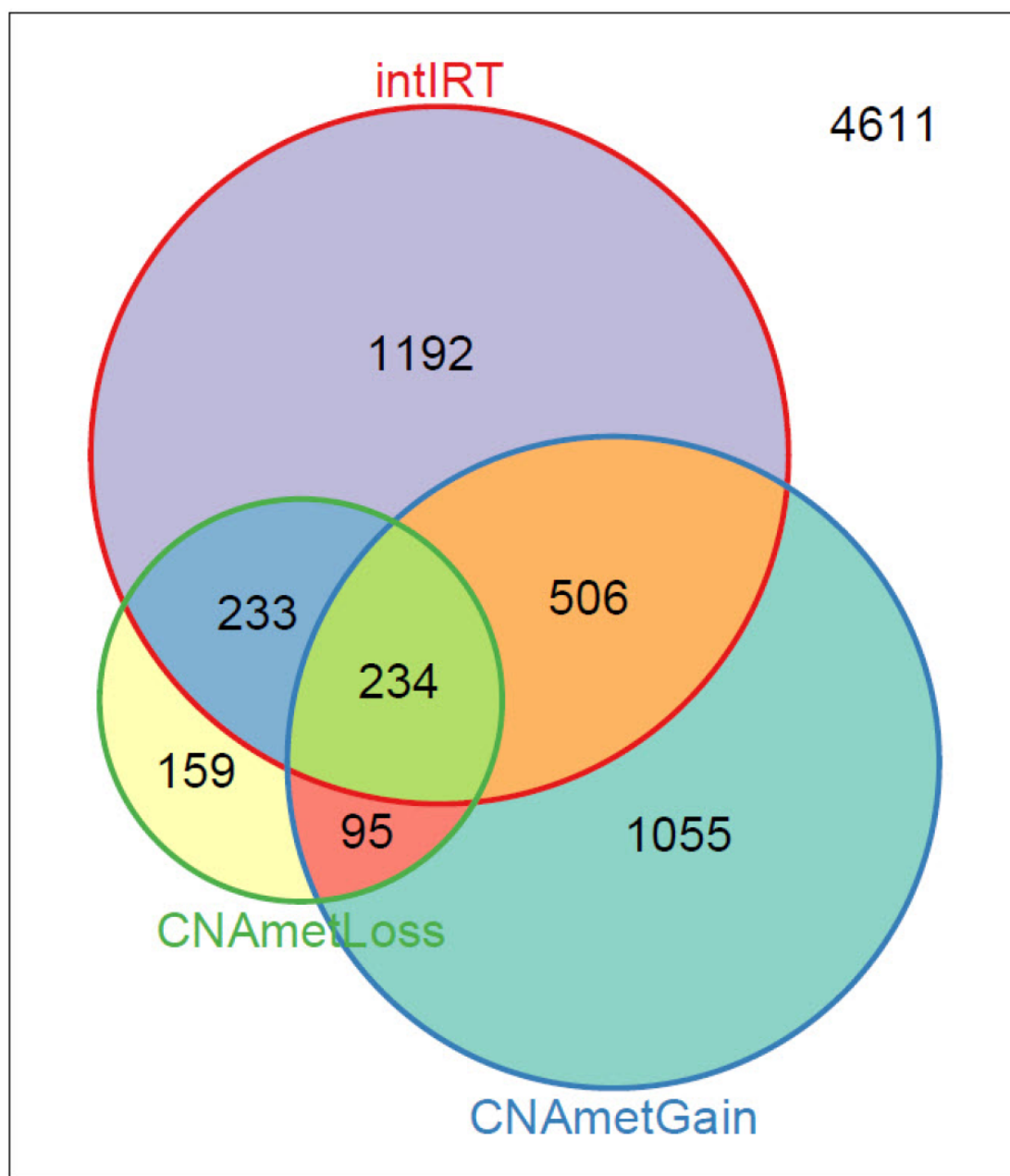


Figure 3.14: Venn diagram of identified genes from **intERTy** and **CNAmet** in OV data. At FDR=0.01, the genes selected by **intERTy** and **CNAmet** (both gain and loss analysis) are shown. Roughly half of the genes identified by **intERTy** and **CNAmet** are shared. Genes identified by **intERTy** but missed by **CNAmet** are usually altered without synergistic regulation while genes missed by **intERTy** but found by **CNAmet** show little difference between tumor and normal. Specific examples found in Fig. 5 of the main text (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)

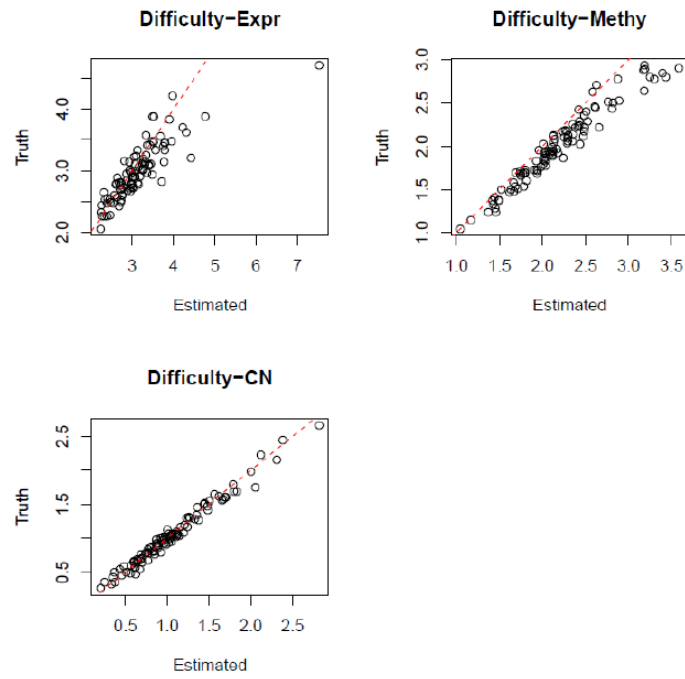


Figure 3.15: Comparison of item difficulty estimates and truth in simulation study (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)

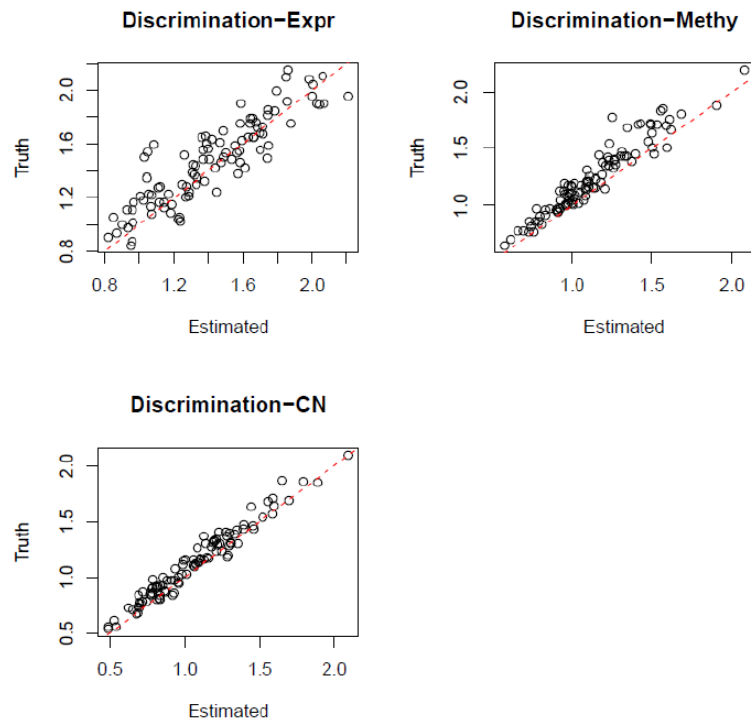


Figure 3.16: Comparison of item discrimination estimates and truth in simulation study (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)



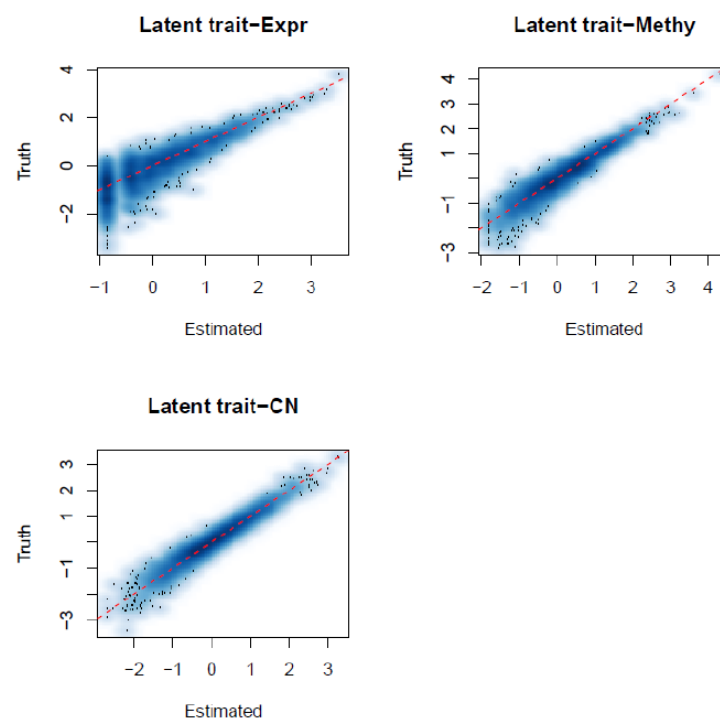


Figure 3.17: Comparison of latent trait estimates and truth in simulation study (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)

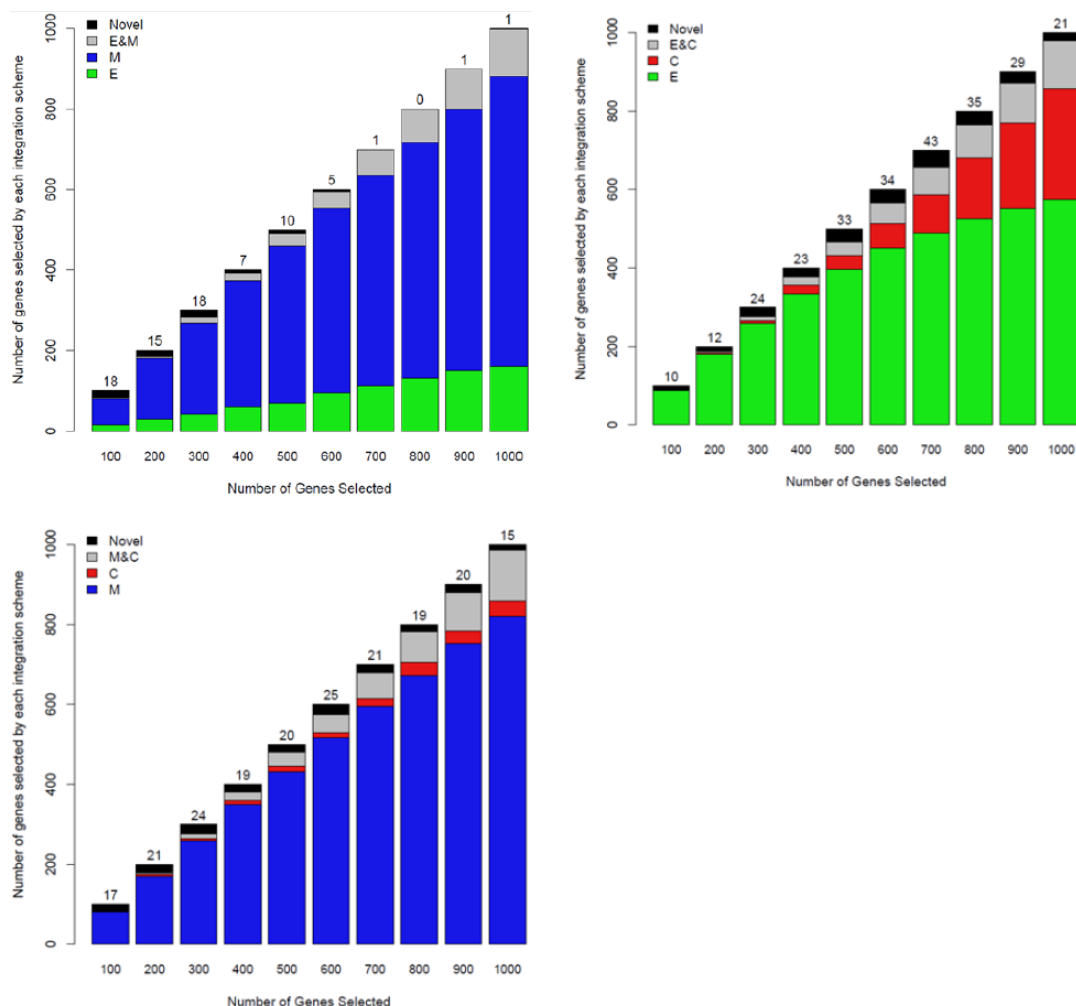


Figure 3.18: Top (100 to 1000) genes from 2-way integrated analysis and from 1-way analysis (no integration) are compared. E: expression, M: methylation, C: copy number. Each bar is equivalent to a Venn diagram showing how many of the top genes from the integrated analysis came from. Black regions and numbers at the top of each bar count the number of “novel” genes that only appear on the list from the integrated analysis. When integrating methylation with either expression or copy number, methylation contributes the majority of the top genes. When expression is integrated with copy number, expression contributes most of the top genes. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2012)

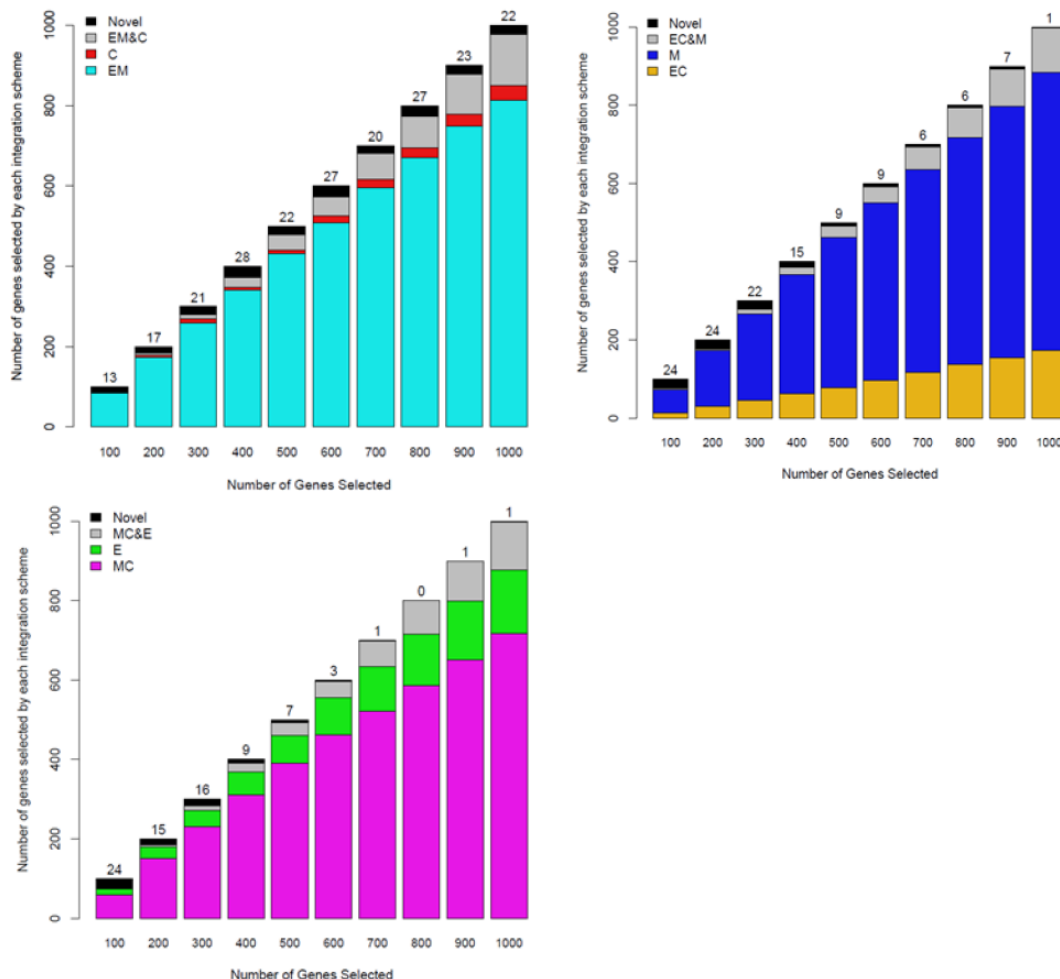


Figure 3.19: Top (100 to 1000) genes from 3-way integrated analysis are compared with genes from 1-way and 2-way integration. Each bar is equivalent to a Venn diagram showing how many of the top genes from the integrated analysis came from. Adding a third assay to 2-way integration introduces several novel genes (black boxes) that are mostly highly altered. Adding methylation to EC integration brings in many genes that are unique to methylation data. This suggests methylation is the most informative assay for gene alteration. (*Figure reprinted from Tong, P. et al, Bioinformatics, 2012*)

### 3.5.5 Supplemental Tables

Table 1: Latent trait and rank for top 20 genes selected by integrated data in BRCA (*Table reprinted from Tong, P. et al, Bioinformatics, 2012*)

Genes	Integrated		Expression		Copy Number	
	Latent trait	Rank	Latent trait	Rank	Latent trait	Rank
CYP4Z1	2.67	1(E)	2.67	9	1.99	92
S100A8	2.56	2(E)	2.93	3	1.37	319
IGHG1	2.54	3(E)	1.96	80	2.82	16
CXXC4	2.29	4(C)	1.74	142	2.44	41
GRB7	2.19	5(E)	2.56	16	1.17	461
CYP4X1	2.18	6(E)	2.52	17	1.24	403
SERPINB5	2.14	7(E)	2.10	58	1.69	171
CD79A	2.13	8(E)	2.18	47	1.55	224
HIST2H2BE	2.11	9(E)	2.40	28	1.21	422
NDRG1	2.10	10(E)	2.42	24	1.15	476
WNT11	2.09	11(C)	1.67	163	2.25	57
SELENBP1	2.09	12(I)	1.91	100	1.87	121
GJB1	2.08	13(E)	2.70	7	0.8	867
EDIL3	2.06	14(I)	1.88	110	1.85	127
POSTN	2.05	15(E)	2.14	55	1.45	273
IFIT1	2.04	16(E)	2.42	23	1.03	594
CEACAM5	2.03	17(E)	2.57	15	0.81	838
CRYM	2.02	18(E)	2.01	72	1.57	219
NPNT	2.00	19(E)	1.94	86	1.62	199
SLCO2A1	1.99	20(C)	1.33	358	2.56	28

Table 2: Latent trait and rank for top 20 genes selected by integrated data in GBM (*Table reprinted from Tong, P. et al, Bioinformatics, 2012*)

	Integrated		Expression		Methylation		Copy Number	
Genes	Latent trait	Rank	Latent trait	Rank	Latent trait	Rank	Latent trait	Rank
PPP2R2D	3.39	1(I)	3.5	321	3.08	283	2.43	262
HECW1	3.24	2(I)	3.89	239	2.81	467	1.26	799
NOS3	3.15	3(I)	3.94	231	2.66	547	1.06	1035
POU6F2	2.95	4(I)	3.44	349	2.35	740	1.27	783
CALN1	2.91	5(I)	3.57	301	1.84	1126	1.35	575
DNAJC12	2.89	6(I)	2.91	533	3.31	137	2.27	340
HK1	2.75	7(I)	3.14	440	1.60	1371	2.43	267
UROS	2.73	8(C)	3.01	494	1.65	1314	2.74	50
CCNY	2.52	9(I)	2.63	675	2.53	621	1.70	392
SLC13A4	2.43	10(M)	2.61	683	3.31	6	1.21	874
ANK3	2.35	11(I)	3.22	415	1.16	2135	2.38	305
C7orf51	2.12	12(I)	2.67	658	1.41	1684	1.28	731
HOXA3	2.12	13(I)	2.65	666	1.41	1676	1.33	605
OPN4	2.11	14(I)	2.12	1053	1.81	1161	2.57	169
MPP7	2.09	15(I)	3.83	255	0.92	2479	1.65	435
NUDT1	2.03	16(I)	2.42	819	1.62	1354	1.08	1026
DIP2C	1.96	17(I)	1.88	1295	3.16	205	1.57	485
AASS	1.96	18(I)	2.41	825	1.42	1590	1.25	833
PKD2L1	1.85	19(I)	1.80	1398	1.66	1310	2.68	106
SNCG	1.79	20(E)	5.51	50	0.63	2897	2.53	201

Table 3: Model selection using BIC (*Table reprinted from Tong, P. et al, Bioinformatics, 2012*)

Dataset	1PL	2PL	3PL
TCGA OV Expr	1493380.0	1483020.0	<b>1482962.0</b>
TCGA OV Methy	906278.7	<b>827892.6</b>	832165.0
TCGA OV CN	927403.8	<b>891969.3</b>	906220.7
BRCA Expr	191718.7	<b>191416.2</b>	191685.1
BRCA CN	222289.8	<b>221699.0</b>	221848.3
TCGA GBM Expr	3210477.0	3137962.0	<b>3136744.0</b>
TCGA GBM Methy	2106436.0	1897085.0	<b>1897061.0</b>
TCGA GBM CN	2772672.0	<b>2467917.0</b>	2535804.0

The minimum BIC value for each dataset is bolded. Note that the difference of BIC values for 2PL and 3PL models in TCGA OV Expr, TCGA GBM Expr and TCGA GBM Methy data is quite small. Hence, for easy comparison and interpretation, we argue 2PL is preferred.

## Chapter 4

### Bimodality identification from RNAseq Data

*(Most of the materials in this chapter have been published online in Bioinformatics, January 2013: Tong, P. et al, “SIBER: systematic identification of bimodally expressed genes using RNAseq data”. According to the journal policy, the author retains the right to include the published article in full or in part in a dissertation.)*

#### 4.1 Background

Chapter 3 introduces the `integIRTy` approach to identify altered genes through integrative analysis. The idea is based on integrating binary signals extracted from different data sources. Besides the extracted binary signal, there are actually natural binary signals in living organisms. One of the most important binary signals is bimodal expression. In this chapter, we develop a novel method to identify bimodal genes from RNAseq data. As will be shown in later chapters, identifying bimodal genes is very important for accurate cancer classification since bimodal genes contain most of the information needed for predicting clinical outcome. Therefore, this chapter serves as a basis towards our goal of building

accurate prediction models by integrating multiple data sources.

By definition, bimodal expression requires the distribution of expression to have two modes representing the baseline expression and a deviation from the baseline such as over-expression or under-expression. It has been found that bimodally expressed genes can capture the heterogeneity among samples and present clear separations between different subgroups in prostate cancer patients [Tomlins et al., 2005]. These features make bimodal genes good candidates for diagnostic and prognostic markers especially in the era of personalized medicine.

The mechanisms driving bimodal expression can be different. These include DNA copy number change, microRNA regulation, DNA methylation, transcription factor binding, histone modification, tissue heterogeneity and batch effects due to technical artifact [Biggar and Crabtree, 2001, Louis and Becskei, 2002, Chen and Widom, 2005]. Similarly, the impact of bimodal expression can be quite different. In normal cells, bimodal expression is required for tissue-specific and temporal-specific expression such that different sets of genes can be turned on and off in different tissues and during different developmental stages. Bimodal expression is also critical to maintain cell signaling. On the other hand, disrupted bimodal expression can lead to uncontrolled cell proliferation and ultimately malignant cancer. It should be noted that technical effects can also lead to bimodal expression. For example, batch effects usually lead to a high percentage of bimodal genes ( $>50\%$ ). Our current analysis assumes the data has high quality such that any bimodal expression is attributable to the underlying biology.

Due to its importance, identifying bimodally expressed genes from whole genome expression assays has become an important topic. Generally speaking, existing methods can be grouped into two categories: nonparametric and normal mixture models. For example, COPA (Cancer Outlier Profile Analysis) is the first



method designed to search for genes with “outlier” expression patterns [Tomlins et al., 2005]. COPA applies a simple transformation of the original data and uses the transformed value to rank the genes. In particular, the expression of each gene is subtracted by its median and scaled with its median absolute deviation, and then ranked by a pre-specified quantile (e.g., 75%, 90% or 95%) of the transformed data. The standardization using median and median absolute deviation rather than the mean and standard deviation is to ensure robustness. This method was applied to 132 datasets and the fusion of ERG and ETV1 was found to be over-expressed in 57% of prostate cancer patients [Tomlins et al., 2005].

The second category models gene expression levels through a mixture of normal distributions. For example, PACK (Profile Analysis using Clustering and Kurtosis) first filters unimodal genes based on model selection with BIC and then ranks bimodal genes using kurtosis [Teschendorff et al., 2006]. Bimodal genes identified by PACK were found to be linked to breast cancer prognosis [Teschendorff et al., 2007]. Rather than using BIC for model selection, the method proposed by Ertel and Tozeren [2008] applies the likelihood ratio test (LRT) to identify bimodal genes. Since the exact null distribution is not available, the original authors use a  $\chi^2$  distribution with six degrees of freedom obtained from simulation studies to calculate approximate p-values. Recently, Wang et al. [2009] proposed a new metric called the Bimodality Index (BI) to rank genes without the need for model selection. This method provides a consistent ranking for all genes using the BI metric that is computed from the fitted parameters from a two-component normal mixture model. The BI approach not only effectively identifies bimodal genes but also provides an intuitive interpretation. It has been shown that the bimodal genes identified by BI define a subset of triple negative breast cancers that might benefit from immune augmentation [Karn et al., 2012].

Existing methods designed for microarray data do not directly work with RNAseq. As next generation sequencing (NGS) becomes more and more popular,

it is important to develop a working method for identifying bimodally expressed genes from RNAseq data. Unlike microarray data that is usually modeled by normal distribution, RNAseq is discrete count data. Usually, discrete distributions such as Poisson and binomial distributions are used to model RNAseq counts. When the samples present large heterogeneity, it is usually necessary to use the negative binomial or generalized Poisson distribution to deal with the observed over-dispersion. The intrinsic difference between RNAseq and microarray data motivates us to develop a method specifically tailored for RNAseq data.

In practice, investigators may first transform the RNAseq data (e.g., log or square root transformation) and treat the transformed data as if it came from microarray. However, the validity of this approach has not been evaluated when identifying bimodal genes. We therefore formally investigate the performance of this approach by applying microarray based approaches including COPA and PACK on the transformed data.

Our method generalizes the original BI approach such that it also works on a mixture of arbitrary distributions such as mixtures of negative binomial, generalized Poisson, or log normal distributions. We evaluate the proposed method through both simulation and real data analysis.

## 4.2 Methods

We propose a two-step procedure to identify bimodally expressed genes. The first step is to fit a two-component mixture model. Specifically, three candidate mixture models are considered. Two of these models explicitly account for the discrete nature of the RNAseq data, whereas the third model treats the data as continuous after some transformation. The second step is to calculate the Bimodality Index corresponding to the assumed mixture distribution.

### 4.2.1 Mixture Models For RNAseq Count Data

We model the observed raw counts using a two-component mixture model, as in Wang et al. [2009]. Denote the raw count for gene  $g$  in sample  $s$  by  $C_{g,s}$  and the true expression by  $\mu_{g,c(s)}$  depending on the component (or cluster) membership  $c(s)$  that sample  $s$  belongs to. Here,  $c(s) = k$  (for  $k = 1, 2$ ) means that sample  $s$  belongs to component  $k$  with mean expression  $\mu_{g,k}$ . To avoid model non-identifiability, we require  $\mu_{g,1} \leq \mu_{g,2}$ . Since each gene is studied separately, we may suppress the index  $g$  for simplicity of notation. We consider three different mixture models.

**Negative Binomial mixture:** Our first model is motivated by the Negative Binomial (NB) distribution which is widely used to model RNAseq data in differential gene expression analysis [Robinson and Smyth, 2007, Anders et al., 2010, Hardcastle and Kelly, 2010, Di et al., 2011]. Of note, we prefer NB over the Poisson distribution because the former can account for the overdispersion observed in RNAseq data. Specifically, the probability of observing count  $C_s$  can be formulated as:

$$\Pr(C_s) = \pi f_{\text{NB}}(C_s; \mu_1, \phi) + (1 - \pi) f_{\text{NB}}(C_s; \mu_2, \phi) \quad (4.1)$$

where  $f_{\text{NB}}(\cdot; \mu, \phi)$  is the probability mass function for the Negative Binomial distribution with mean  $\mu$  and dispersion  $\phi$  (variance =  $\mu + \phi\mu^2$ ):

$$f_{\text{NB}}(y; \mu, \phi) = \frac{\Gamma(\frac{1}{\phi} + y)}{\Gamma(y + 1)\Gamma(\frac{1}{\phi})} \left(\frac{1}{\phi\mu + 1}\right)^{\frac{1}{\phi}} \left(1 - \frac{1}{\phi\mu + 1}\right)^y,$$

and  $\mu_1$  and  $\mu_2$  are the true expression levels for the two components. The parameter  $\phi$  affects the within-group variability. Note that we assume equal dispersion in the two distributions,<sup>1</sup> similar to the tagwise dispersion mode in EdgeR [Robinson and Smyth, 2007, Robinson et al., 2010]. When the dispersion parameter  $\phi = 0$ , equation(4.1) reduces to a mixture of Poisson distributions. The parameters  $(\pi, \mu_1, \mu_2, \phi)$  can be estimated by maximizing the likelihood function:

$$L(\pi, \mu_1, \mu_2, \phi | C_s) = \prod_{s=1}^n \{\pi f_{\text{NB}}(C_s; \mu_1, \phi) + (1 - \pi) f_{\text{NB}}(C_s; \mu_2, \phi)\}$$

The Expectation-Maximization (EM) algorithm or direct optimization can be used for this purpose.

**Generalized Poisson mixture:** The Generalized Poisson (GP) dis-

---

<sup>1</sup>We do not assume equal variance because, for NB distribution, the variance depends on the mean. Assuming equal variance would impose an undesirable constraint on the component means.

tribution is another model used to describe RNAseq count data [Srivastava and Chen, 2010]. Under the two-component mixture framework, it can be formulated similarly as:

$$\Pr(C_s) = \pi f_{\text{GP}}(C_s; \mu_1, \phi) + (1 - \pi) f_{\text{GP}}(C_s; \mu_2, \phi) \quad (4.2)$$

where  $f_{\text{GP}}(\cdot; \mu, \phi)$  is the probability density function for the Generalized Poisson distribution with mean  $\mu$  and dispersion  $\phi$  (variance  $= \phi\mu$ )

$$f_{\text{GP}}(y; \mu, \phi) = \frac{\mu}{\sqrt{\phi}} \left\{ \frac{\mu}{\sqrt{\phi}} + \left(1 - \frac{1}{\sqrt{\phi}}\right)y \right\}^{y-1} \exp\left\{-\frac{\mu}{\sqrt{\phi}} - \left(1 - \frac{1}{\sqrt{\phi}}\right)y\right\} / y!,$$

and  $\mu_1$  and  $\mu_2$  are the true expression levels for the two components. For similar reasons to the NB model, we assume equal dispersion between the two components. Note that the variance of the GP distribution is a linear function of its mean, whereas the variance of the NB distribution is a quadratic function of the mean. When  $\phi = 1$ , the GP distribution reduces to Poisson. As a result, a mixture of Poisson distribution is automatically included in the GP model.

**Normal mixture with Box-Cox transformation:** Instead of accounting for the discrete nature of the RNAseq data as in models (1) and (2), we could treat the data as normal after some transformation. A wide class of transformations was proposed by Box and Cox [1964], known as the Box-Cox transformation or power transformation,

$$y_i^{(\lambda)} = \begin{cases} (y_i^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

As suggested by the data empirically, the optimal choice of  $\lambda$  for RNAseq data considered in this paper is  $\lambda = 0$ , which corresponds to the log-transformation (details in Section 4.3.2.1). This leads to our third model as a mixture of lognormal (LN) distributions,

$$\Pr(C_s) = \pi f_{\text{LN}}(C_s; \mu_1, \sigma^2) + (1 - \pi) f_{\text{LN}}(C_s; \mu_2, \sigma^2) \quad (4.3)$$

where  $f_{\text{LN}}(\cdot; \mu, \sigma^2)$  is the probability density function for Lognormal distribution with mean  $\mu$  and variance  $\sigma^2$  at the log scale. The variances of the two log transformed distributions are assumed to be equal, similar to the mixture of normals considered in Wang et al. [2009]. We note that the log-transformation has been used previously to analyze RNAseq experiments [Cloonan et al., 2008, Lee et al., 2011, McIntyre et al., 2011].

### 4.2.2 Generalized Bimodality Index

The Bimodality Index [Wang et al., 2009] is defined as:

$$\text{BI} = \sqrt{\pi(1-\pi)} \frac{|\mu_1 - \mu_2|}{\sigma} \quad (4.4)$$

where  $\delta = |\mu_1 - \mu_2|/\sigma$  is the effect size that measures the distance between the two-components. The coefficient  $\sqrt{\pi(1-\pi)}$  is maximized at  $\pi = 0.5$  and hence penalizes unbalanced allocation into the two components. A limitation of the original BI is that it is defined based on a normal mixture with equal variance. It does not apply to normal mixtures with unequal variance or to genes whose expression values do not follow normal distributions (e.g., discrete distributions as in RNAseq data). In order to deal with these situations, here we generalize the original BI. The definition of BI in Wang et al. [2009] was motivated by sample size considerations. For a normal mixture with unequal variances, similar calculations tell us:

$$\text{BI}^2 = \frac{\pi(1-\pi)(\mu_1 - \mu_2)^2}{(1-\pi)\sigma_1^2 + \pi\sigma_2^2} = \frac{(Z_{\alpha/2} + Z_\beta)^2}{N} \quad (4.5)$$

Hence, the generalized BI can be calculated by:

$$\text{BI} = \sqrt{\pi(1-\pi)} \frac{|\mu_1 - \mu_2|}{\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}} \quad (4.6)$$

Formula (4.6) is quite similar to formula (4.4) except the effect size is modified as:  $\delta = |\mu_1 - \mu_2|/\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}$ . Note that when  $\sigma_1 = \sigma_2$ , the generalized BI formula reduces to (4.4).

For other mixture models, deriving the exact BI formula directly from sample size considerations is a tedious task, and the resulting BI may be quite complicated. In general, there is no closed form for BI. However, we can instead obtain a formula for BI under large sample approximation. It turns out that, even for a mixture of discrete distributions, we can obtain the same BI formula as in (4.6) by using the Central Limit Theorem. Details are provided in Appendix Section 4.5.1.

The generalized BI for normal mixtures with unequal variance has a sample size interpretation, as indicated in formula (4.5) where  $\alpha$  and  $\beta$  are the type I and type II error and  $N$  is the sample size.  $Z_\alpha$  determines the quantile for a standard normal distribution that has right tail probability being  $\alpha$ . In a typical microarray or RNAseq experiment, the sample size  $N$  is predetermined and BI can be computed for each gene. Different values of BI then represent different type I and type

II error. To reliably detect useful bimodal genes, BI needs to exceed a certain threshold that is determined by  $N$ .

### 4.2.3 Adjusting For Library Size and Gene Length Effect

Similar to microarray data, RNAseq data requires proper normalization in order to make meaningful comparisons between samples. A common practice is to scale the raw counts by both the gene length  $L_g$  for gene  $g$  and the total reads  $T_s$  in sample  $s$ , giving the so-called RPKM value [Mortazavi et al., 2008]. For this reason, we introduce a normalization term,  $d_{g,s}$ , into our mixture models. This term accounts for technical effects including lane, flow-cell, and library preparation effects. In the case of RPKM,  $d_{g,s} = L_g T_s$ . Directly scaling the count data by  $d_{g,s}$  would transform the data onto a continuous scale that cannot be modeled by NB or GP distributions. Instead, we incorporate  $d_{g,s}$  through the expected count as:

$$E[C_{g,s}|d_{g,s}, c(s)] = d_{g,s} \mu_{g,c(s)}$$

Hence, we only need to replace the component distribution  $f(C_s; \mu_{c(s)}, \phi)$  in Section 4.2.1 with  $f(C_s; d_{g,s} \mu_{c(s)}, \phi)$ . The rest of the inference remains the same. As pointed out by Bullard et al. [2010], RPKM normalization performs poorly when there are highly differentially expressed genes. More robust normalization methods such as TMM [Robinson and Oshlack, 2010] and the method used in DESeq or DEXseq can be applied [Anders et al., 2010, 2012]. Inclusion of such normalization methods to our models is quite similar, only adding a scaling factor to the component means.

*(This section is an excerpt from Tong, P. et al, “SIBER: systematic identification of bimodally expressed genes using RNAseq data”, Bioinformatics, 2013)*

## 4.3 Results

We let NB, GP, and LN denote the three models described above. For each model, let  $BI_{NB}$ ,  $BI_{GB}$ , and  $BI_{LN}$  be the generalized BI computed with respect to that model. The fundamental question to be addressed is which model yields more robust and more reliable identification of bimodal genes from RNAseq data. To evaluate the performance of different models, and to compare to alternative methods, we first

conduct simulation studies. We generate artificial RNAseq data from one of the three models. Regardless of which model is used to generate the data, we compute BI using all three models. This procedure allows us to evaluate the performance of BI under misspecified models; this step is important because the true underlying model for RNAseq data is often unknown in practice.

In the second scenario, we look at TCGA data where both microarray and RNAseq data are available for the same set of breast cancer samples. We establish the “true” bimodal status for a subset of genes by applying the existing methods to the microarray data, then manually confirming the results by visually inspecting the distributions. For this subset of genes, the misclassification rates (of genes as bimodal or unimodal) are expected to be low. We then compute BI from the RNAseq data using all three models. Because there is a good correspondence between microarray and RNAseq data [Marioni et al., 2008], we can evaluate the performance of the BI models by constructing receiver operating characteristic (ROC) curves that test their ability to correctly match the microarray-based gene classifications.

### 4.3.1 Simulation Study

In this subsection, we consider RNAseq data generated from one of the three mixture models, which will be referred to as NB, GP and LN datasets, respectively.

#### 4.3.1.1 NB, GP and LN Datasets

For each of the NB, GP, and LN datasets, we simulate both bimodal and unimodal genes, which are generated from two-component and one-component mixture models, respectively. To cover a spectrum of settings in practice, we allow the mixture proportion parameter  $\pi$ , the effect size  $\delta$ , and the sample size to vary. Since we know the true status of the generated gene data, we can construct ROC curves that evaluate the ability of the BI models to correctly predict the true status. The performance of  $BI_{NB}$ ,  $BI_{GP}$  and  $BI_{LN}$  will be evaluated using the area under the corresponding ROC curves (AUC).

**Bimodal genes:** For the bimodally expressed genes, we choose different combinations of parameters in order to represent a wide range of bimodal shapes. Specifically,  $\pi$  takes values between 0.1 and 0.5 with

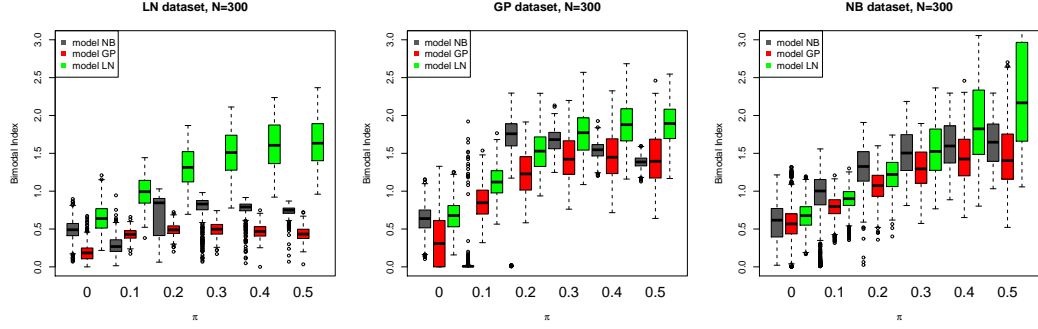


Figure 4.1: Bimodal Index (BI) as a function of the size ( $\pi$ ) of the smaller group. For each of the three models, we simulated datasets with a range of different distances ( $\delta = 2.5, 3, 3.5, 4$ ) and applied all three models to compute BI. Boxplots for  $\pi = 0$  give the distribution of BI when the data is simulated from a unimodal distribution. Performance under the correctly specified model is similar for all three methods, with equal splits ( $\pi = 0.5$ ) yielding the largest BI values. The NB model (gray) performs extremely poorly under misspecified models, with BI values for  $\pi = 0.1$  clearly less than the unimodal BI values and peak BI when  $\pi = 0.2$ . The GP model (red) performs poorly on data simulated from the LN model. The LN model (green) performs consistently regardless of how the data is simulated. (*Figure reprinted from Tong, P. et al, Bioinformatics, 2013*)

a step of 0.1 ( $\pi = 0.6, \dots, 0.9$  are omitted by symmetry). In practice,  $\pi = 0.1$  or  $0.2$  leads to an unbalanced mixture while  $\pi = 0.3, 0.4$  or  $0.5$  leads to more balanced mixture distribution. We also use a range of effect sizes,  $\delta = 2.5, 3, 3.5, 4$ . To simulate genes that have different expression levels, we set  $\mu_1 = 5$  for the LN model (corresponding mean on the exponential scale is 244.7),  $\mu_1 = 100, 1000, 5000, 10000$  for the NB model and  $\mu_1 = 100, 1000, 2000, 4000$  for the GP model. For LN, we set  $\sigma = 1$  due to the equal variance assumption (the corresponding variance on the exponential scale is 34.5). We assume equal dispersion between the two groups for both NB and GP models. As a result, we set the dispersion parameter  $\phi = 0.1$  for the NB model and  $\phi = 0.5\mu_1$  for the GP model. This implies that in both NB and GP models, the variance is a quadratic function of the mean, as typically seen in RNAseq experiments (see the mean-variance relationship in Supplemental Figure 4.19). We use equation (4.6) to solve for  $\mu_2 = \mu_1 + \delta\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}$ . All possible combinations of  $\pi, \mu_1$  and  $\delta$  are considered in each model, which results in 20 ( $5 \times 1 \times 4$ ) settings in LN datasets and 80 ( $5 \times 4 \times 4$ ) settings in NB and GP datasets. The parameters were chosen to mimic real data. For each setting, we simulate 100 genes which leads to 2000 bimodal genes in LN datasets and 8000 genes in NB or GP datasets. We choose four different sample size settings ( $N=50, 100, 200$  and  $300$ ) for each dataset. Data generated from the LN model is continuous, but is rounded to the nearest integer.

**Unimodal genes:** The unimodal genes are simulated from a one-component model. To match the parameter settings for the bimodal genes, we generate unimodal genes from the larger component corre-



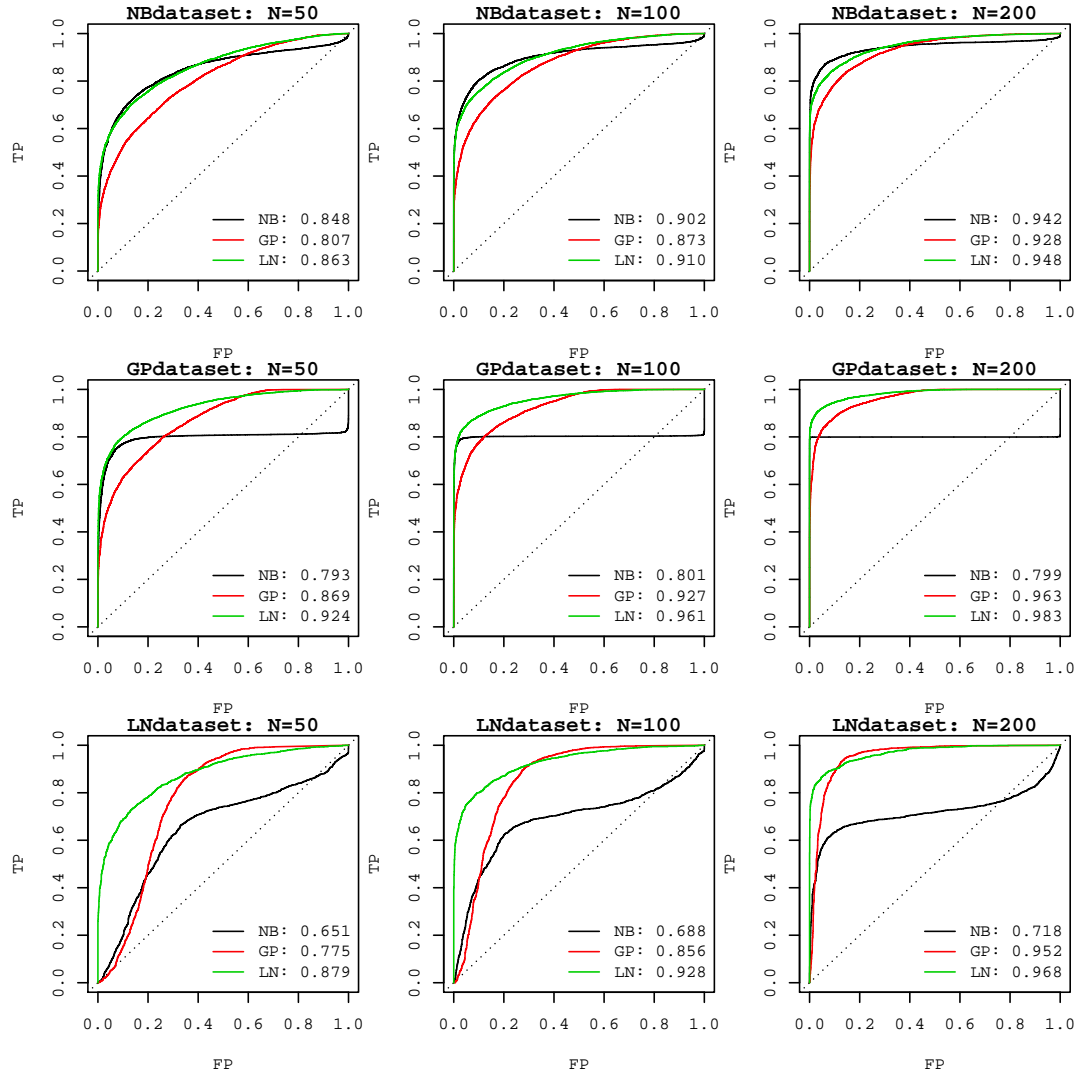


Figure 4.2: Robustness of NB, GP and LN models. ROC curves for the three mixture models fitted on NB, GP and LN datasets are compared under sample size  $N=50, 100$ , and  $200$  ( $N=300$  is omitted due to space limitations). Various bimodal shapes as characterized by different distances ( $\delta = 2.5, 3, 3.5, 4$ ) and component size ( $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ ) are simulated to mimic real data. The LN model is most robust and provides satisfactory performance even when the model is misspecified. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

sponding to the mixture proportion greater than 0.5. Equal numbers of unimodal and bimodal genes are simulated and combined to form LN, NB and GP datasets.

#### 4.3.1.2 Effect of the Mixture Proportion

To examine the effect of the parameter  $\pi$  that describes the proportion of samples in the smaller group, we prepared box-and-whisker plots of BI as a function of  $\pi$  (Figure 4.1). For each of the three datasets (for  $N = 300$ ), we fit three mixture models to obtain parameter esti-

mates and calculate BI. The distributions when the data come from a unimodal distribution are included at  $\pi = 0$ . Ideally, we expect BI to increase as  $\pi$  increases from 0 to 0.5. We see this behavior when we analyze each dataset using the same model that was used to simulate it. The NB model, however, exhibits different behavior when the model is misspecified; BI values for  $\pi = 0.1$  are lower than in the unimodal case and peak when  $\pi = 0.2$ . The GP model behaves well on the GP and NB datasets, but behaves poorly when the true model is LN. The LN model, by contrast, has similar performance regardless of which model was used to simulate the data.

#### 4.3.1.3 Performance Evaluation Metrics

A useful measure to evaluate the performance of BI is the ROC curve. Figure 4.2 and Supplemental Figure 4.5 plot the ROC curves for the three BIs, namely  $BI_{NB}$ ,  $BI_{GP}$  and  $BI_{LN}$ , in NB, GP and LN datasets for different sample sizes. Specifically, Supplemental Figure 4.5 shows the performance when the assumed distribution is correct, whereas Figure 4.2 shows the performance under misspecified models. Note that the evaluation using AUC is limited since it puts the performance under different False Positive (FP) rates on an equal footing. Hence, to explicitly control FP, we also evaluate the performance by looking at the power under predefined FP rates in Tables 4.1-4.3.

#### 4.3.1.4 Performance Under Correctly Specified Models

Supplemental Figure 4.5 shows that when the model is correctly specified, the three methods perform similarly in terms of AUC. Even when the sample size  $N = 50$ , the ROC curve is still satisfactory. For FP rate and power under the correctly specified model, details are given in column  $BI_{NB}$  in Table 4.1, column  $BI_{GP}$  in Table 4.2 and column  $BI_{LN}$  in Table 4.3. Note that the largest power in each row is bolded. We see that under the correctly specified model,  $BI_{NB}$ ,  $BI_{GP}$  and  $BI_{LN}$  all perform reasonably well. In all three models, increased sample size improves power by reducing the cutoff of BI when controlling the same type I error. For the same sample size and type I error, the LN model has slightly larger power (in average 1.9% larger) than the NB model. Both the LN and the NB model perform significantly better than the GP model (the power of NB is 8.9% larger than GP) when  $FP < 0.1$ . Nevertheless, the GP model has larger power at larger FP which makes its AUC better than the NB model and almost matches that of the LN model (Supplemental Figure 4.6). There is a minor

decline of AUC in the NB and GP models. This happens because the BI formula for the NB and GP models relies on large sample approximation and hence loses some efficiency. Note that the FP rate and power under different sample sizes for the LN model agrees with the results in Wang et al. [2009].

#### 4.3.1.5 Performance Under Misspecified Models

In practice, the true underlying model is often unknown. It is important to investigate the robustness of the proposed BIs under a misspecified model. Here we compare the performance of the three models and study the effect of model misspecification. Figure 4.2 shows the ROC curves under misspecified models. We see that the performance varies suggesting different robustness among the three models.

For the NB datasets where the true model generating the simulated data is the NB model, Table 4.1 shows the smallest BI needed to achieve a given type I error and corresponding power to detect bimodal genes for the three mixture models with  $N=50, 100, 200$  or  $300$ .  $BI_{LN}$  provides competitive performance compared to the true model while  $BI_{GP}$  performs much worse. For the GP datasets,  $BI_{LN}$  even dominates  $BI_{GP}$  at all FP rate and sample sizes listed in Table 4.2 (however, not at all possible FP rates).

For the LN datasets, the power of  $BI_{NB}$  and  $BI_{GP}$  stemming from misspecified models is much lower than that of  $BI_{LN}$  under the same FP rate (Table 4.3). More importantly, when the sample size is small or moderate, the power of  $BI_{NB}$  and  $BI_{GP}$  is even smaller than the FP rate or half of the power achieved by  $BI_{LN}$  at best. When the sample size is relatively large, i.e.  $N=200$ , the performance of  $BI_{GP}$  improves and almost matches  $BI_{LN}$  (AUC: 0.95 versus 0.97, see Figure 4.2). However, increasing sample size only improves the power of  $BI_{NB}$  at low FP rate while decreasing the power at high FP rate. Overall, the AUC of  $BI_{NB}$  only increases slightly with sample size. The reason is that the fitted NB model fails to detect most bimodal genes with  $\pi = 0.1$  in the GP and LN datasets (Figure 4.1). These results suggest that  $BI_{NB}$  and  $BI_{GP}$  are highly sensitive to model misspecification. Hence, from the spectrum of settings considered,  $BI_{LN}$  outperforms the other two methods in terms of power under the correctly specified model as well as robustness under a misspecified model.

Table 4.1: Performance on NB datasets (*Table reprinted from Tong, P. et al, Bioinformatics, 2013*)

FP	N	<b>BI<sub>NB</sub></b>		<b>BI<sub>GP</sub></b>		<b>BI<sub>LN</sub></b>	
		BI cutoff	Power	BI cutoff	Power	BI cutoff	Power
0.01	50	1.605	0.395	1.540	0.264	1.608	<b>0.438</b>
	100	1.347	<b>0.590</b>	1.312	0.384	1.325	0.589
	200	1.144	<b>0.772</b>	1.123	0.549	1.145	0.707
	300	1.054	<b>0.842</b>	1.006	0.666	1.042	0.771
0.05	50	1.410	0.582	1.344	0.418	1.394	<b>0.584</b>
	100	1.199	<b>0.732</b>	1.138	0.556	1.181	0.698
	200	1.030	<b>0.856</b>	0.966	0.706	1.026	0.798
	300	0.946	<b>0.901</b>	0.880	0.788	0.948	0.851
0.10	50	1.301	<b>0.678</b>	1.226	0.524	1.287	0.659
	100	1.111	<b>0.802</b>	1.039	0.652	1.100	0.757
	200	0.958	<b>0.891</b>	0.886	0.786	0.958	0.850
	300	0.882	<b>0.924</b>	0.814	0.854	0.894	0.890

#### 4.3.1.6 Difficulty in Identifying the True Model

In general, it is desirable to identify the true underlying model (e.g., NB, GP, LN, or other models). However, this task is extremely challenging (and perhaps impossible) in practice. For example, when BIC is used as the criterion for model selection, the BICs from NB, GP and LN models are almost indistinguishable for all three simulated datasets (Supplemental Figure 4.8). Compared to misspecified models, the true model does not show a clear advantage in terms of BIC. In this sense, each of the three models provides similar fits for the data, despite the fact that they have different performance in terms of identifying bimodal genes. This finding suggests that robustness of BI is important due to the practical difficulty in identifying the true model.

#### 4.3.1.7 Robustness to Outlier Data

In practice, microarray and RNAseq data often contain outliers due to various technical artifacts such as library preparation and amplification bias as well as biological variations that makes the expression (RNAseq data after log transformation) deviate from the assumed normal distribution. Ignoring these outliers might lead to false positive calls. Therefore, in addition to examine the robustness to model misspecification, we also examine the robustness to outlier data. We consider two kinds of outlier data, namely data of heavy tailed dis-

Table 4.2: Performance on GP datasets (*Table reprinted from Tong, P. et al, Bioinformatics, 2013*)

FP	N	<b>BI<sub>NB</sub></b>		<b>BI<sub>GP</sub></b>		<b>BI<sub>LN</sub></b>	
		BI cutoff	Power	BI cutoff	Power	BI cutoff	Power
0.01	50	1.521	0.510	1.515	0.336	1.604	<b>0.586</b>
	100	1.275	<b>0.759</b>	1.277	0.520	1.356	<b>0.759</b>
	200	1.093	0.799	1.154	0.632	1.157	<b>0.869</b>
	300	0.992	0.799	1.031	0.742	1.068	<b>0.915</b>
0.05	50	1.323	0.714	1.296	0.521	1.406	<b>0.732</b>
	100	1.126	0.796	1.102	0.688	1.205	<b>0.846</b>
	200	0.969	0.799	0.929	0.829	1.047	<b>0.920</b>
	300	0.898	0.799	0.839	0.877	0.970	<b>0.952</b>
0.10	50	1.223	0.770	1.183	0.628	1.298	<b>0.799</b>
	100	1.042	0.800	1.001	0.775	1.122	<b>0.887</b>
	200	0.903	0.799	0.841	0.883	0.979	<b>0.945</b>
	300	0.844	0.799	0.764	0.918	0.914	<b>0.967</b>

tribution such as t distributions and data containing extreme values. The detailed summary of our investigation is in Section 4.5.2. Extensive simulation studies suggest that BI is robust to both heavy tailed distributions and extreme values (Supplemental Figure 4.9-4.10).

#### 4.3.1.8 Comparison to Alternative Approaches

Although there are no existing methods specifically designed to identify bimodal genes in RNAseq data, it is still meaningful to compare the performance of BI with naive methods that treat the RNAseq data as similar to microarray data after some transformation (“log(data+1)”). To this end, we compare BI<sub>LN</sub> with PACK and COPA. (Full details are provided in Section 4.5.3.) When there are no outliers, Supplemental Figure 4.13 shows the performance of PACK is better than BI<sub>LN</sub> or COPA in most cases. However, PACK has difficulty detecting bimodal genes with 20%-80% or 30%-70% split, since the kurtosis values in these cases are near zero (Supplemental Figure 4.14). The reason PACK still achieves a good ROC curve is mostly attributable to the model selection step. When the data contains outliers, the performance of BI and COPA is more robust than PACK (Supplemental Figure 4.17-4.18). The reason is that model selection by BIC would flag most unimodal genes with outliers as bimodal candidates which makes it difficult for PACK to classify them correctly. In fact, BIC would claim that around 40% of the genes are bimodal candidates in the breast cancer data in the section. Supplemental Figure 4.15 shows that COPA fails to detect bimodal genes with 50%-50% or 10%-90% split at the chosen 10% quantile. We have to mention that COPA

Table 4.3: Performance on LN datasets (*Table reprinted from Tong, P. et al, Bioinformatics, 2013*)

FP	N	<b>BI<sub>NB</sub></b>		<b>BI<sub>GP</sub></b>		<b>BI<sub>LN</sub></b>	
		BI cutoff	Power	BI cutoff	Power	BI cutoff	Power
0.01	50	1.218	0.007	1.100	0.005	1.561	<b>0.410</b>
	100	0.976	0.052	0.818	0.008	1.304	<b>0.613</b>
	200	0.820	0.322	0.574	0.136	1.109	<b>0.790</b>
	300	0.763	0.482	0.459	0.520	1.034	<b>0.856</b>
0.05	50	1.014	0.086	0.840	0.049	1.367	<b>0.592</b>
	100	0.862	0.263	0.628	0.143	1.160	<b>0.750</b>
	200	0.737	0.568	0.409	0.734	1.005	<b>0.868</b>
	300	0.689	0.628	0.360	0.871	0.932	<b>0.914</b>
0.10	50	0.931	0.197	0.700	0.152	1.265	<b>0.688</b>
	100	0.797	0.436	0.511	0.430	1.094	<b>0.800</b>
	200	0.693	0.634	0.354	0.886	0.954	<b>0.899</b>
	300	0.653	0.650	0.308	<b>0.956</b>	0.878	0.938

has a tuning parameter, that is the quantile used to rank the genes. If this parameter changes, it is possible to identify a different set of bimodal genes (Supplemental Figure 4.16). However, the downside of using different quantiles is that it is difficult to obtain a consensus ranking of the genes as well as evaluate the false positive rate. Based on our simulation studies, we recommend the use of BI in practice for its ability of detecting a wide variety of bimodal genes, having no blind spots and being robust to outliers.

### 4.3.2 Real Data Analysis

We applied our methods to the TCGA Breast Cancer Dataset (BRCA) that contains 341 breast cancer samples for which both microarray and RNAseq data are available. The microarray data can serve as a reference to the RNAseq data in detecting bimodal genes.

#### 4.3.2.1 LN Model Fits Best For RNAseq Data

In order to examine which of the three models is most appropriate for real RNAseq data (and to identify the optimal  $\lambda$  in the Box-Cox power transformation), we need to identify reliable bimodal and unimodal genes in this dataset with high fidelity. For this purpose, we use the microarray data to guide our search. Since genes with null expression are usually beyond the detection limit of microarray

technology that may mislead the training set, we only looked at genes with mean expression  $> 1.5$  (Supplemental Figure 4.20). We then selected 142 candidate unimodal genes with  $BI < 0.5$ , which ensures that there is no apparent bimodality. For candidate bimodal genes, we used  $BI > 1.2$  as minimum requirement and found 181 candidates. All these genes passed manual examination. The complete curated gene list is provided in Supplemental Table 3.

Figure 4.3(a) shows an example gene where we used profile likelihood to identify the optimal transformation indexed by  $\lambda$ . Figure 4.3(b) shows the histogram of optimal  $\lambda$  for all genes. Figure 4.3(c) shows that the optimal  $\lambda$  for the candidate unimodal genes is concentrated at 0, suggesting that a log-transformation is optimal. Figure 4.3(d) shows that the LN model recovers almost all bimodal and unimodal genes in the curated dataset, while the performance of the NB and GP models is quite limited. This suggests that the LN model (with log transformation of the normalized counts) provides a better fit of real RNAseq data for the purpose of identifying bimodal genes.

#### 4.3.2.2 Bimodal Genes Identified Using RNAseq Data

Figure 4.4(a) shows the distribution of the mixture parameters ( $\pi$  and  $\delta$ ) in the BRCA RNAseq data after fitting the LN model. The red curve is the contour where  $FDR=0.01$  ( $BI=1.093$ ); genes identified as bimodal by the LN model are above this curve and circled in purple. We present the distributions of log-transformed count data for three genes known to be bimodally expressed in breast cancer (Figure 4.4(b)-(d)). Supplemental Table 1 shows the number of genes identified at different BI cutoffs, with FDRs obtained through simulation. A complete list of BI values for all genes is listed in Supplemental Table 2.

*(This section is an excerpt from Tong, P. et al, "SIBER: systematic identification of bimodally expressed genes using RNAseq data", Bioinformatics, 2013)*

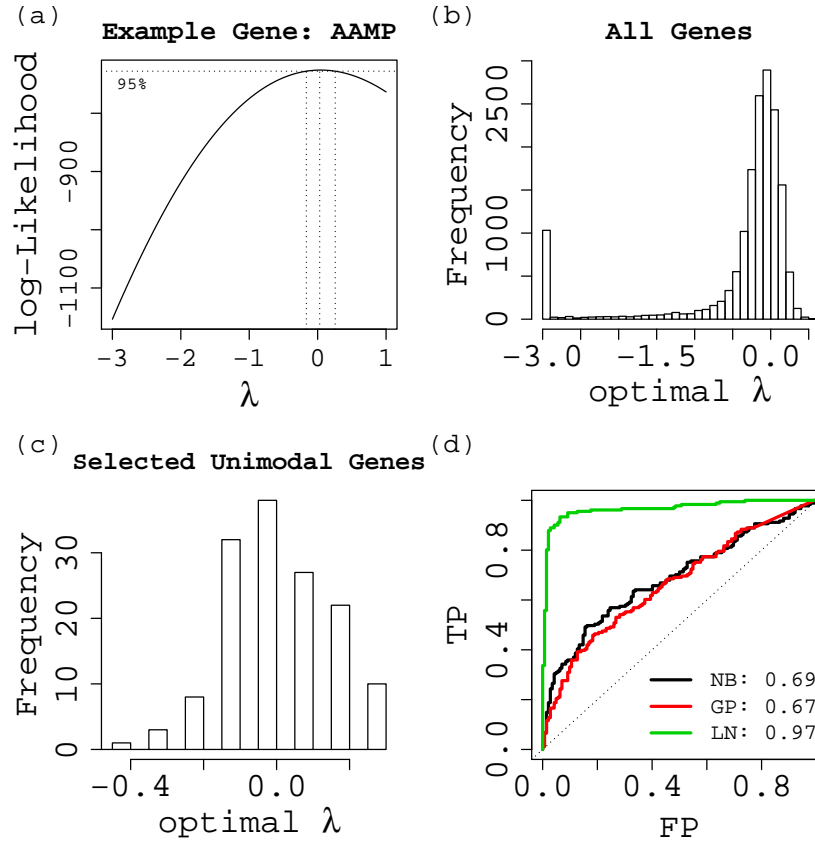


Figure 4.3: RNAseq data best fit by LN model. (a) An example showing log-transformation ( $\lambda = 0$ ) is identified as optimal by profile likelihood. Vertical dash lines indicate a 95% confidence interval for the optimal  $\lambda$ . (b) Histogram of optimal  $\lambda$  for all genes in RNAseq data.  $\lambda$  is concentrated at 0, suggesting log-transformation is optimal for the majority of genes. (c) Histogram of optimal  $\lambda$  values for the unimodal genes from curated dataset.  $\lambda$  values smaller than -3 are truncated at -3. Log-transformation is optimal for all these curated unimodal genes. (d) ROC curve for LN, NB and GP models fitted on RNAseq data for manually curated unimodal and bimodal genes. The performance of the LN model dominates that of the NB and GP models, suggesting the data is fitted best by the LN model. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

## 4.4 Discussion

During the last few decades, numerous studies have been published studying the transcriptome using microarray and, more recently, RNAseq technology with the hope of identifying biomarkers that can discriminate important phenotypes such as disease status, therapy response or even patient survival. As the paradigm for patient treatment shifts to personalized medicine, identifying clinically actionable biomarkers that are robust and have sufficient discriminatory power and dynamic range becomes an urgent task.



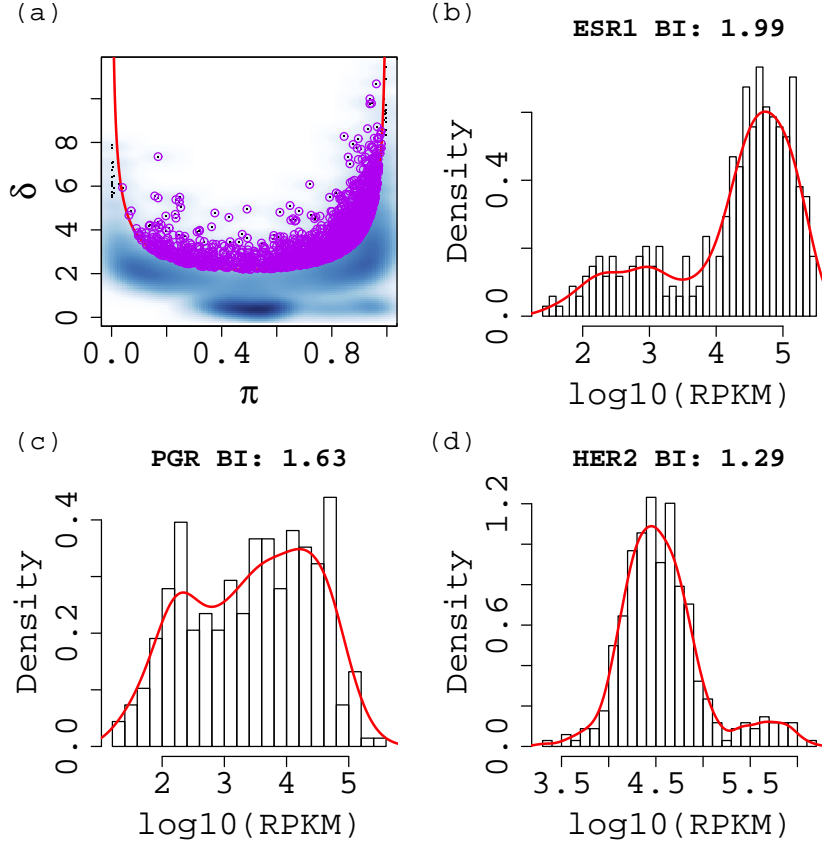


Figure 4.4: Example bimodal genes (a) Genes identified by the LN model in BRCA data. Genes with bimodal expression are circled in purple under FDR=0.01 (corresponding BI=1.093).  $\pi$  is the size of first component;  $\delta$  defines the distance between the two components as in (4.6). (b)-(d) Results of know breast cancer bimodal genes including ESR1, PGR, and HER2. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

In this chapter, we have proposed a novel method, **SIBER**, to systematically identify bimodal genes from RNAseq data. Our method is based on the BI approach [Wang et al., 2009] but extends the original definition of BI to mixtures of unequal variance and mixtures of discrete distributions. We consider different types of discrete distributions to model RNAseq and evaluate their performance under both true models and misspecified models. We also investigate their performance with real data.

Following the same line as BI, **SIBER** preserves nice interpretation. Strong bimodal genes identified by **SIBER** either have balanced component size  $\pi$  around 0.5 or a large distance between the two modes as defined by  $\delta$ . Although the exact BI formula cannot be derived for mixture of negative binomial and generalized

Poisson distributions, we have obtained an approximate formula that works well in both simulation and real data analysis. Our simulation shows there is only minimum power loss in this approximation. The idea for BI is quite general and hence it can be applied in different settings. For example, it is straightforward to derive the BI formula for a mixture of t-distributions or other distributions. It is also possible to derive a non-parametric version of BI as done in [Abdullatif Alwatban and Zheng Rong Yang \[2012\]](#). In all cases, the resulting BI formula is invariant under shifting and scaling which is a quite appealing feature for real data analysis since we should not change our idea of bimodality simply because the data is processed differently.

The LN model turns out to be the most effective model in both simulation study and real data analysis. Although all three models perform reasonably well under the true model, only the LN model performs well under misspecified models. In terms of recovering the bimodal status from the curated training set based on microarray data, the LN model performs better compared to the NB and GP models. We further show that the optimal transformation for curated unimodal genes is indeed a log transformation. This partially explains the superior performance of the LN model. However, an intrinsic reason might be that the nature of true expression levels is better described by a lognormal distribution, no matter whether it is measured by microarray (where the intensity value is modeled by lognormal distribution) or RNAseq.

Although the lognormal model performs best for identifying bimodal genes, it remains an open question whether the lognormal model also performs well for differential expression analysis in RNAseq data. In fact, the nature of the two tasks is quite different. While identifying bimodally expressed genes is unsupervised, differential expression analysis requires knowing the treatment condition and hence is supervised. Current comparison studies for RNAseq differential expression analysis only discuss the discrete models such as generalized

Poisson, negative binomial and two-stage Poisson [Kvam et al., 2012, Bullard et al., 2010]. Further study is needed to assess the performance of lognormal model in differential expression analysis.

Both the EM algorithm and Markov-chain Monte Carlo (MCMC) can be used to estimate the parameters in a mixture model. It has been shown that the two methods provide similar inference for BI [Wang et al., 2009]. Considering the computational efficiency issue, **SIBER** implements the EM algorithm. To further boost the computation speed, the **SIBER** package also provides parallel computing capability.

The BI based approach has been shown to be the most effective method for identifying bimodal genes. Once the bimodal genes have been identified, many follow up studies can be performed. For example, we can examine what GO categories or pathways are enriched in the bimodal gene list. Since many bimodal genes share similar patterns, it is biologically appealing to decompose the set of bimodal genes into a smaller number of binary signals that are distinct. A more important task would be to examine the predictive power of bimodal genes and see how well they perform for predicting clinical outcome. We defer this important task to the next chapter.

## 4.5 Appendix

*(Excerpts in this Appendix section are from the supplemental materials published online from: Tong, P. et al (2013). “SIBER: systematic identification of bimodally expressed genes using RNAseq data.”, Bioinformatics.)*

### 4.5.1 Derivation of the Generalized Bimodality Index

We first generalize the BI formula to a normal mixture with unequal variance. Then we define BI for mixture of arbitrary distributions.

#### 4.5.1.1 Normal Mixture with Unequal Variance

Suppose we have  $n_1 = N\pi$  and  $n_2 = N(1 - \pi)$  samples from normal distributions  $N(\mu_1, \sigma_1^2)$  and  $N(\mu_2, \sigma_2^2)$ , respectively.

Let the sample means be  $\bar{x}_1$  and  $\bar{x}_2$ . The null hypothesis is that the two components have equal mean when the expression is unimodal. Hence,

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

We have

$$\bar{x}_1 - \mu_1 \sim N(0, \frac{\sigma_1^2}{n_1}),$$

$$\bar{x}_2 - \mu_2 \sim N(0, \frac{\sigma_2^2}{n_2}).$$

Define the test statistic to be:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under  $H_0, T \sim N(0, 1)$

To achieve type I error  $\alpha$ , we set the critical value for rejecting  $H_0$  as  $c$ . Then,

$$P_{H_0}(|T| > c) = \alpha$$

where  $z_{\alpha/2}$  is the quantile of the standard normal distribution such that the upper tail probability is  $\frac{\alpha}{2}$ . Hence,  $c = z_{\frac{\alpha}{2}}$

Similarly, under  $H_1$  where the gene is bimodal, a proper  $c$  would be chosen to control type II error as  $\beta$ :

$$P_{H_1}(|T| < c) = \beta$$

Since

$$P_{H_1}(|T| < c) = P_{H_1}\left(\left|\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2) + (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| < c\right)$$

where  $\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$  under  $H_1$ , we have:

$$c - \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z_\beta.$$

$$\text{Plug in } c = z_{\frac{\alpha}{2}}, (z_{\frac{\alpha}{2}} + z_\beta)^2 = \frac{(\mu_1 - \mu_2)^2}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

which means

$$\frac{(z_{\frac{\alpha}{2}} + z_\beta)^2}{N} = \frac{\pi(1-\pi)(\mu_1 - \mu_2)^2}{(1-\pi)\sigma_1^2 + \pi\sigma_2^2} \triangleq BI^2.$$

Therefore,

$$BI = \sqrt{\pi(1-\pi)} \frac{|\mu_1 - \mu_2|}{\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}} = \sqrt{\pi(1-\pi)}\delta \quad (4.7)$$

where  $\delta = \frac{|\mu_1 - \mu_2|}{\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}}$  measures the distance between the two modes.

(An excerpt from Tong, P. et al, “SIBER: systematic identification of bimodally expressed genes using RNAseq data”, *Bioinformatics*, 2013)

#### 4.5.1.2 Negative Binomial Mixture

We assume the data is generated from a 2-component Negative Binomial (NB) Mixture with means  $\mu_1, \mu_2$  and variances  $\sigma_1^2, \sigma_2^2$ . We assume  $\sigma_1^2, \sigma_2^2$  are known parameters. Similar to the normal mixture case, we formulate our hypothesis as:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

The most powerful test would be the Wald test which requires the MLE of  $\mu_1, \mu_2$ , denoted as  $\hat{\mu}_1, \hat{\mu}_2$ :

$$\frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\sqrt{\text{var}(\widehat{\mu}_1) + \text{var}(\widehat{\mu}_2)}}$$

When  $\sigma_1^2, \sigma_2^2$  are known,  $\widehat{\mu}_1, \widehat{\mu}_2$  are not just the sample means. In fact, they are complicated functions of the data and  $\sigma_1^2, \sigma_2^2$ . Instead, we can use alternative test based on Central Limit Theory (CLT) as below:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under  $H_0, T \sim N(0, 1)$  approximately. This is the same as the normal mixture case, except the distribution of T is an approximate.

With similar arguments as Section 4.5.1.1, this approximation leads to exactly the same BI formula as the normal mixture:

$$BI = \sqrt{\pi(1 - \pi)} \frac{|\mu_1 - \mu_2|}{\sqrt{(1 - \pi)\sigma_1^2 + \pi\sigma_2^2}}$$

The above formula relies on CLT and hence relies on asymptotic normality. Motivated by this, we can first transform (i.e. Box-Cox transformation) the data and use normal mixtures to identify bimodality. The Box-Cox transformation is defined as below [Box and Cox, 1964]:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

where  $y_i$  is the original data and  $y_i^\lambda$  is the transformed data. Real data shows taking logarithm (as indicated by  $\lambda = 0$ ) is optimal, which means the original data is fitted by a mixture of lognormal (LN) distribution, which we define as the LN model in the manuscript. (An excerpt from Tong, P. et al, “SIBER: systematic identification of bimodally expressed genes using RNAseq data”, *Bioinformatics*, 2013)

#### 4.5.1.3 Generalized Poisson Mixture

The data can be fitted with a mixture of Generalized Poisson Distributions. Similar procedures would give us the approximate BI as:

$$BI = \sqrt{\pi(1-\pi)} \frac{|\mu_1 - \mu_2|}{\sqrt{(1-\pi)\sigma_1^2 + \pi\sigma_2^2}} = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1^2}{\pi} + \frac{\sigma_2^2}{1-\pi}}}$$

#### 4.5.1.4 Zero-Inflation

RNAseq has a high resolution to measure null expression where exact count of zero can be found. This is a special case in both our mixture modeling as well as BI formula. In terms of fitting mixture models, one component would be a point mass at zero while the other component might be any of NB, GP, LN or other models. Both NB and GP models can automatically fit such a zero-inflated model by fitting  $\mu_1 = 0$  since a point mass at zero is a special case of both NB and GP distribution. However, for the LN model the point mass at zero can not be automatically fitted as the EM algorithm would not return  $\mu_1 = 0$ ,  $\sigma_1 = 0$ . To deal with this situation, we empirically detect genes with 0-inflation (i.e. >20% of the samples have count zero) and fit a univariate LN model on the nonzero counts.

For the BI formula in the zero-inflated case, it degenerates as follows by setting  $\mu_1 = 0$  and  $\sigma_1 = 0$  in formula 4.6 (*An excerpt from Tong, P. et al, "SIBER: systematic identification of bimodally expressed genes using RNAseq data", Bioinformatics, 2013*):

$$BI = \sqrt{1-\pi} \frac{\mu_2}{\sigma_2}$$

#### 4.5.2 Investigation of Outlier Data

The comparison of NB, GP and LN models examines the robustness to model misspecification. Now we examine the robustness to outlier data points. Since previous comparison shows LN model performs best in terms of power and robustness under various scenarios, we just focus our investigation on LN model (or equivalently the normal model after log-transformation). Real microarray data is usually exemplified by heavy tailed distributions and extreme expression. This motivates us to use a t-distribution to simulate data with heavy tail. The severity of the heavy tail can be controlled by the degree of freedom (df). Both unimodal and bimodal genes can be generated under similar settings as in Section 4.3.1.1 (therefore, we borrow the notation introduced before). Note that for a t-distribution, the standard deviation ( $\sigma$ ) is determined by  $\sigma = \sqrt{\frac{df}{df-2}}$  and mean  $\mu_1 = 0$ . To generate a mixture of t-distribution with different bimodality, we can shift a t-distribution with  $\mu_2 = \sigma * \delta$  to represent the second component

where  $\delta = 2.5, 3, 3.5, 4$  quantifies the distance between the two components. The proportion of the first component  $\pi_1$  is also set to be between 0.1 and 0.5 with a step length of 0.1. To simulate data with extreme expression, we simulate the ‘normal part’ with  $\text{Normal}(5, 1)$  and add extreme values from a uniform distribution  $U[3\sigma, 4\sigma] + \mu_1$ . The number of extreme values is chosen to be 1, 2 or 4. Both unimodal and bimodal genes with extreme values are simulated. The presence of both a heavy tailed distribution and extreme values tends to generate more small BI estimates in both unimodal and bimodal genes (see Supplement Figure 4.11, 4.12). This effect is stronger on the simulated unimodal genes. The overall performance of BI is quite robust to both heavy tailed distributions and extreme values (4.9, 4.10). (*An excerpt from Tong, P. et al, SIBER: systematic identification of bimodally expressed genes using RNAseq data, Bioinformatics, 2013*)

### 4.5.3 Comparison with COPA and PACK

Both COPA [Tomlins et al., 2005] and PACK [Teschendorff et al., 2006] are designed for microarray data. Since there is no existing method dealing with RNAseq data for bimodal gene identification, we compare our method with a naive approach that simply transforms the RNAseq data and treats it as microarray data. In particular, we compare  $BI_{LN}$  with COPA and PACK after log-transformation. Note that the log-transformation is shown to be optimal for RNAseq data and frequently used in other studies.

We implement PACK with the `vabayelMix` package. In particular the `unbiasedKurt()` function is used to compute the Kurtosis of log-transformed data. We perform PACK analysis without clustering so that all genes can be assigned with a rank. In theory, samples from a normal distribution would have Kurtosis of 0. In contrast, bimodal genes arising from a mixture of normal distributions would have nonzero Kurtosis (either positive or negative). We use the absolute value of Kurtosis to rank the genes and hence construct an ROC curve.

COPA applies a simple transformation of the data before ranking the genes. Suppose the log-transformed values (or microarray measurements) is a vector  $\mathbf{y} = (y_1, y_2, \dots, y_N)$  where  $y_i$  is the gene expression in the  $i$ th sample for a given gene. Let the median of  $\mathbf{y}$  be  $m$  and its median absolute deviation (MAD) be  $d$ . COPA first applies the following transformation for each  $y_i$ :

$$y'_i = \frac{y_i - m}{d}$$



Then, COPA uses the quantile of the transformed values  $\mathbf{y}' = (y'_1, y'_2, \dots, y'_N)$  to rank the genes. Originally, the 75%, 90% and 95% quantiles are used which lead to three lists of rankings for identifying genes with outlier *over-expression*. In our setting, the “outlier” group is the first component since  $\pi_1 \leq 0.5$ . Because we assume  $\mu_1 \leq \mu_2$  which means the “outlier” group has *under-expression*, it is equivalent to choose 25%, 10% and 5% quantiles to rank the genes with outlier *under-expression*. In our representation, we choose the 10% quantile as the COPA score which is always negative. Hence, smaller COPA score means stronger bimodality. This COPA score is then used to rank the genes and to construct an ROC curve. (*An excerpt from Tong, P. et al, “SIBER: systematic identification of bimodally expressed genes using RNAseq data”, Bioinformatics, 2013*)

#### 4.5.4 Supplemental Figures

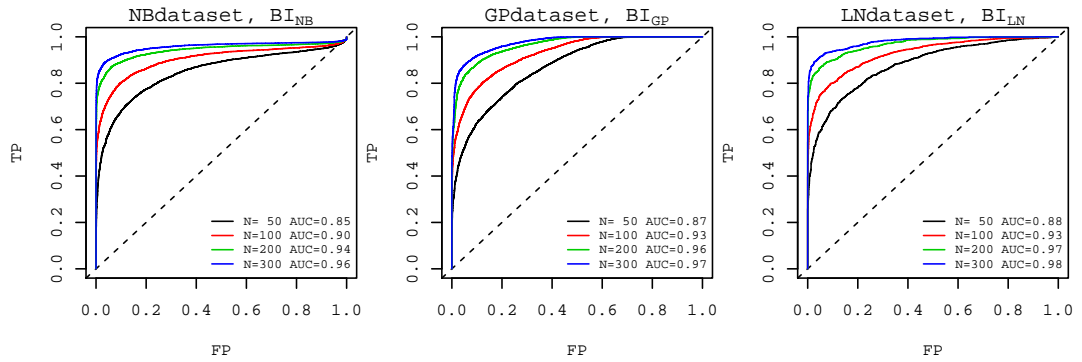


Figure 4.5: Performance under the correctly specified (“true”) model. ROC curves fitted by the true model are compared for each of NB, GP and LN datasets under sample size  $N=50$ , 100, 200 and 300. Various bimodal shapes as characterized by different distances ( $\delta = 2.5, 3, 3.5, 4$ ) between the two components and component size ( $\pi = 0.1, 0.2, 0.3, 0.4, 0.5$ ) are simulated to mimic real data. Under the true model, the three methods have similar performance. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

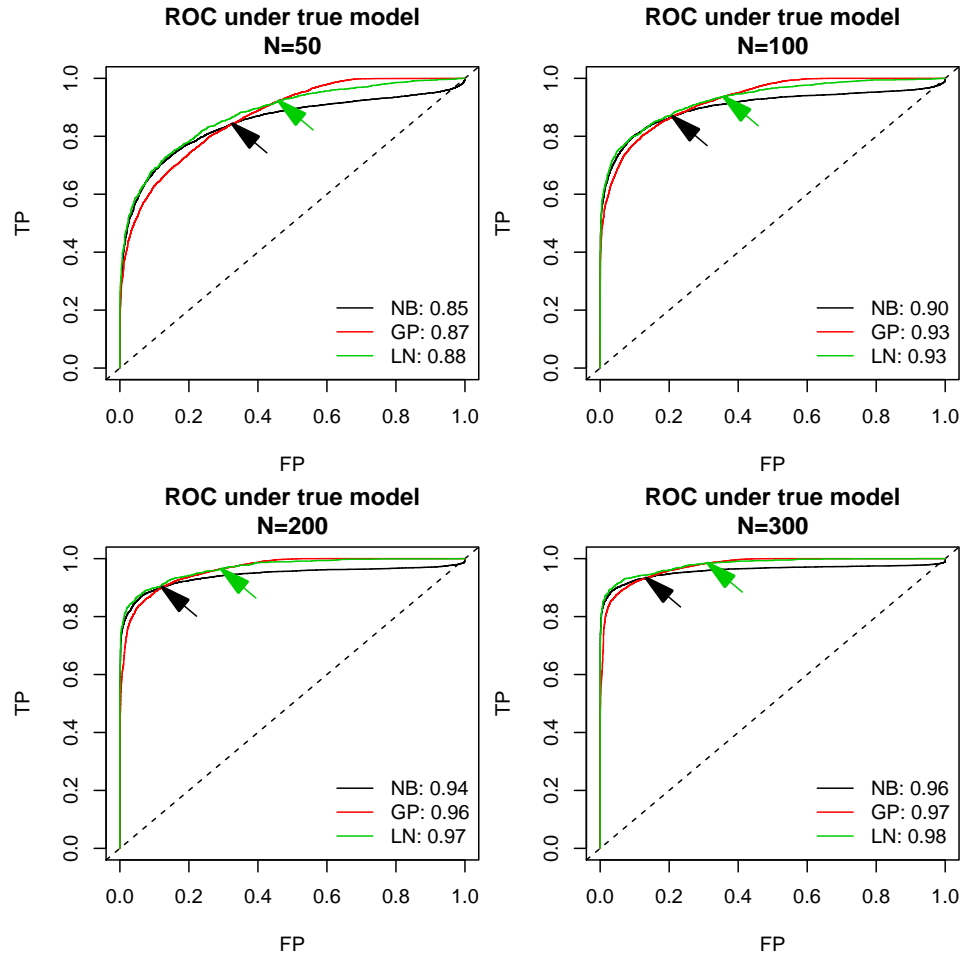


Figure 4.6: ROC curves under true models. ROC curves under correct model specification for the three models. Black arrows indicate the point where ROC curves of GP models intersect NB models; Green arrows indicate the intersection points of GP models and LN models. FP rates at the intersection points between GP and NB models are 0.326, 0.210, 0.122 and 0.133 under  $N=50, 100, 200$  and  $300$ , respectively; similarly, FP rates at the intersection points between GP and LN models are 0.460, 0.358, 0.294 and 0.321 under  $N=50, 100, 200$  and  $300$ , respectively. From a practical point of view where FP rate is not allowed to be large, the larger power at low FP rate seen with the LN and NB models is preferred. Compared to the LN and NB models, the TP rate of the GP model increases faster. However, the overall performance of the three models in terms of AUC is quite similar. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

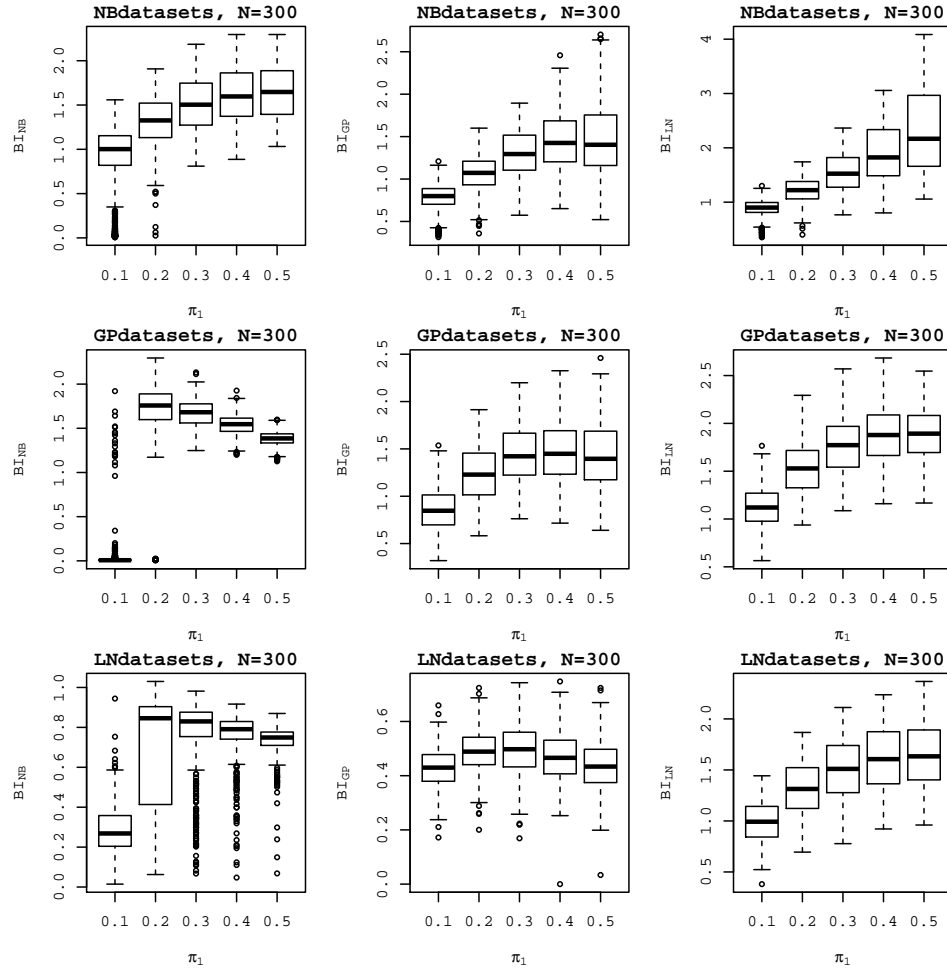


Figure 4.7: Boxplot of fitted BI by the NB, GP and LN models stratified by different  $\pi_1$  for simulated bimodal genes (denoted as  $H_1$  in the y-axis). Boxplot of estimated  $BI_{NB}$ ,  $BI_{GP}$  and  $BI_{LN}$  from simulated bimodal genes stratified by  $\pi_1$  from each dataset at  $N=300$ . First row represents NB dataset; second row for GP dataset and third row for LN dataset.  $BI_{NB}$  fails to detect bimodal genes with  $\pi_1=0.1$  under misspecified model. Estimated BI increases with  $\pi_1$  except  $BI_{NB}$  in LN and GP datasets and  $BI_{GP}$  in LN dataset due to model misspecification. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

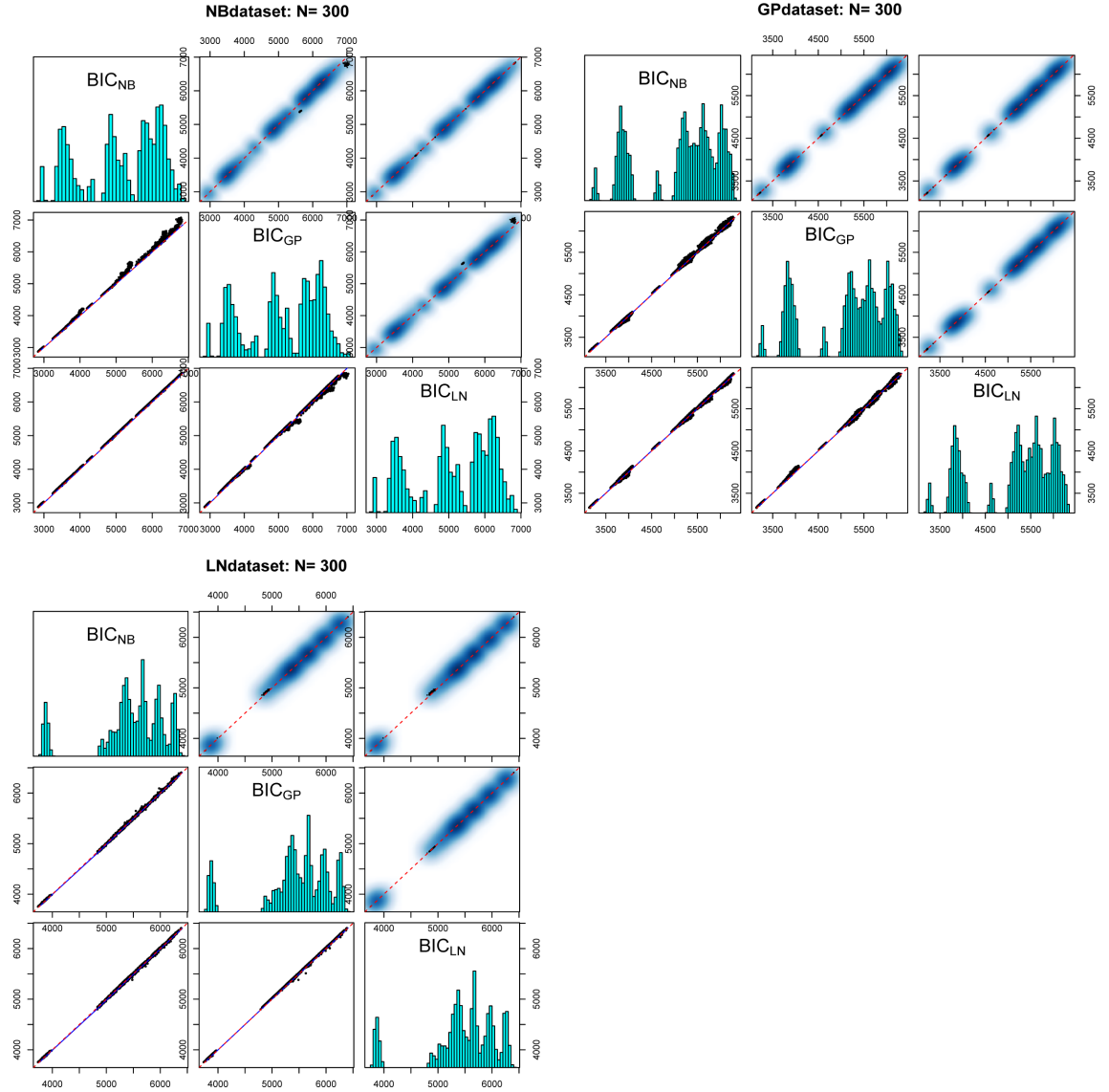


Figure 4.8: Cases where it is impossible to identify the data generating model (true model) by BIC. Comparison of BIC from fitting three mixture models on NB, GP and LN datasets. Sample size  $N=300$  is chosen for illustration purposes. Upper panel: smoothed density plot with higher density indicated by thicker cloud; Lower panel: scatter plot; diagonal panel: histogram of the BIC values. For all three datasets, BICs from the three models are almost identical, suggesting they provide similar fits to the data, despite their varying performance for bimodality identification. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

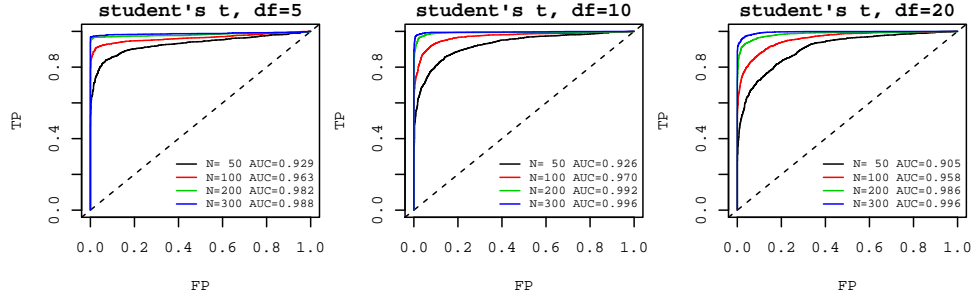


Figure 4.9: Robustness to heavy tailed distributions. Unimodal and bimodal genes are simulated from t or a mixture of t distribution with different degrees of freedom (df) to mimic the effect of heavy tailed distribution frequently seen from real data. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

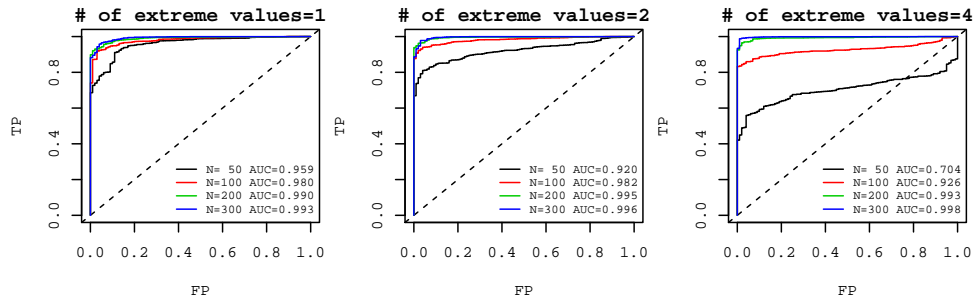


Figure 4.10: Robustness to extreme values. Different numbers of extreme values are simulated such that the robustness to them can be examined. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

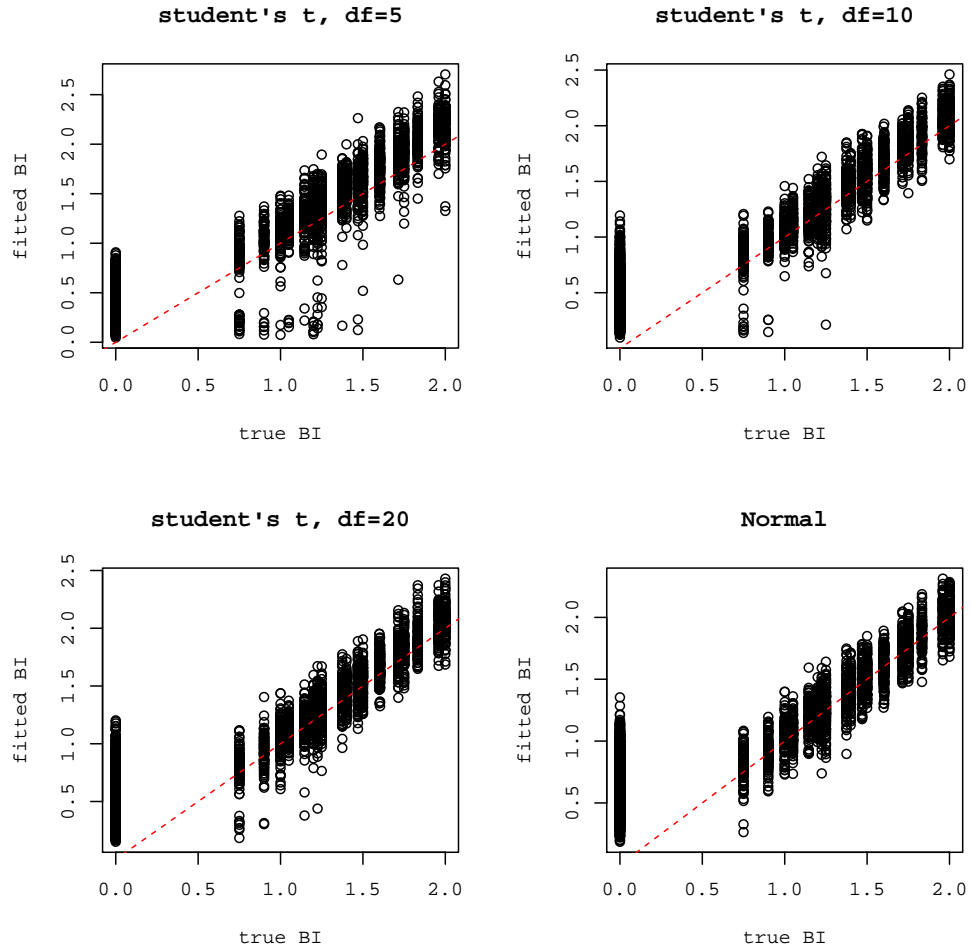


Figure 4.11: Effect of heavy tailed distributions on the estimation of BI. The estimated BI is plotted against the true BI for simulated unimodal (true BI=0) and bimodal genes (BI > 0) from t-distribution with different degrees of freedom (sample size N=200 is used for illustration). The presence of heavy tails tends to generate more smaller BI values in both unimodal and bimodal genes compared to the normal distribution. Overall, the shift in unimodal genes is more obvious. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

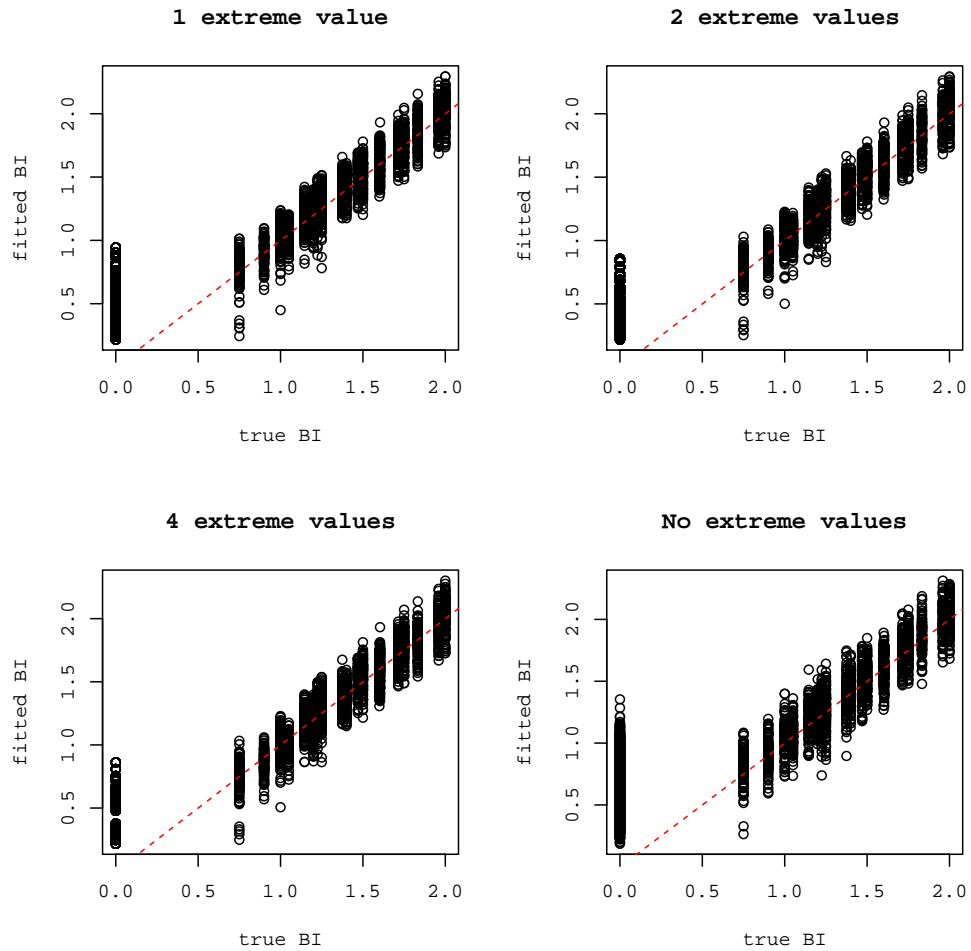


Figure 4.12: Effect of extreme values on the estimation of BI. The estimated BI is plotted against the true BI for simulated unimodal (true BI=0) and bimodal genes (BI > 0) containing different number of extreme values (N=200). The presence of extreme values generates more smaller BI estimates in both unimodal and bimodal genes compared to the normal distribution. Overall, the extreme values seems to affect the estimated BI more in unimodal genes. (*Figure reprinted from Tong, P. et al, Bioinformatics, 2013*)



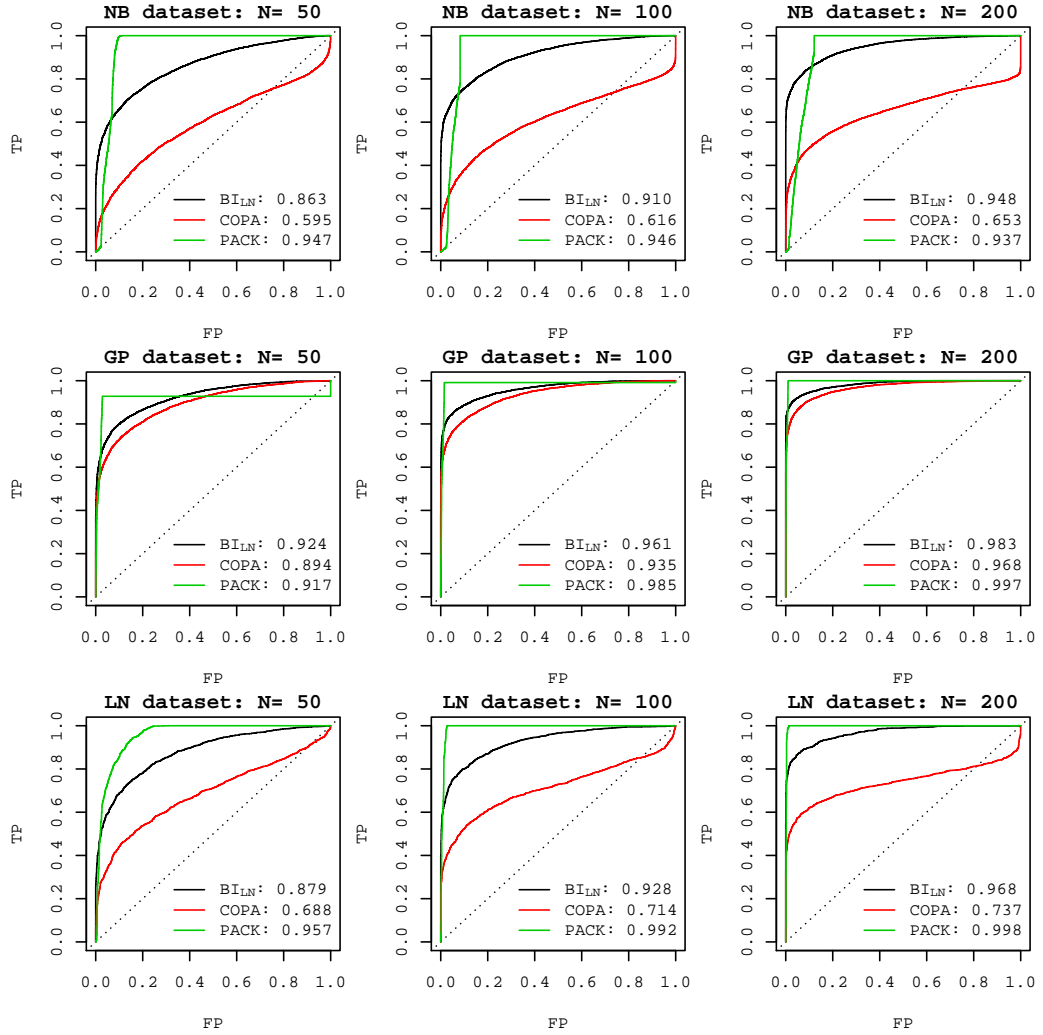


Figure 4.13: ROC curves for  $BI_{LN}$ , PACK and COPA. The performance of  $BI_{LN}$ , PACK and COPA are compared under sample size  $N=50$ , 100 and 200. AUC is shown at bottom-right of each panel. Since PACK and COPA works on normally distributed data, the count data is first log-transformed before applying PACK and COPA. For PACK, the absolute value of Kurtosis is used to rank the genes after model selection with BIC. For COPA, the 10% quantile of the transformed values (centered by median and scaled by MAD) is used for ranking. Lower COPA score leads to stronger bimodal expression. When there are no outliers, PACK performs best followed by BI and then COPA. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

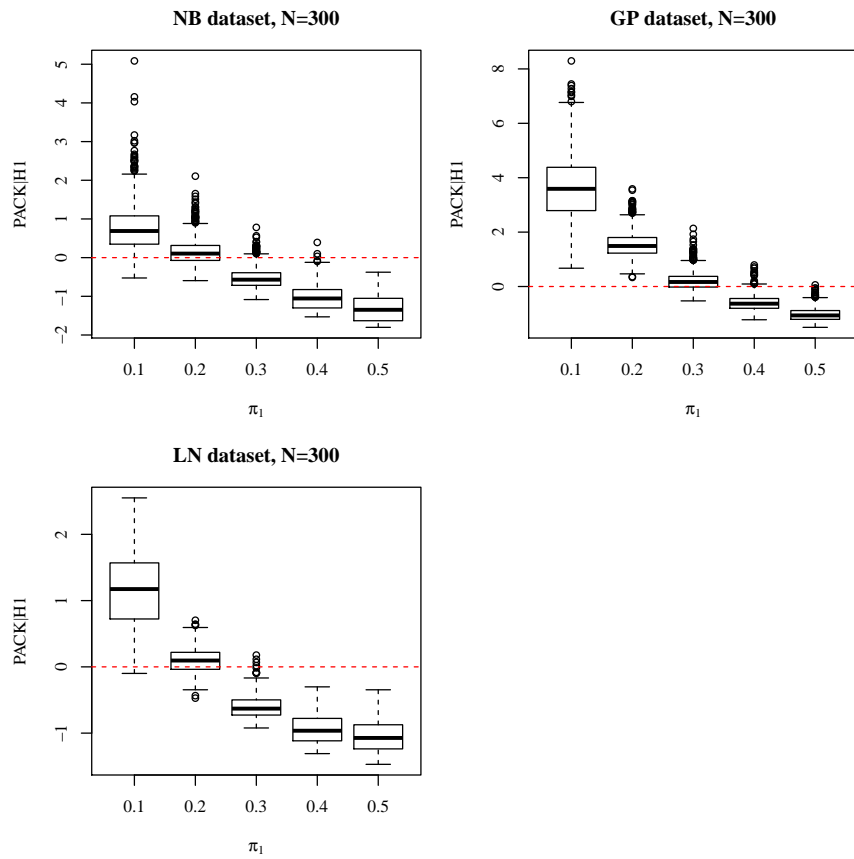


Figure 4.14: PACK finds it difficult to detect bimodal genes with a 20%-80% or 30%-70% split. Boxplot of kurtosis for simulated bimodal genes stratified by  $\pi_1$  in each dataset with sample size  $N=300$ . For unimodal genes, kurtosis is theoretically 0 which is indicated by the horizontal red line. However, most kurtosis at  $\pi_1=0.2$  in LN and NB dataset as well as  $\pi_1=0.3$  in GP dataset is centered at 0. (*Figure reprinted from Tong, P. et al, Bioinformatics, 2013*)

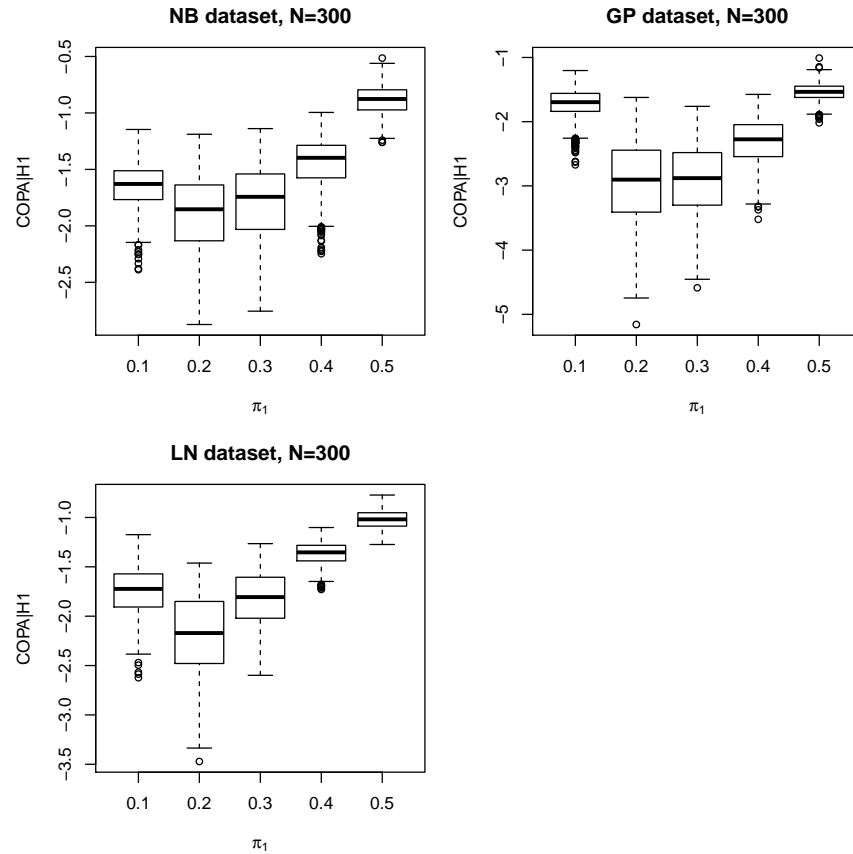


Figure 4.15: COPA does not work well when bimodal expression is 50%-50% or 10%-90% split at the chosen quantile for ranking. Boxplot of COPA scores (10% quantile of transformed value) is shown for simulated bimodal genes stratified by  $\pi_1$  in each dataset with sample size  $N=300$ . Lower COPA scores indicate stronger bimodal expression. At  $\pi_1=0.5$  in the LN and NB datasets, COPA score is largest which overlaps with COPA score from unimodal genes. Therefore, COPA fails to detect these genes. Similarly, at  $\pi_1=0.1$  and  $\pi_1=0.5$  in the GP dataset, COPA assigns a large score to the simulated bimodal genes which makes COPA fail to detect these genes. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

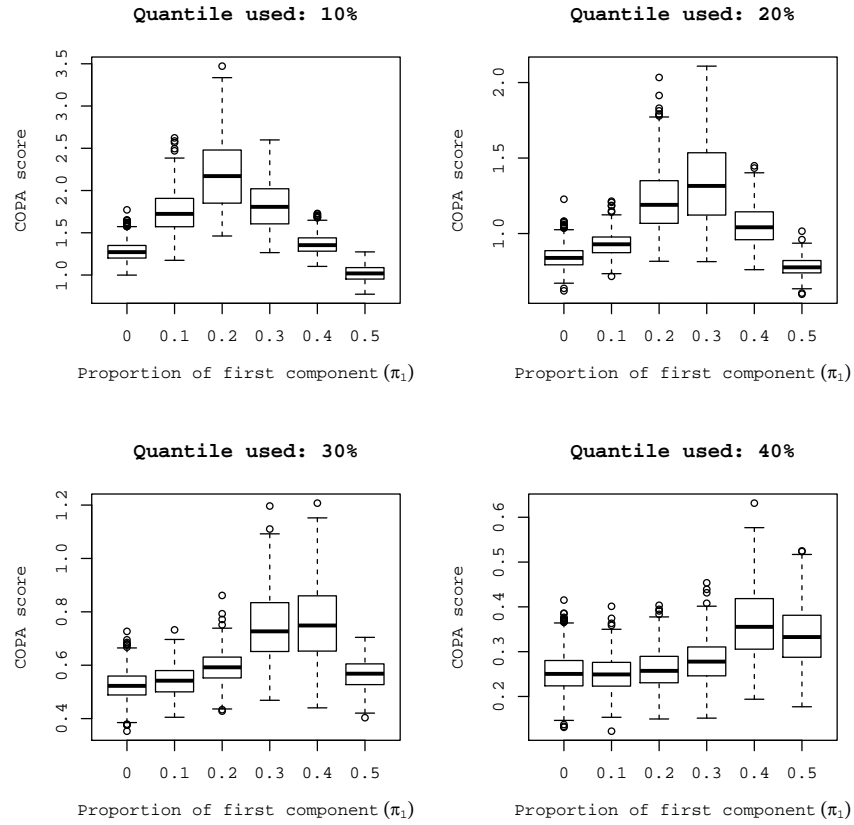


Figure 4.16: COPA detects different set of bimodal genes as the quantile used for ranking changes. Boxplot of COPA score under different choice of quantiles for simulated bimodal genes with sample size  $N=300$  is shown. Lower COPA score indicates stronger bimodal expression. COPA is good at detecting bimodal genes with  $\pi_1 = 0.1 \sim 0.3$  when 10% quantile is used, with  $\pi_1 = 0.2 \sim 0.4$  when 20% or 30% quantile is used and with  $\pi_1 = 0.4 \sim 0.5$  when 40% quantile is used, respectively. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

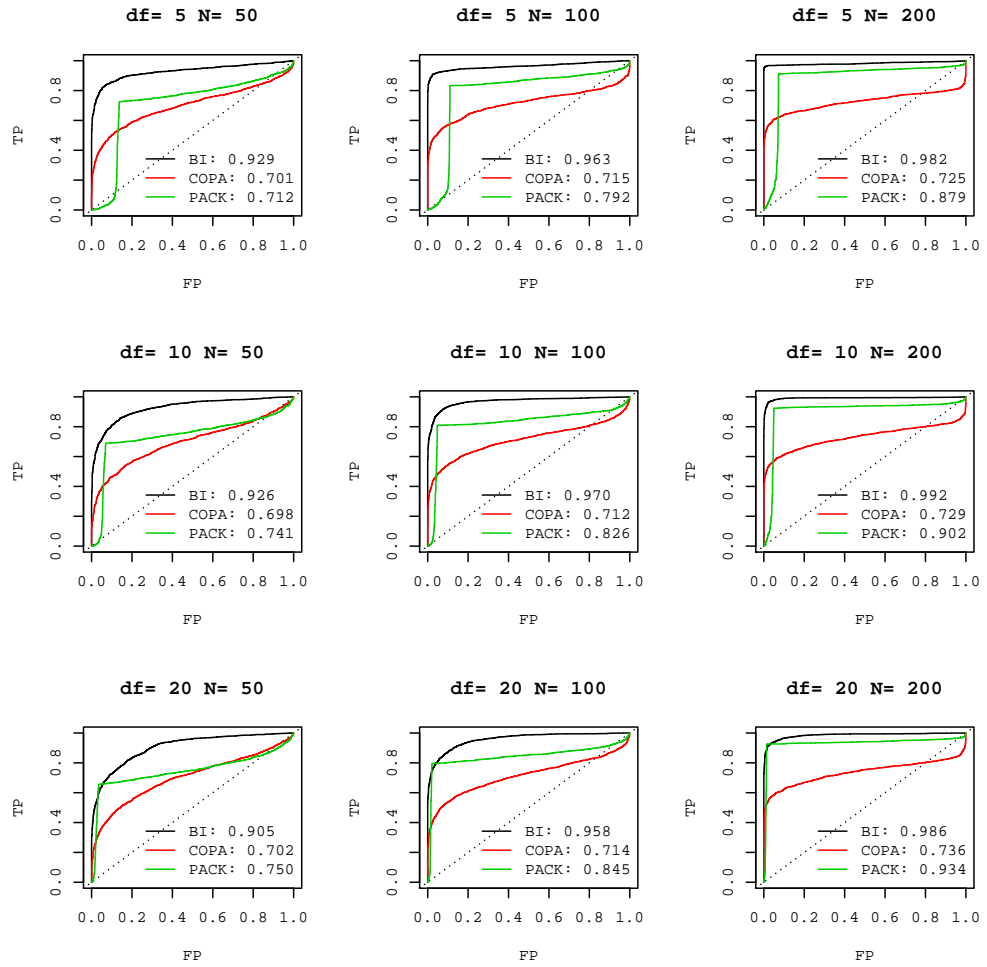


Figure 4.17: Robustness to heavy tailed distributions. Unimodal and bimodal genes are simulated from student's  $t$  or a mixture of student's  $t$  distribution with different degrees of freedom (df) to mimic the effect of a heavy tailed distribution. Both BI and COPA are robust to data with heavy tails while PACK is not. As df increases, the performance of PACK also improves. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

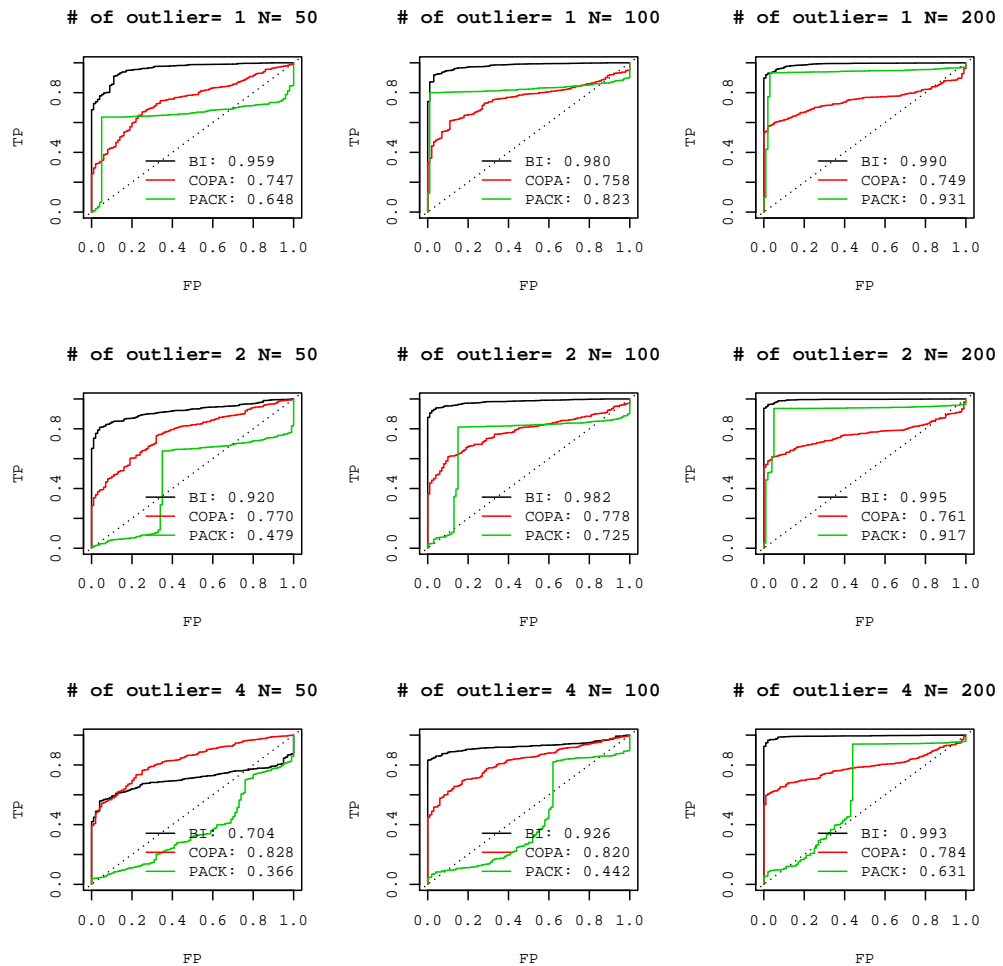


Figure 4.18: Comparison of robustness to extreme values. Different numbers of extreme values are simulated such that the robustness to them can be examined. Both BI and COPA are robust to extreme values while PACK is not, mostly due to model selection step that flag the unimodal genes as bimodal. The AUCs at  $N=50$  with 4 outliers in all three methods are not good. This is because 8% of the samples are simulated to be extreme values which in reality is not likely to happen. It is included here to demonstrate the trend. (*Figure reprinted from Tong, P. et al, Bioinformatics, 2013*)

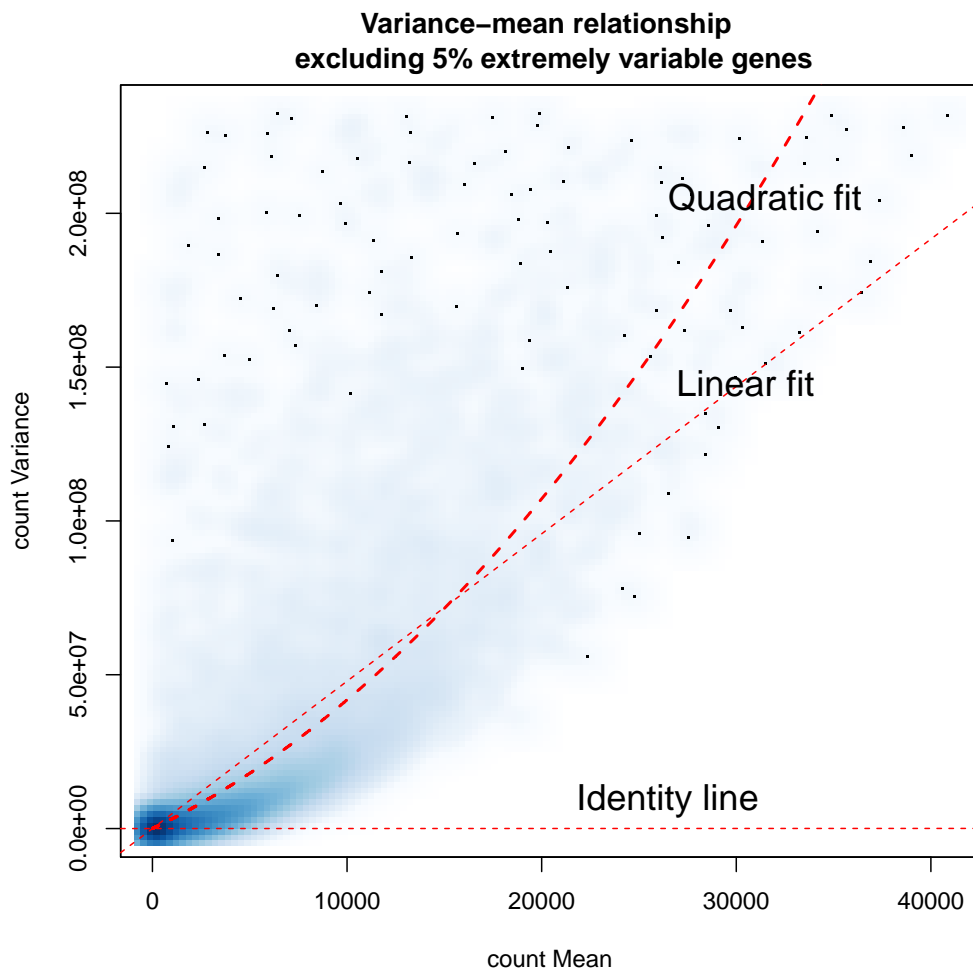


Figure 4.19: Mean-variance relationship in the BRCA RNAseq data. ANOVA shows the p value of choosing the quadratic fit over linear fit is less than  $2.2 \times 10^{-16}$ . (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)

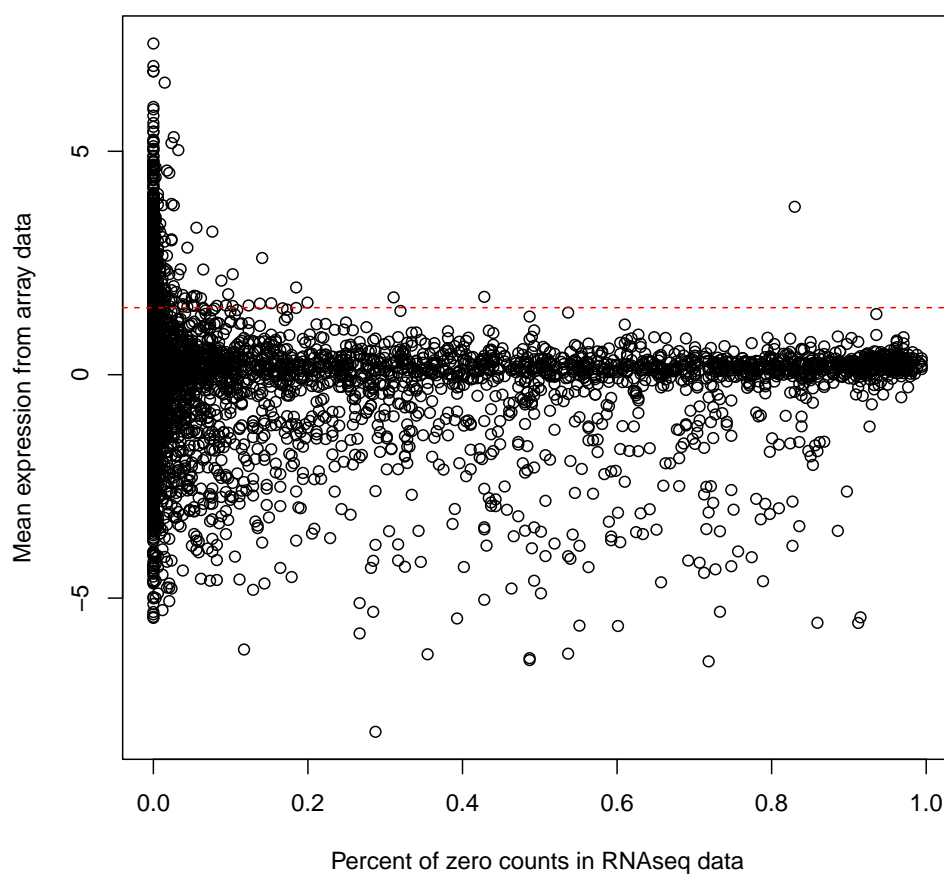


Figure 4.20: Curated genes with known bimodal status are not null expressed. Red dashed line indicates mean expression  $> 1.5$ . Curated bimodal and unimodal genes have mean expression level over 1.5 from microarray data. Most of these have percent of zero counts less than 5% which confirms they are expressed. We check expression with percent of zero counts rather than median count because median count is confounded with gene length effect. (Figure reprinted from Tong, P. et al, *Bioinformatics*, 2013)



#### 4.5.5 Supplemental Tables

Table 1: Genes identified in BRCA data at different BI threshold  
(Table reprinted from Tong, P. et al, *Bioinformatics*, 2013)

BI Cutoff	Number of Bimodal Genes Identified	FDR
1.0	2589	0.066
1.1	1805	0.008
1.2	1206	0.001
1.3	805	$\approx 0$
1.4	554	$\approx 0$
1.5	368	$\approx 0$
1.6	231	$\approx 0$
1.7	140	$\approx 0$
1.8	95	$\approx 0$

Table 2: **SIBER** analysis for all genes in BRCA data. See the online supplementary file [SupplementaryTable2.csv](#). The fitted parameters in mixture model as well as BI are listed.  $\log_{10}(\text{data}+1)$  is applied before fitting the mixture model.

Table 3: Curated unimodal and bimodal gene list. See the online supplementary file [SupplementaryTable3.csv](#). This table contains the gene names as well bimodality status for 181 bimodal and 142 unimodal genes.

## Chapter 5

# Bimodal genes contain most information for predicting outcome

### 5.1 Background

In Chapter 4, we developed the SIBER approach to systematically identify bimodal genes for RNAseq data. These bimodal genes can be good candidates for therapeutic targets. It remains a question how these bimodal genes predict clinical outcome. This chapter is devoted to investigating the predictive power of bimodal genes. We will discuss the utility of bimodal genes in the context of cancer classification. Further, after establishing the predictive power of bimodal genes, we propose an integrative approach for cancer classification based on discrete signals extracted from multiple sources.

In the literature, it is reported that gene expression signatures derived from whole transcriptome profiling using microarray and next generation sequencing have the potential to impact patient care including accurate diagnosis and prognosis and ultimately realize the promise of personalized medicine. Over the years, many computational methods have been proposed to classify patients into

different risk groups or even predict survival with varying degrees of success [Golub et al., 1999, Dudoit et al., 2002]. Existing approaches for class prediction involve feature selection followed by model building and validation. The feature selection step is needed to filter out irrelevant genes for better accuracy and avoidance of over-fitting. The selected features are then used to train a predictive model usually in the form of a classifier or regression model. The performance is usually evaluated on a validation set or through cross-validation when independent validation data is not available.

In addition to prediction accuracy, it is realized that a good signature should help illuminate the underlying biology and more importantly, be clinically actionable. The development of a clinically actionable biomarker involves multiple phases [Pletcher and Pignone, 2011]. At the early phase, statistical association should be established with the clinical outcome of interest. The risk of disease needs to be accounted for to ensure that a large enough patient population exists. Studies to a later stage investigate whether the novel biomarker would alter the practice of physicians prescribing treatment decisions. Most methods to date primarily focus on the association of a biomarker with the outcome while ignoring the incidence rate of disease. In contrast, biomarkers based on bimodal expression implicitly deal with incidence rate and hence are good candidates for clinically actionable biomarkers [Wang et al., 2009].

Genes with bimodal expression exhibit strong contrast of expression difference between two patient groups [Ertel and Tozeren, 2008, Teschendorff et al., 2006]. It has been shown that genes with prognostic power are enriched in bimodal genes where the expression among the samples forms two distinct clusters representing low and high expression [Hellwig et al., 2010]. Although the technique to identify bimodal genes has become a mature technology [Wang et al., 2009, Tong et al., 2013], there is no published work evaluating the predictive power of bimodal genes.

Here we propose to predict clinical outcome with binary signals arising from bimodal expression. Classifiers built by this approach have a potential to be more clinically useful and easy to operate. Through extensive evaluation using multiple public benchmark datasets, we show that prediction models built from bimodal genes have the same accuracy as models built with all genes. The remainder of this chapter is organized as follows: In Section 5.2 we describe the three benchmark data sets as well as the pipeline to evaluate the performance of bimodal genes in predicting clinical outcome. The approach for bimodal gene identification will be briefly reviewed. In Section 5.3 we present the performance of models built from bimodal genes, unimodal genes and all genes and conclude the effectiveness of outcome prediction using bimodal genes.

## 5.2 Methods

### 5.2.1 Datasets

We have assembled three benchmark datasets from public repositories. These datasets have been frequently used to evaluate the performance of different prediction models.

**MAQC-II data** We download the MAQC-II data [Shi et al., 2010] from Gene Expression Omnibus (GEO) with series accession number GSE16716. MAQC-II includes six datasets corresponding to thirteen binary endpoints coded as A through M (Table 5.1). This dataset serves as a benchmark for predicting binary outcome. The first three datasets are related to toxicogenomics in rodents. The last three datasets are related to prognosis in human cancer including breast cancer, multiple myeloma and neuroblastoma. Four of the thirteen endpoints are artificially designed to serve as positive and negative controls: endpoints I and M are negative controls where the class label is randomly assigned and hence

impossible to predict; endpoints H and L are positive controls representing sex of the patients that is highly predictable from gene expression data. The binary survival status (F, G, J, K) is compiled in the MAQC-II study based on a threshold for the survival time. The MAQC-II dataset encompasses diverse outcomes with different characteristics which makes it a good benchmark dataset for evaluating classification methods.

**Tan et al data** For multi-class outcome data, we use the same datasets as in [Tan et al. \[2005\]](#). There are 10 datasets in total all related to human cancer including leukemia, lung, colorectal, prostate, breast, central nervous system, lymphoma, bladder, melanoma, renal, uterus, pancreas, ovary and mesothelioma. A summary of the Tan et al data is provided in Table 5.2. The number of classes in each dataset ranges from 3 to 14.

**NCI Director’s Challenge Lung Cancer data** The NCI Director’s Challenge Consortium released comprehensive lung cancer data with blinded training/testing split collected from multiple sites mimicking the true patient population [[Shedden et al., 2008](#)]. The primary goal for the Director’s Challenge project was to evaluate whether microarray measurements of gene expression could be used to predict overall survival for lung cancer patients. We download the released data and use the original assignment of training/testing status to evaluate the predictive performance for different classifiers. To construct receiver operating characteristic (ROC) curves for each classifier, [Shedden et al. \[2008\]](#) recommend dichotomizing the survival into binary classes based on 3-year survival. We summarize the assembled data in Table 5.3. There are two test sets for the NCI Director’s Challenge Lung data. The Test1 data is broadly similar to the training set in the initial evaluation of gene expression data [[Shedden et al., 2008](#)]. Test2 data has reduced signal intensity and is more challenging.

Date set name	Endpoint code	Endpoint description	Microarray platform	Sample size (training)	Sample size (validation)
Hamner	A	Lung tumorigen vs. non-tumorigen (mouse)	Affymetrix	70	88
Iconix	B	Liver carcinogens vs. non-carcinogens (rat)	GE CodeLink	216	201
NIEHS	C	Liver toxicants vs. non-toxicants (rat)	Affymetrix	214	204
Breast cancer (BR)	D	Pathologic complete response status	Affymetrix	130	100
	E	Estrogen receptor status		130	100
Multiple myeloma (MM)	F	Overall survival milestone outcome	Affymetrix	340	214
	G	Event-free survival milestone outcome		340	214
	H	Positive control		340	214
	I	Negative control		340	214
Neuro-blastoma (NB)	J	Overall survival milestone outcome	Agilent	238	177
	K	Event-free survival milestone outcome		239	193
	L	Positive control		246	231
	M	Negative control		246	253

Table 5.1: Summary of MAQC-II dataset

Dataset Name	Platform	Number of classes	Sample size (training)	Sample size (validation)	Reference
Leukemia1	Affymetrix	3	38	34	(Golub et al., 1999)
Leukemia2	Affymetrix	3	57	15	(Armstrong et al., 2002)
Lung1	Affymetrix	3	64	32	(Beer et al., 2002)
SRBCT	cDNA	4	63	20	(Khan et al., 2001)
Breast	Affymetrix	5	54	30	(Perou et al., 2000)
Lung2	Affymetrix	5	136	67	(Bhattacharjee et al., 2001)
DLBCL	cDNA	6	58	30	(Alizadeh et al., 2000)
Leukemia3	Affymetrix	7	215	112	(Yeoh et al., 2002)
Cancers	Affymetrix	11	100	74	(Su et al., 2001)
GCM	Affymetrix	14	144	46	(Ramaswamy et al., 2001)

Table 5.2: Summary of Tan et al dataset

	Training	Test1	Test2
Short Survival	96	23	25
Long Survival	152	68	52

Table 5.3: Summary of Director's Challenge Lung Cancer data

### 5.2.2 Identifying Bimodally Expressed Genes

Several methods have been proposed to identify genes with bimodal expression for microarray data. An extensive comparison of these methods can be found in [Hellwig et al. \[2010\]](#). We choose the Bimodality Index (BI) to identify bimodal genes here for two reasons: (1) BI is shown to perform well for a wide range of bimodal shapes; (2) under the BI framework, samples can be easily split into two groups representing low and high expression.

Below we briefly review the BI method. For a given gene, let the expression value in sample  $i$  be  $y_i$  for  $i=1, 2, \dots, N$  samples. The expression values are modelled through a two component mixture model:

$$f(y_i) = \pi f(y_i; \mu_1, \sigma) + (1 - \pi) f(y_i; \mu_2, \sigma)$$

where  $f(y_i; \mu_1, \sigma)$  is the density function for a normal distribution with mean  $\mu_1$  and standard deviation  $\sigma$  and  $\pi$  is the proportion of samples in the first component. After estimating the parameters  $(\pi, \mu_1, \mu_2, \sigma)$  using the Expectation-Maximization method, BI can be calculated as:

$$BI = \sqrt{(\pi(1 - \pi))} \frac{|\mu_1 - \mu_2|}{\sigma}$$

The above model is fit for each gene such that bimodal genes can be identified by setting a threshold of BI that is related to the sample size  $N$ . Based on the computed BI, we can build three sets of genes for later comparison: bimodal genes, unimodal genes, and all genes. Gene expression array data usually contains noise that might jeopardize classification accuracy. For this reason, we apply a filtering step before any computation. In particular, we compute the variance

of each gene and remove genes that have low variability. We filter out 50% of the genes based on their variances. Microarray data might also suffer from batch effects. An extensive study about batch effects in MAQC-II data can be found in [Luo et al. \[2010\]](#). We examined possible batch effect related to run date and use mean-centering for batch effect correction when possible.

Given a bimodal gene, it is natural to draw a cutoff and dichotomize the expression values. Doing so will make it easier to interpret the result as well as greatly facilitate the development of diagnostic/prognostic devices. To dichotomize the expression values of bimodal genes, we choose a cutoff such that the probability of belonging to either component is 0.5. This leads to a cutoff computed as the mean of two component means estimated from the mixture model. Expression values smaller than this cutoff will be coded as 0 while larger values will be coded as 1. There are also other methods to dichotomize bimodal genes. For example, the posterior probability of belonging to either component can also be used. This metric takes into account the component size but gives similar results as using the mean of component means.

### 5.2.3 Classification Methods

We use PAM (Prediction Analysis of Microarrays) [[Tibshirani et al., 2002](#)] for classification since it is widely used and performs well in extensive comparison studies [[Lee et al., 2005](#), [Wessels et al., 2005](#)]. Another good feature of PAM classifier is that PAM uses soft-thresholding for feature selection. Therefore, there is no limit on the number of features to be explored comparing alternative classifiers such as Support Vector Machine and K Nearest Neighbours. We didn't include other classifiers because our primary focus is to evaluate whether bimodal genes are enough for microarray based classification compared to all genes. In this regard, we base our conclusion on the relative performance between using all



genes and using only bimodal genes. For each endpoint or dataset, PAM was first fit on the training set and then frozen to make predictions on the validation set. PAM implements cross-validation to select the number of features. Three-fold cross-validation is used for all datasets in the training stage.

We also evaluate how the bimodal genes perform when the data is dichotomized. The PAM classifier is designed for continuous features and hence inappropriate for binary features. We therefore evaluate alternative classifiers including Naive Bayes, Classification and Regression Tree (CART) and Bayesian Network (BN). CART is chosen since it can generate simple decision rules to make prediction. We also evaluate BN since it is able to model the dependency among bimodal genes and the clinical outcome. Naive Bayes is a simple yet effective classifier that treats the features as independent.

#### 5.2.4 Performance Evaluation Metrics

It is critical to use a valid metric to compare the performance of the all-gene and bimodal-gene models. Classification accuracy is one of the mostly widely used metrics. Although Matthews correlation coefficient (MCC) was used in the MAQCII study [Shi et al., 2010], it only applies to binary predictions. Usually the two metrics give similar results. Therefore, for the sake of simplicity, we only present results based on accuracy while MCC is omitted in this presentation. We compare the accuracy of validation set from classifiers built with all genes and bimodal genes. To specifically test which of the two classifiers is better, we applied the Net Reclassification Index (NRI) and Integrated Discrimination Index (IDI) method [Pencina et al., 2008]. The result from NRI and IDI test may be controversial in the sense that the classifier with higher accuracy can be declared to be significantly worse than the other classifier with lower accuracy (data not shown). Further, both NRI and IDI are criticized for ignoring the

sampling variability [Pepe et al., 2007]. These considerations lead us to declare a winning classifier when it has higher accuracy. Note that this criterion is also used in Tan et al. [2005].

## 5.3 Results

### 5.3.1 MAQC-II Binary-Class Data

Table 5.4 shows the accuracy on test set for models using bimodal, unimodal and all genes. The largest accuracy among the three models for each endpoint is bolded to aid visual inspection. The endpoints H and L are positive controls and hence the accuracy is high for all three models. In contrast, the endpoints I and M are negative controls and thus accuracy around 0.5 is expected. After excluding the negative controls, the bimodal gene model achieves 7 largest accuracies out of 11 endpoints while the all gene model achieves 6 largest accuracies. The unimodal gene model only performs best on two of the 11 endpoints. This result shows that for binary classification, the bimodal-gene model performs similarly to all-gene model and better than the unimodal-gene model in terms of classification accuracy.

### 5.3.2 Tan et al Multi-Class Data

Table 5.5 shows the performance of the three models on multi-class classification task on Tan et al data. Among the 10 data sets, the bimodal-gene model performs best in 6 of them. The performance for the all-gene model is similar (7 best out of 10). The unimodal-gene model performs much worse. Note that the performance of all three models on the GCM data is poor. The reason is that the outcome of GCM data contains 14 different categories. The sample size on the training set is only 144 which means around 10 samples are used to

	Bimodal	Unimodal	All
A	<b>0.716</b>	0.659	<b>0.716</b>
B	0.672	<b>0.721</b>	0.711
C	<b>0.917</b>	0.863	0.897
D	<b>0.710</b>	0.650	0.640
E	<b>0.890</b>	0.870	<b>0.890</b>
F	<b>0.696</b>	0.654	0.678
G	0.626	<b>0.650</b>	0.631
H	0.869	0.822	<b>0.874</b>
I	0.477	0.500	0.491
J	<b>0.848</b>	0.842	<b>0.848</b>
K	0.782	0.798	<b>0.808</b>
L	<b>0.987</b>	0.727	<b>0.987</b>
M	0.490	0.478	0.502

Table 5.4: Result of MAQC-II dataset

	Bimodal	Unimodal	All
Leukemia1	<b>0.940</b>	0.940	0.820
Leukemia2	<b>0.800</b>	0.670	<b>0.800</b>
Lung1	<b>0.780</b>	0.750	<b>0.780</b>
SRBCT	<b>1.000</b>	0.950	<b>1.000</b>
Breast	0.970	0.900	<b>1.000</b>
Lung2	0.960	<b>0.970</b>	<b>0.970</b>
DLBCL	<b>0.930</b>	0.900	0.900
Leukemia3	<b>0.910</b>	0.780	0.900
Cancers	0.860	0.840	<b>0.880</b>
GCM	0.410	<b>0.570</b>	<b>0.570</b>

Table 5.5: Result of Tan et al dataset

characterize each of the 14 categories. Although the accuracy is less than 0.6, it is much better than random guess (the expected accuracy would be 0.07). over all the datasets, the bimodal-gene model performs almost the same as the all-gene model.

### 5.3.3 Director's Challenge Lung Cancer Data

We have shown examples suggesting that bimodal genes contain enough information for both binary and categorical classification tasks. We now go one step further to evaluate how the bimodal genes perform when the features are

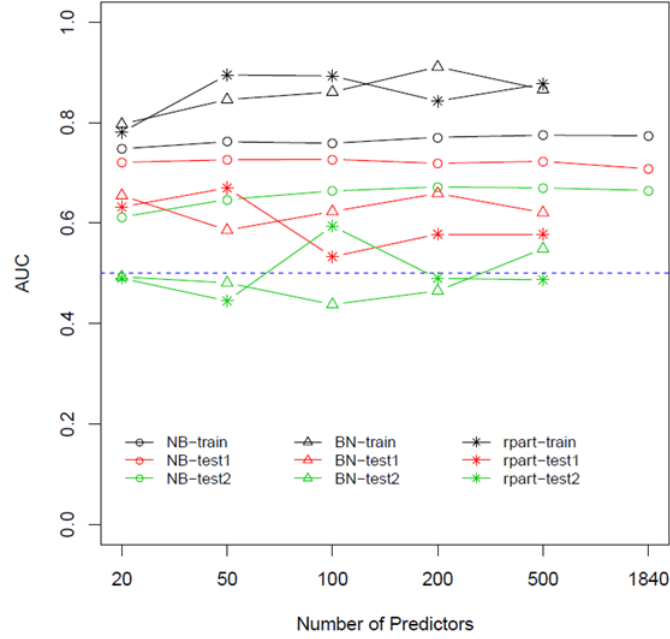


Figure 5.1: Comparison of naive bayes (NB), Bayesian network (BN) and CART (rpart) on Director's challenge lung cancer data. As the number of top predictive genes changes, the AUC of the three classification methods also varies. The performances on both training (labeled as train in black) and test set (labeled as test1 in red or labeled as test2 in green) are illustrated.

dichotomized. The reason to dichotomize the data is because binary inputs make it easier to interpret. Also, if classification based on binary features is still satisfactory, this means we do not need exact measurements for the biomarkers. Only the grouping information is needed for prediction. This will greatly facilitate the development of diagnostic/prognostic device.

Figure 5.1 shows the performance of NB, CART and BN classifiers on the Director's Challenge lung cancer data. The performance seems to be rather stable when the number of features changes. Both BN and CART achieve larger AUC on the training set than NB classifier does. However, NB performs much better than BN and CART on both test sets. Overall, BN and CART tend to be overfitting. This suggests that when bimodal genes are dichotomized, NB is a good choice for building classifiers.

We specifically investigate why BN performs so poorly on the test set. Figure 5.2(a) shows the network constructed by the BN classifier with genes

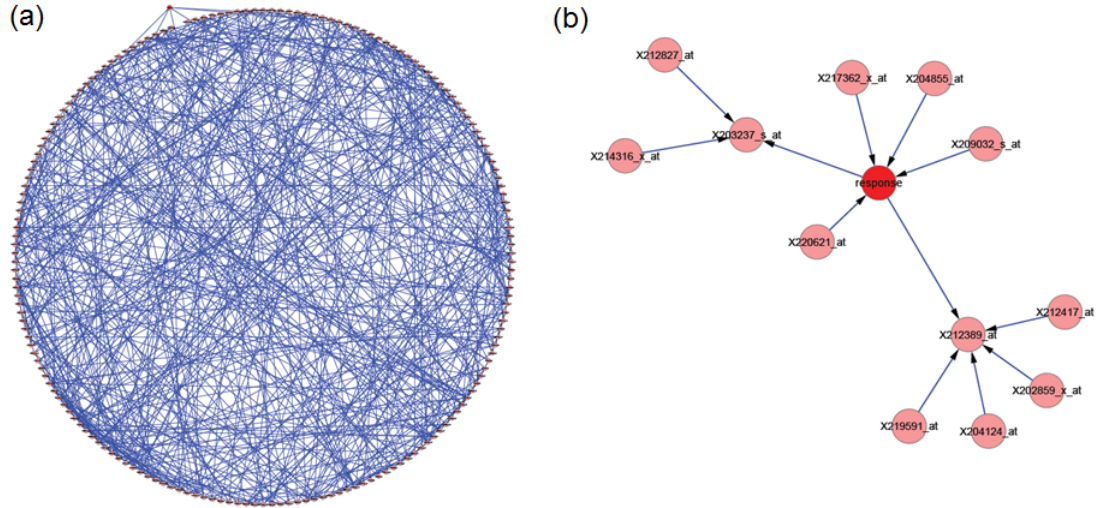


Figure 5.2: The dependency network learnt by Bayesian network on Director’s challenge lung data. (a) The network with 181 genes and clinical outcome. The red node on top left indicates the outcome node. (b) The Markov blanket for the outcome node. This is the network that is needed to predict outcome.

significantly associated with outcome ( $p$  value  $< 0.05$ ). The red point on the top left indicates the outcome variable. We see that there are many edges between the nodes. Since there are only 248 samples used in training, most of these edges might be dubious. In fact, not all nodes and edges learnt in BN are used for making predictions. Only nodes serving as the Markov blanket (shown in Figure 5.2(b)) of the outcome variable are needed to make predictions. This further explains why BN tends to overfit microarray data.

## 5.4 Discussion

Genes with bimodal expression play very important roles in various biological processes especially in carcinogenesis. The two modes of expression are a reflection of switch-like regulation [Ertel and Tozeren, 2008]. Multiple studies have suggested that bimodal genes can separate patients with different survival [Tomlins et al., 2005] or cancer subtypes [Teschendorff et al., 2006]. The analysis done by Hellwig et al. 2010 evaluates the performance of different methods for identifying bimodal genes and examine how the identified genes correlate to

clinical outcome. The Bimodality Index approach [Wang et al., 2009] turns out to outperform other methods by identifying more predictive genes. It remains a question whether bimodal genes alone are enough for building predictive models.

In this chapter we formally evaluate the performance of classifiers built from bimodal genes, unimodal genes and all genes using several benchmark datasets. These benchmark datasets cover a wide variety of endpoints, species and array platforms. We choose the PAM classifier for all models to eliminate the bias introduced by different classification methods. The built-in feature selection in PAM further removes artifacts from fine tuning parameters during the training process. We also check potential batch effects for each dataset to ensure data quality as recommended in Luo et al. 2010. Practically, genes with low variation are deemed as noise and filtered out before building classifiers. We also adopt this practice to ensure a fair comparison.

Through extensive evaluation, we confirm that bimodal genes contain the same information as all genes in predicting various binary and categorical outcome. In the MAQC-II data, classifiers built from bimodal genes perform best in 7 out of the 11 outcomes. Classifiers built from all genes perform best in 6 outcomes. There are 4 tied accuracies between the bimodal-gene model and the all-gene model. In terms of classification accuracy, the bimodal-gene model is slightly better than the all-gene model and both are much better than the unimodal-gene model. The result on Tan et al data is quite similar. Classifiers built from bimodal genes perform best in 6 of the 10 data sets while the classifiers built from all genes perform best in 7 of the data sets with 3 data sets having tied accuracies. The accuracies in our analysis is similar to those reported publicly.

After establishing the predictive power of bimodal genes, we further evaluate how these genes perform when data is dichotomized. For the dichotomized

data, PAM becomes inappropriate. We therefore choose three representative classifiers including NB, CART and BN. The NB classifier achieves the best performance among the three classifiers, and it is comparable to those reported in [Shedden et al. \[2008\]](#). Both CART and BN tend to overfit. We find that BN infers too many edges even when the data is limited which explains why BN tends to be overfit.

The effectiveness of building prediction models solely with bimodal genes has great implications. Our analysis has established the predictive power of bimodal genes. This means we can extract natural binary signals such as bimodal expression for prediction. Further, where there are multiple assays, it is straightforward to integrate the data when they are discretized. In terms of classification, we can build classifiers with discretized data such as categorized copy number change, methylation change and mutation data.

Our comparison is based on empirical criteria. A rigorous statistical test for asserting the performance difference would be attractive. However, there is no effective test developed yet. Both the NRI and IDI tests [[Pencina et al., 2008](#)] do not work well in our context.

Our current analysis does not use gene modules that explicitly incorporate dependency among the genes. Due to the strong contrast of expression in the bimodal genes, inferring regulatory networks from bimodal genes and building classifiers with networks would be quite interesting.

## Chapter 6

### Conclusions and future research

#### 6.1 Conclusions

Throughout this thesis, we focus on developing statistical approaches for integrating multiple high throughput assays. The key question we try to address is how to extract biological insights and formulate testable hypothesis based on the combined information. We have developed a variety of methodologies for integrative analysis that cover both supervised and unsupervised learning. We devote Chapters 2 and 3 to integrative biomarker identification and Chapters 4 and 5 to integrative classification.

Chapter 2 introduces a regression based approach to identify biomarkers by integrating multiple assays including gene expression, methylation and copy number data. This method allows us to evaluate the predictive power of each individual assay as well as the combined data. We implement penalized regression so that correlated measurements can be dealt with. To specifically identify a subset of the measurements that is most predictive, we adopt a stepwise model selection procedure. An application to the TCGA ovarian cancer data shows that



gene expression, methylation, and copy number have different power to predict either therapy response status or overall survival. Interestingly, genes predictive of therapy response also differ from genes predictive to overall survival despite significant overlap. We also find that the prognostic genes identified through our integrated analysis rarely overlap with known cancer genes characterized by mutation.

Chapter 3 shifts gears to identify biomarkers based on gene alteration. In particular, we have developed a latent trait model for identifying altered genes accounting for different mechanisms. This model automatically adjusts for the heterogeneity among different assay types and samples such that the latent traits for different genes are placed on a common scale. Compared to conventional methods, our method is able to identify altered genes that are more reliable and biologically meaningful. Further, our method can identify novel altered genes that cannot be found by looking at individual assay separately.

Chapters 4 and 5 approach the data integration problem in the classification setting. Both chapters resemble Chapter 3 in the sense that they perform data integration with discrete signals. Chapter 4 proposes a novel method to extract binary signals from RNAseq expression data. We present the Bimodality Index (BI) approach which generalizes a previous method developed for microarray data. The proposed method compares favorably with other methods in both simulation and real data analysis. Chapter 5 evaluates the predictive power of bimodal genes. Through extensive analysis on several benchmark datasets, we find that bimodal genes contain the same amount of information as all genes for predicting various endpoints. Further, even after converted into binary, bimodal genes still provide accurate classification. For these binary features, it is found that the Naive Bayes classifier performs better than several other candidate classifiers in terms of both the accuracy and ROC curve.

## 6.2 Future research

Data integration has become an ongoing challenge in biomedical research. The explosion of high throughput profiling technologies has enabled cheaper and faster data generation. For example, both array-based and sequencing-based platforms have been used to obtain profiles of whole genome expression, methylation and copy number. Over the years, the biomedical community has collected various sources of information stored in diverse repositories. How to link and integrate the collected data is a big challenge. We have developed several methods for integrative analysis. Still, there are many topics that need further research.

An immediate project expanding our research would be classification with multiple data sources. It is expected that by integrating information from different sources, classification performance would be greatly enhanced. Our analysis has shown the predictive power of binary signals. It is straightforward to build classifiers with discretized copy number, methylation and mutation data. In our evaluation, we choose to investigate the PAM, Naive Bayes, Bayesian Network and CART classifiers. In terms of integrative classification, it is likely that different data sources might prefer different classifiers. To unleash the power of integrated analysis, we can apply the boosting algorithm such that the weights of different data sources can be learnt [Schapire, 2002].

Our research on data integration has explored both biomarker identification and classification. What we have not touched is data integration using gene networks. Network-based biomarker identification and classification is a natural generalization of our work. In terms of integrative analysis, a network-based approach is quite attractive. We can first identify functional modules from different data sources. These functional modules, jointly modeled through a dependency network, will illuminate the inherent structure within the data. Classification based on network modules would further help us understand how pathways as a

whole affect phenotype.

## References

- Abdullatif Al watban and Zheng Rong Yang. Bimodal Gene Prediction Via Gap Maximisation. In *The 2012 International Conference on Bioinformatics and Computational Biology*, July 16-19 2012.
- Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.
- Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe’er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, 2010.
- Simon Anders, Alejandro Reyes, and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.
- Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-Seq data. *Genome Research*, 2012.
- Erling B Andersen. *Discrete statistical models with social science applications*. North-Holland Publishing Company, 1980.
- David Andrich. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573, 1978.

- Masato Aoki, Masahiko Kanamori, Kazuo Ohmori, Mikiro Takaishi, Nam-ho Huh, Shigeharu Nogami, and Tomoatsu Kimura. Expression of developmentally regulated endothelial cell locus 1 was induced by tumor-derived factors including VEGF. *Biochemical and Biophysical Research Communications*, 333(3):990–995, 2005.
- B Aranda, P Achuthan, Y Alam-Faruque, I Armean, A Bridge, C Derow, M Feuermann, AT Ghanbarian, S Kerrien, J Khadake, et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(suppl 1):D525–D531, 2010.
- Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25, 2000.
- Gary D Bader, Doron Betel, and Christopher WV Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Research*, 31(1):248–250, 2003.
- Frank B Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*, volume 176. CRC, 2004.
- Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, et al. NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.
- Dennis A Benson, Mark S Boguski, David J Lipman, and James Ostell. GenBank. *Nucleic Acids Research*, 25(1):1–6, 1997.
- Silvio Bicciato, Roberta Spinelli, Mattia Zampieri, Eleonora Mangano, Francesco Ferrari, Luca Beltrame, Ingrid Cifola, Clelia Peano, Aldo Solari, and Cristina

- Battaglia. A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Research*, 37(15):5057, 2009.
- Stephen R Biggar and Gerald R Crabtree. Cell signaling can direct either binary or graded transcriptional responses. *The EMBO Journal*, 20(12):3167, 2001.
- George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society Series B*, 26:211–252, 1964.
- Michelle D Brazas, Joseph T Yamada, and BF Francis Ouellette. Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory. *Nucleic Acids Research*, 38(suppl 2):W3–W6, 2010.
- James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.
- Lynne Cassimeris. The oncoprotein 18/stathmin family of microtubule destabilizers. *Current Opinion in Cell Biology*, 14(1):18–24, 2002.
- Arnaud Ceol, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38(suppl 1):D532–D539, 2010.
- Guoan Chen, Hong Wang, Charles T Miller, Dafydd G Thomas, Tarek G Gharib, David E Misek, Thomas J Giordano, Mark B Orringer, Samir M Hanash, and David G Beer. Reduced selenium-binding protein 1 expression is associated with poor outcome in lung adenocarcinomas. *The Journal of Pathology*, 202(3):321–329, 2004.

- Lingyi Chen and Jonathan Widom. Mechanism of transcriptional silencing in yeast. *Cell*, 120(1):37–48, 2005.
- Wei Chu, Zoubin Ghahramani, Francesco Falciani, and David L Wild. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 21(16):3385–3393, 2005.
- Han-Yu Chuang, Laura Rassenti, Michelle Salcedo, Kate Licon, Alexander Kohlmann, Torsten Haferlach, Robin Foà, Trey Ideker, and Thomas J Kipps. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood*, 120(13):2639–2649, 2012.
- Nicole Cloonan, Alistair RR Forrest, Gabriel Kolle, Brooke BA Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–619, 2008.
- Yanming Di, Daniel W Schafer, Jason S Cumbie, and Jeff H Chang. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology*, 10(1):24, 2011.
- Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.
- Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pages 71–103, 2003.
- Adam Ertel and Aydin Tozeren. Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC Genomics*, 9(1):3, 2008.

Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Pauline A Fujita, Brooke Rhead, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Melissa S Cline, Mary Goldman, Galt P Barber, Hiram Clawson, Antonio Coelho, et al. The UCSC genome browser database: update 2011. *Nucleic Acids Research*, 39(suppl 1):D876–D882, 2011.

Carole Goble, Robert Stevens, et al. State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687–693, 2008.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Lesley H Greene, Tony E Lewis, Sarah Addou, Alison Cuff, Tim Dallman, Mark Dibley, Oliver Redfern, Frances Pearl, Rekha Nambudiry, Adam Reid, et al. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research*, 35(suppl 1):D291–D297, 2007.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinform-*



*matics*, 11(1):422, 2010.

Birte Hellwig, Jan G Hengstler, Marcus Schmidt, Mathias C Gehrman, Wiebke Schormann, and Jörg Rahnenführer. Comparison of scores for bimodality of gene expression distributions and genome-wide evaluation of the prognostic relevance of high-scoring genes. *BMC Bioinformatics*, 11, 2010.

Norman Huang, Parantu K. Shah, and Cheng Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13 (3):305–316, 2012.

Chung J Hung, David G Ginzinger, Raza Zarnegar, Hajime Kanauchi, Mariwil G Wong, Electron Kebebew, Orlo H Clark, and Quan-Yang Duh. Expression of vascular endothelial growth factor-c in benign and malignant thyroid tumors. *Journal of Clinical Endocrinology & Metabolism*, 88(8):3694–3699, 2003.

Iñaki Inza, Pedro Larrañaga, Rosa Blanco, and Antonio J Cerrolaza. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine*, 31(2):91–103, 2004.

Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010.

Thomas Karn, Lajos Pusztai, Eugen Ruckhaberle, Cornelia Liedtke, Volkmar Muller, Marcus Schmidt, Dirk Metzler, Jing Wang, Kevin R Coombes, Regine Gatje, Lars Hanker, Christine Solbach, Andre Ahr, Uwe Holtrich, Achim Rody, and Manfred Kaufmann. Melanoma antigen family A identified by the bimodality index defines a subset of triple negative breast cancers as candidates for immune response augmentation. *European Journal of Cancer*, 48(1):12–23, 2012.

Hyunki Kim, Hyun Ju Kang, Kwon Tae You, Se Hoon Kim, Kang Young Lee, Tae Il Kim, Chul Kim, Si Young Song, Hye-Jung Kim, Cheolju Lee, et al. Suppression of human selenium-binding protein 1 is a late event in colorectal carcinogenesis and is associated with poor survival. *Proteomics*, 6(11):3466–3476, 2006.

Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany*, 99(2):248–256, 2012.

Leo Lahti, Martin Schafer, Hans-Ulrich Klein, Silvio Bicciato, and Martin Dugas. Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. *Briefings in Bioinformatics*, 2012.

Kim-Anh Lê Cao, Pascal GP Martin, Christèle Robert-Granié, and Philippe Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, 10(1):34, 2009.

Jae Won Lee, Jung Bok Lee, Mira Park, and Seuck Heun Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885, 2005.

Soohyun Lee, Chae Hwa Seo, Byungho Lim, Jin Ok Yang, Jeongsu Oh, Minjin Kim, Sooncheol Lee, Byungwook Lee, Changwon Kang, and Sanghyuk Lee. Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Research*, 39(2):e9–e9, 2011.

Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Springer, 1998.

Riku Louhimo and Sampsa Hautaniemi. CNAmet: an R package for integrating

- copy number, methylation and expression data. *Bioinformatics*, 27(6):887–888, 2011.
- Riku Louhimo, Tatiana Lepikhova, Outi Monni, and Sampsa Hautaniemi. Comparative analysis of algorithms for integration of copy number and expression data. *Nature Methods*, 9, 2012.
- Matthieu Louis and Attila Becskei. Binary and graded responses in gene networks. *Science STKE*, 2002(143):33, 2002.
- J Luo, M Schumacher, A Scherer, D Sanoudou, D Megherbi, T Davison, T Shi, W Tong, L Shi, H Hong, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *The Pharmacogenomics Journal*, 10(4):278–291, 2010.
- David Magis. catR: An R Package for Computerized Adaptive Testing. *Applied Psychological Measurement*, 2011.
- Elaine R Mardis. The \$1,000 genome, the \$100,000 analysis. *Genome Medicine*, 2(11):84, 2010.
- John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.
- Donald W Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612, 1970.
- Peter McCullagh and John A Nelder. *Generalized linear models*. Chapman & Hall/CRC, second edition, 1989.
- Lauren M McIntyre, Kenneth K Lopiano, Alison M Morse, Victor Amin, Ann L

- Oberg, Linda J Young, and Sergey V Nuzhdin. RNA-seq: technical variability and sampling. *BMC Genomics*, 12(1):293, 2011.
- Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G Van Meir, Daniel J Brat, Gena M Mastrogianakis, Jeffrey J Olson, Tom Mikkelsen, Norman Lehman, Ken Aldape, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.
- Gideon J Mellenbergh. A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, 29:223–236, 1994.
- Renée X Menezes, Marten Boetzer, Melle Sieswerda, Gert-Jan B van Ommen, and Judith M Boer. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*, 10(1):203, 2009.
- Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008.
- Irini Moustaki. A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49(2):313–334, 2011.
- Eiji Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied pPsychological Measurement*, 16(2):159–176, 1992.
- Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933.
- Yvonnick Noel and Bruno Dauvier. A beta Item Response Model for continuous

bounded responses. *Applied Psychological Measurement*, 31(1):47–73, 2007.

Helen Parkinson, Ugis Sarkans, Nikolay Kolesnikov, Niran Abeygunawardena, Tony Burdett, Mirosław Dyląg, Ibrahim Emam, Anna Farne, Emma Hastings, Ele Holloway, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research*, 39(suppl 1):D1002–D1004, 2011.

Michael J Pencina, Ralph B D’Agostino Sr, Ralph B D’Agostino Jr, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172, 2008.

Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4(1):53–77, 2010.

MS Pepe, Z Feng, and JW Gu. Comments on ‘Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by MJ Pencina et al., *Statistics in Medicine*. *Statistics in Medicine*, 27(2):173–181, 2007.

Jeremy D Peterson, Lowell A Umayam, Tanja Dickinson, Erin K Hickey, and Owen White. The comprehensive microbial resource. *Nucleic Acids Research*, 29(1):123–125, 2001.

Mark J Pletcher and Michael Pignone. Evaluating the Clinical Utility of a Biomarker. *Circulation*, 123(10):1116–1124, 2011.

Jonathan R Pollack, Therese Sørbye, Charles M Perou, Christian A Rees, Stefanie S Jeffrey, Per E Lonning, Robert Tibshirani, David Botstein, Anne-Lise

- Børresen-Dale, and Patrick O Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, 99(20):12963, 2002.
- Stan Pounds and Stephan W Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, 19(10):1236–1242, 2003.
- Kim D Pruitt, Tatiana Tatusova, William Klimke, and Donna R Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(suppl 1):D32–D36, 2009.
- Dimitris Rizopoulos. ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25, 2006.
- Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25, 2010.
- Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.
- Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.
- Peter W Rose, Bojan Beran, Chunxiao Bi, Wolfgang F Bluhm, Dimitris Dimitropoulos, David S Goodsell, Andreas Prlić, Martha Quesada, Gregory B Quinn, John D Westbrook, et al. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Research*, 39(suppl 1):D392–D401, 2011.

Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(suppl 1):D449–D451, 2004.

Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1969.

Eric W Sayers, Tanya Barrett, Dennis A Benson, Evan Bolton, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, 2011.

Robert E Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Nonlinear Estimation and Classification*, 2002.

Kerby Shedden, Jeremy MG Taylor, Steven A Enkemann, Ming-Sound Tsao, Timothy J Yeatman, William L Gerald, Steven Eschrich, Igor Jurisica, Thomas J Giordano, David E Misek, et al. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14(8):822–827, 2008.

Leming Shi, Gregory Campbell, Wendell D Jones, Fabien Campagne, Zhining Wen, Stephen J Walker, Zhenqiang Su, Tzu-Ming Chu, Federico M Goodsaid, Lajos Pusztai, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28(8):827, 2010.

Sudeep Srivastava and Liang Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38(17):e170–e170, 2010.

Barbara E Stranger, Matthew S Forrest, Mark Dunning, Catherine E Ingle,

- Claude Beazley, Natalie Thorne, Richard Redon, Christine P Bird, Anna de Grassi, Charles Lee, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848, 2007.
- Aik Choon Tan, Daniel Q Naiman, Lei Xu, Raimond L Winslow, and Donald German. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.
- Andrew E Teschendorff, Ali Naderi, Nuno L Barbosa-Morais, and Carlos Caldas. PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer. *Bioinformatics*, 22(18):2269–2275, 2006.
- Andrew E Teschendorff, Ahmad Miremadi, Sarah E Pinder, Ian O Ellis, and Carlos Caldas. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, 8(8):157, 2007.
- Zivana Tevzak, Marina V Kondratovich, and Elizabeth Mansfield. US FDA and personalized medicine: in vitro diagnostic regulatory perspective. *Personalized Medicine*, 7(5):517–530, 2010.
- The Cancer Genome Atlas Research Network. INTEGRATED GENOMIC ANALYSIS OF OVARIAN CARCINOMA. *Nature*, 474:609–615, 2011.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 73:267–288, 1996.
- Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.



- Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310(5748):644–648, 2005.
- Pan Tong and Kevin R Coombes. integIRTy: a method to identify genes altered in cancer by accounting for multiple mechanisms of regulation using Item Response Theory. *Bioinformatics*, 2012.
- Pan Tong, Yong Chen, Xiao Su, and Kevin R Coombes. SIBER: systematic identification of bimodally expressed genes using RNAseq data. *Bioinformatics*, 2013.
- Sandra Waaijenborg, Philip C Verselewe de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7(1):3, 2008.
- Jing Wang, Sijin Wen, W Fraser Symmans, Lajos Pusztai, and Kevin R Coombes. The Bimodality Index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. *Cancer Informatics*, 7:199–216, 2009.
- Xin Victoria Wang, Roel GW Verhaak, Elizabeth Purdom, Paul T Spellman, and Terence P Speed. Unifying gene expression measures from multiple platforms using factor analysis. *PloS One*, 6(3):e17691, 2011.
- Lodewyk FA Wessels, Marcel JT Reinders, Augustinus AM Hart, Cor J Veenman, Hongyue Dai, Yudong D He, and Laura J Van’t Veer. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–3762, 2005.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515, 2009.

Xiwei Wu, Tibor A Rauch, Xueyan Zhong, William P Bennett, Farida Latif, Dietmar Krex, and Gerd P Pfeifer. CpG island hypermethylation in human astrocytomas. *Cancer Research*, 70(7):2718–2727, 2010.

Meiheng Yang and Arthur J Sytkowski. Differential expression and androgen regulation of the human selenium-binding protein gene hSP56 in prostate cancer cells. *Cancer Research*, 58(14):3150–3153, 1998.

Zhang Zhang and Jeffrey P Townsend. The filamentous fungal gene expression database (FFGED). *Fungal Genetics and Biology*, 47(3):199, 2010.

Zhang Zhang, Vladimir B Bajic, Jun Yu, Kei-Hoi Cheung, and Jeffrey P Townsend. Data integration in bioinformatics: Current efforts and challenges. *Bioinformatics-Trends and Methodologies*, pages 41–56, 2011.

## VITA

Pan Tong was born in Wuhan city, Hubei province, China on November 28, 1985, the son of Weizheng Tong and Jinwen Shen. After completing his degree at Xinzhou Number 2 Middle School at Xinzhou district, Wuhan in 2004, he entered Huazhong University of Science and Technology at Wuchang district, Wuhan where he completed a Bachelor of Engineering degree in Bioinformatics in 2008. In August 2008, he moved to Houston, Texas in the United States to pursue his Ph.D. degree in Biostatistics at the University of Texas Health Science Center at Houston and the University of Texas MD Anderson Cancer center. In September 2009, he joined the laboratory of Dr. Kevin Coombes for his Ph.D. thesis after finishing several tutorials with the faculties in the Division of Quantitative Sciences at MD Anderson Cancer center.

Permanent Address:

Wangji street, Wangxin road, room 51, Xinzhou district, Wuhan, Hubei, China,  
430418