Prognostic biomarker detection, machine learning bias correction, and differential coexpression module detection

by

Ying Ding

B.E. in Bioinformatics, Huazhong University of Science and Technology, 2007

Submitted to the Graduate Faculty of

The School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH

School of Medicine

This dissertation was presented

by

Ying Ding

It was defended on

April 8, 2014

Dr. Etienne Sibille, Associate Professor, Department of Psychiatry

Dr. Daniel E. Weeks, Professor, Departments of Human Genetics and Biostatistics

Dr. Takis Benos, Associate Professor, Departments of Computational and Systems Biology

Dr. Ziv Bar-Joseph, Associate Professor, Department of Machine Learning

Dissertation Advisor: Dr. George Tseng, Associate Professor, Department of Biostatistics

Prognostic biomarker detection, machine learning bias correction, and differential coexpression module detection

Ying Ding, PhD

University of Pittsburgh, 2014

In this thesis, we present three projects on prognosis biomarker detection, machine learning bias correction and identification of differential coexpression modules in complex diseases. In the first project, we aimed to identify fusion transcripts that are of predictive value on prostate cancer prognosis, an important task to avoid overtreatment to patients. We discovered eight fusion transcripts from 19 RNA-seq datasets and validated its predictive value on >200 patients from three sites (Pittsburgh, Stanford and Wisconsin). The constructed prediction model showed consistently high accuracy on predicting prostate cancer recurrence and aggressiveness in all three cohorts. In the second project, we consider a common practice to apply many (up to several hundred) machine learning classifiers to a dataset and report the best cross-validated accuracy. We demonstrated a downward bias using this approach and proposed an inverse power law (IPL) method to correct the bias. The method was compared with several existing methods using simulation and real datasets and showed superior performance. For the third study, we developed a computational algorithm (MetaDiffNetwork) to identify coexpressioin modules that are consistently differential across disease conditions in multiple transcriptomic studies. We demonstrated good performance of the algorithm using simulated data and applied it to combine eight major depressive disorder studies (cases vs. controls) and four breast cancer studies (ER+ vs. ER-). The identified modules were validated by existing knowledge of disease pathways. These modules can be used to help generate new hypotheses regarding suspected disease genes. In conclusion, the three areas of research covered in this thesis are critical bioinformatic elements for biomedical applications and can be used to help understand the underlying disease mechanism and ultimately improve patient treatment.

TABLE OF CONTENTS

PREFACE 12

1.0 INTRODUCTION 13

1.1 BACKGROUND AND MOTIVATION 13

1.2 HIGH-THROUGHPUT TECHNIQUES FOR BIOMARKER DISCOVERY 16

1.2.1 Microarrays 17

1.2.2 RNA sequencing 18

1.3 BIOINFORMATICS ANALYSES ON HIGH-THROUGHPUT DATA 19

1.3.1 From differential expression to differential coexpression 19

1.3.2 Differential coexpression network 20

1.3.3 Pathway analysis 21

1.3.4 Machine learning methods 22

1.4 OVERVIEW OF THE THESIS 23

2.0 TRANSLOATION AND FUSION TRANSCRIPTS IN PROGRESSIVE PROSTATE CANCER 28

2.1 INTRODUCTION 32

2.1.1 Fusion genes 32

2.1.2 Existing tools on fusion gene/transcript detection with RNA-Seq 33

2.1.3 Existing measures for predicting prostate cancer recurrence.................. 35

2.2 DESIGN AND HYPOTHESIS 37

2.3 METHODS.. 38

2.3.1 Tissue samples.. 38

2.3.2 Whole genome and transcriptome sequencing library preparation and sequencing.. 39

2.3.3 Read alignment 39

2.3.4 Fusion transcript detection 40

2.3.5 Machine learning classifier to predict recurrence status.......................... 40

2.3.6 RT-PCR and FISH 41

2.4 RESULTS.. 41

2.4.1 Fusion transcripts discovered by RNA and whole genome sequencing ... 41

2.4.2 Fluorescence in situ hybridization suggests genome recombination underlying fusion transcript formation 43

2.4.3 Fusion transcripts associate with PCa recurrence 44

2.5 DISCUSSION.. 47

3.0 BIAS CORRECTION FOR SELECTING THE MINIMAL-ERROR CLASSIFIERS FROM MANY MACHINE LEARNING MODELS.................................... 54

3.1 INTRODUCTION 57

3.2 METHODS.. 60

3.2.1 Problem setting and formulation under simulation scheme...................... 60

3.2.2 Illustration by a 2-D toy model.. 61

3.2.3 Properties and asymptotic theorems of MEC bias 64

3.2.4 Three existing bias correction methods 65

3.2.5 The resampling-based inverse power law (IPL) method 67

3.3 RESULTS 69

3.3.1 Bias correction of the simulated 2D example 69

3.3.2 Application on five GEO datasets 70

3.3.3 TCGA and METABRIC breast cancer data 71

3.4 CONCLUSION AND DISCUSSION 73

4.0 DIFFERENTIAL COEXPRESSED MODULE DETECTION 81

4.1 INTRODUCTION 82

4.2 METHODS 84

4.2.1 Network construction 84

4.2.2 Target function 85

4.2.3 Optimization with simulated annealing 86

4.2.4 Seed module generation before optimization 87

4.2.5 Control of False Discovery Rate (FDR) 88

4.2.6 Assembly of small modules 89

4.2.7 Weight parameter tuning 90

4.3 RESULTS 90

4.3.1 Simulation datasets 91

4.3.2 Breast cancer datasets 92

4.3.3 Major depressive disorder (MDD) example 94

4.4 CONLUSION 97

5.0 CONCLUSIONS AND FUTURE DIRECTIONS ... 108

5.1 SUMMARY OF CONTRIBUTIONS ... 108

5.2 FUTURE DIRECTIONS ... 110

5.2.1 Combining fusion gene with existing measures to improve prediction ... 111

5.2.2 Power analysis on fusion gene analysis using RNA-Seq ... 111

5.2.3 Increasing the speed of IPL and taking into consideration the dependence structure of classifiers in bias correction ... 112

5.2.4 Improving module assembly process ... 113

5.2.5 Summarizing differential coexpressed modules to aid biological exploration ... 113

APPENDIX A - SUPPLEMENTARY INFORMATION FOR CHAPTER 2 ... 115

APPENDIX B - SUPPLEMENTARY INFORMATION FOR CHAPTER 3 ... 118

BIAS CORRECTION FOR SELECTING THE MINIMAL-ERROR CLASSIFIERS FROM MANY MACHINE LEARNING MODELS ... 118

BIBLIOGRAPHY ... 122

LIST OF TABLES

Table 2-1 A summary of eight fusion detection tools.. 35

Table 3-1 Error rate estimates on 5 GEO datasets.. 79

Table 4-1 Data description of four breast cancer datasets 99

Table 4-2 Top pathways enriched after module assembly for breast cancer datasets (densely connected in ER+ group). 100

Table 4-3 Top pathways enriched after module assembly for breast cancer datasets (densely connected in ER- group). 101

Table 4-4 Data description of eight MDD microarray studies. 104

Table 4-5 Top pathways enriched after module assembly for MDD datasets (densely connected in control)... 106

Table 4-6 Top pathways enriched after module assembly for MDD datasets (densely connected in MDD).. 107

LIST OF FIGURES

Figure 1-1 An overview of high-throughput data analysis. 16

Figure 1-2 Summary of molecular, cellular, tissue and technical regulatory sources of observed gene-gene correlations/coexpression links. 26

Figure 1-3 Gene expression patterns translate regulatory changes into networks links. 27

Figure 2-1 Unique fusion gene events. 49

Figure 2-2 Fluorescence in situ hybridization suggests genome recombination in prostate cancer cells. 50

Figure 2-3 Fusion transcripts expressed in prostate cancer. 51

Figure 2-4 Fusion genes predict recurrence of prostate cancer. 52

Figure 3-1 The framework for estimating the true optimal classification error rate, MEC error rate and its bias from independent simulated data sets. 75

Figure 3-2 Error rate estimates in two dimensional toy model. 76

Figure 3-3 Illustration of the downward bias of MEC error rate 77

Figure 3-4 Comparison of minimal error rate estimates from all four bias correction methods. . 78

Figure 3-5 IPL estimates along with extrapolation. 79

Figure 3-6 Error ratas estimates on real datasets. 80

Figure 4-1 Module output before and after optimization in simulation datasets. 98

Figure 4-2 Energy trace plot in simulated annealing. 99

Figure 4-3 Two example modules in breast cancer datasets. Example module densely connected in ER+ group but not in ER- group. (B) example module densely connected in ER- group but not in ER+ group. 104

Figure 4-4 Two example modules in MDD datasets. (A) Example module densely connected in MDD group but not in control group. (B) example module densely connected in control group but not in MDD group. 106

Figure S-5-1 Circus plots of prostate cancer functional genome translocation. 116

Figure S-5-2 Genome breakpoint analysis of fusion genes 117

PREFACE

I wish to thank the members of my thesis committee for their expertise and for how they have provided me with guidance on the interdisciplinary thesis topic. I want to thank Dr. George Tseng for providing me opportunities working in the high dimensional genomic research as well as involving me in the projects with next generation sequencing. It is his patience advising me on all the research projects these years that make me learn a lot as well as make the progress goes smoothly.

I would like to thank Dr. Etienne Sibille, as my co-advisor, who showed me the passion towards research as well as rigorousness. He provides me the opportunities for different collaborative projects and gives me a broader knowledge and horizon.

I would like to thank Dr. Ziv Bar-Joseph and Dr. Daniel Weeks for their time on providing detailed comments on improving my thesis writing.

I would also like to thank my previous advisor Dr. Daniel Zuckerman for his support for my interest as well as his wish for my thesis. I have gained precious experience in the Monte Carlo simulation and sampling while working in his research group.

I would like to dedicate the work's best aspects to my parents, my wife and my friends.

To my parents, who give me love, education, support and patience and who are the ultimate Harbor for me. To my wife, who accompany me through my Ph.D., gives me encouragement when I face difficulties.

And, to my dear friends, whom I trust, you have accompanied me in different periods of my life; given me happiness and courage to walk along the road and never give up.

1.0 INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

The search for disease-related biomarker detection and their related predictive values is an important task in biomedical research. Biomarkers have wide applications from diagnosis and prognosis to treatment selection and preventative intervention. The definition of biomarkers may vary by different researchers and fields; here we refer to the definition by the Biomarkers Definitions Working Group (Atkinson et al., 2001; Ziegler et al., 2012): "A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacologic responses to a therapeutic intervention." With the advance of new technologies, high-throughput experimental techniques have revolutionized biomedical research with various array based platforms and now next-generation sequencing. An increasing amount of high-throughput datasets are available, which allows a genome-wide screening for potential biomarkers and the related prediction model. The analysis of high-throughput omics data usually has two major objectives (see diagram in Figure 1). The first common objective is to understand the underlying biological or disease mechanisms through analysis of a patient cohort or various treatment/perturbation of samples. Bioinformatics techniques are often used to explore and discover the useful relationships and patterns in biological data such as differential expression (DE) analysis, clustering analysis, pathway analysis and network analysis (Figure 1-1) (Hawkins, Hon and Ren, 2010; Quackenbush, 2001). These analyses identify potential candidate biomarkers, co-expressed module networks and associated biological pathways. The results can help elucidate the underlying biological/disease mechanisms, assist novel hypothesis generation, identify potential drug targets and guide preventive procedures or disease management for patients (Zografos, Liakakos and Roukos, 2013).

For the second common objective, researchers analyze high-throughput omics data to develop classifiers for disease diagnosis or prognosis prediction, treatment selection and preventative intervention. For example, in prostate cancer, biomarker tests can be used to determine whether radio- or chemotherapy can be obviated. Therefore, the overtreatment of this disease can be reduced with the help of the biomarker classifiers. Biomarkers for constructing classifiers can come from the accumulation of biological knowledge from experiments combined with the exploratory bioinformatics analyses, or can come solely from machine learning approaches which demonstrate good prediction performance. Discoveries or translational products from the first objective mentioned above may need extensive validation and evaluation (e.g. multi-center clinical trial) under this approach. Machine learning, clinical trial design and decision theory are common methodologies used for this aim. In many situations, the underlying disease mechanisms related to the gene features used in the classification model are unknown. It is therefore of particular importance to apply unbiased and generalizable machine learning techniques to ensure translational potential of the discovered biomarkers and classification model.

Machine learning methods have been broadly extended in the bioinformatics community with successful applications in many areas (Baek, Tsai and Chen, 2009). High-throughput genomic, proteomic and metabolomic data with a large number of variables and a relatively small number of samples still pose a problem for constructing robust prediction models. Model building with effective feature selection and unbiased performance assessment are two important components in the development of a biomarker classifier.

In summary, developing new methods to discover novel biomarkers which can be used to better understand underlying disease mechanisms (Objective 1) and new methods for evaluation of prediction model performances (Objective 2) can help facilitate the field of personalized medicine and provide clinical utility. In this thesis, we construct prediction models for predicting prostate cancer recurrence and aggressiveness using novel fusion transcripts detected by RNA-Seq technology (Chapter 2) and present a solution to correct the classification error bias when many machine learning methods are applied (Chapter 3). Finally, we present a network-based method by searching differential coexpression patterns between disease and control conditions by combining multiple studies in order to understand potential disease mechanisms (Chapter 4). Taken together, the methods and analyses make a valuable contribution to studying disease mechanisms and biomarker classifier development to ultimately improve disease classification and management.

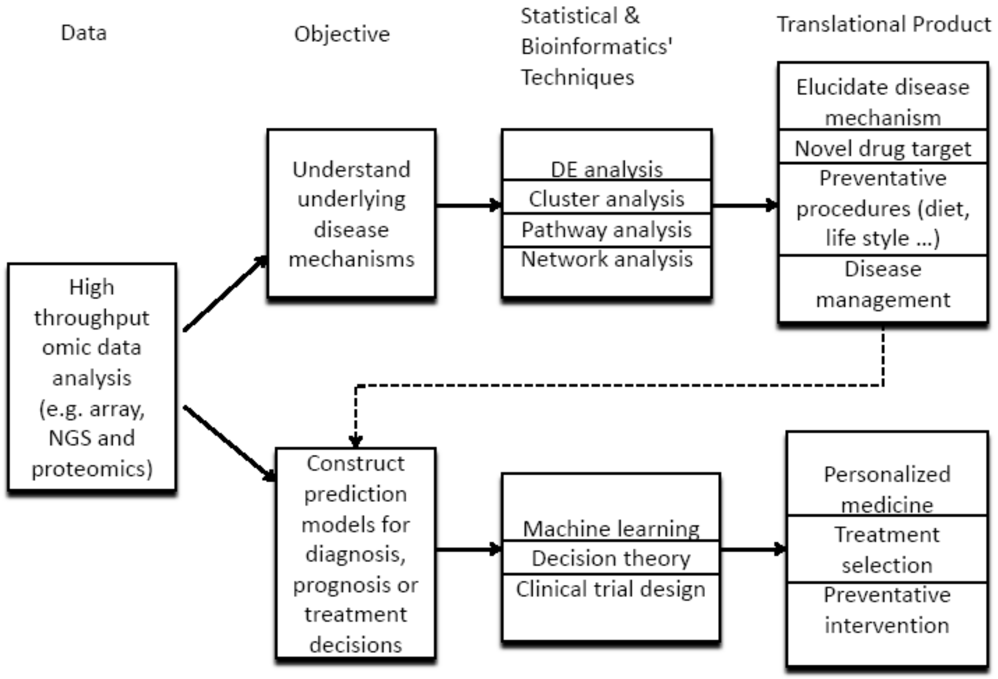
Data
Objective
Statistical & Bioinformatics' Techniques
Translational Product
High throughput omic data analysis (e.g. array, NGS and proteomics)
Understand underlying disease mechanisms
DE analysis
Cluster analysis
Pathway analysis
Network analysis
Elucidate disease mechanism
Novel drug target
Preventative procedures (diet, life style ...)
Disease management
Construct prediction models for diagnosis, prognosis or treatment decisions
Machine learning
Decision theory
Clinical trial design
Personalized medicine
Treatment selection
Preventative intervention

Figure 1-1 An overview of high-throughput data analysis.

1.2 HIGH-THROUGHPUT TECHNIQUES FOR BIOMARKER DISCOVERY

As demonstrated in Figure 1-1, high-throughput omics technologies generate experimental data that help elucidate disease biology and enhance disease treatment. In this section, we will mainly introduce microarray and RNA-seq technologies for gene expression and fusion transcript analyses that are relevant to the later three chapters in the thesis. Many other technologies, such as mass spectrometry, proteomics assays and imaging techniques (e.g. fMRI and PET scan), also generate useful data that apply to methods and concepts we cover in this thesis (especially for the machine learning and network methodologies in Chapter 3 and 4).

1.2.1 Microarrays

DNA microarray is a high-throughput technology that allows for the analysis of the expression of multiple genes in an efficient manner. On a microarray, thousands of probes with known sequence identity are immobilized on a solid support, such as a microscope glass slide or silicon chips. The labeled target sequence binds to probe, allowing for the identification of unknown sequences such that gene expression of thousands of genes can be analyzed simultaneously. In a typical microarray experiment, mRNA is reverse transcribed, amplified and hybridized to the DNA template. The amount of mRNA bound to different sites on the array reflects the expression profiles of thousands of genes, or even the whole genome. Microarrays, with additional specialized design, can also be used to detect single nucleotide polymorphisms (SNPs), copy number variation (CNVs), methylation, and protein-DNA binding.

With microarray technology increasingly used by the scientific community, a multitude of microarray datasets have accumulated, necessitating the creation of databases for storage and public access. The National Center for Biotechnology Information (NCBI) has set up the Gene Expression Omnibus (GEO), which stores many gene expression datasets for ease of public access. Large-scale microarray datasets are also available upon request through METABRIC (Molecular Taxonomy of Breast Cancer International Consortium), which contains a collection of over 2000 clinically annotated primary fresh-frozen breast cancer specimens from tumor banks in the UK and Canada (Curtis et al., 2012).

In this thesis, several GEO microarray datasets as well as METABRIC microarray datasets have been used to demonstrate the performance of IPL, a bias correction tool we developed to correct the bias of reporting the minimal cross-validation error rate when multiple machine learning models are applied. This further enabled us to compare IPL's performance with several other existing bias correction tools (Chapter 3). GEO breast cancer datasets and METABRIC microarray dataset have also been used to combine with other breast cancer gene expression datasets, to demonstrate the performance of MetaDiffNetwork, a method to detect differential modules across disease conditions across multiple studies we developed (Chapter 4).

1.2.2 RNA sequencing

RNA-Seq (RNA Sequencing), which makes use of the capabilities of next-generation sequencing, has been increasingly popular and surpasses microarray as the preferred choice for gene expression analysis at a much higher resolution (McGettigan, 2013). In an RNA-Seq experiment, sample RNA first undergoes fragmentation by Poly-A selection, then RNA fragments are reverse transcribed and amplified into cDNA fragments. After size selection, the fragments are ligated by sequencing primers and sequenced by a sequencing machine such as Illumina Genome Analyzer and Hi-Seq in a massively parallel fashion. Illumina sequencing technology can generate relatively long paired-end reads, up to 100bp, which can enable better mappabilty. After aligning the short reads to the genome, read coverage depth can be counted to estimate gene expression level. The advantages of RNA-Seq over microarrays include but are not limited to: (1) providing much richer information beyond quantification such as alternative exon usage and novel splicing junction detection; (2) non-biased design which is not limited by probe oligo design and can discover novel exons and genes; and (3) providing higher specificity and sensitivity with very low background noise from contaminating genome DNA and higher sensitivity by detecting more genes and differential genes. Paired-end RNA-Seq is also suitable to detect fusion genes. Candidate fusion transcripts can be preliminarily identified by aligning each end of the paired reads to two different genes or two distant regions. The Cancer Genome Atlas (TCGA) projects have also used the RNA-Seq approach to profile thousands of primary tumor samples from 30 different tumor types to understand the underlying mechanism of malignant transformation and progression (http://tcga-data.nci.nih.gov/tcga).

In this thesis, RNA-Seq experiments were conducted on 19 prostate tissues to find tumor specific fusion transcripts (Chapter 2). We also utilized an RNA-Seq breast cancer dataset from TCGA to demonstrate the performance of IPL (Chapter 3) and construct a gene coexpression network (Chapter 4).

1.3 BIOINFORMATICS ANALYSES ON HIGH-THROUGHPUT DATA

In this section, we will introduce several popular bioinformatics analyses that are commonly encountered in high-throughput omics data analysis. We will also link their usage to the three major projects (Chapter 2-4) in this thesis.

1.3.1 From differential expression to differential coexpression

Differential gene expression analysis is typically conducted by comparing gene expression levels, either from microarray or RNA-Seq, between two conditions such as disease and control. It aims to find a set of genes whose mean expression levels are significantly different between two set of samples through statistical tests. Due to the fact that thousands of genes are being tested simultaneously, the multiple hypotheses testing problem is usually solved by controlling the false discovery rate (Reiner, Yekutieli and Benjamini, 2003) . The genes found to have an increase or decrease in mean expression levels may associate with the disease phenotype.

Gene expression datasets contain more information than differential expression studies can extract. In addition to testing differential expression, differential coexpression (generally measured by pairwise correlations) can help reveal the dysfunctional regulation in a disease. The pairwise relationships between gene expression levels can reflect various regulatory sources (Figure 1-2) such as transcription factor (TF) binding (Gaiteri et al., 2014; Marco et al., 2009), the affect of histones on access of transcription factors to DNA (Chen and Zhao, 2005; Deng et al., 2010), miRNA regulation (Baskerville and Bartel, 2005; Dong et al., 2010) and so on. Differential coexpression between two genes or groups of genes may or may not accompany differential expression. As shown in Figure 1-3, two genes with similar mean expression levels between disease and control conditions could have completely different coexpression patterns due to the disruption of transcription factor regulation of one of the genes in the disease condition. Therefore, changes in gene-gene correlation may occur in the absence of differential expression, meaning that a gene may undergo changes in regulatory pattern that would be undetected by traditional differential expression analyses (Gaiteri et al., 2014).

1.3.2 Differential coexpression network

Studies of differential coexpression involve constructing two networks in two groups of samples under two conditions, such as healthy versus disease samples. To build gene coexpression networks, the gene-gene correlations are commonly detected using Pearson correlations or other robust forms (e.g. Spearman correlation or leave-one-out jackknife correlation). For a gene expression dataset of n samples and p genes, a raw correlation matrix of $p \times p$ is calculated. After various transformations, this becomes an adjacency matrix that describes the existence of links between pairs of genes. Transforming correlation matrices into adjacency matrices is important because the selection of links will affect the biological conclusions drawn from network structure. Usually, for an unweighted network, a threshold is used to decide how large the correlation value will create a link between two genes. Taking a fraction (e.g. 1%) of top correlation values has been shown to select genes with related functions (Lee et al., 2003). Comparing the structure of two coexpression networks could provide insight into the disease-relevant regulatory changes underlying the coexpression patterns. One simple way is to look at the genes with differential connectivity (Reverter et al., 2006). The other alternative is to examine the differential relationship at the module level to greatly reduce the number of statistical tests and to check for coherent correlation changes in the disease state (Amar, Safer and Shamir, 2013; Kostka and Spang, 2004).

In this thesis, coexpression networks are the focus in MetaDiffNetwork, a method we developed to identify differentially co-expressed modules that are consistently observed across multiple studies (Chapter 4).

1.3.3 Pathway analysis

When a set of genes is obtained from differential expression analysis or differential coexpression analysis, the next questions are how to interpret the results and how to gain insights into biological mechanisms. Pathway analysis is a useful class of methods to provide such information based on checking over-enrichment of the detected genes in gene sets based on prior biological information such as published biochemical pathways. Many tools have been developed to check gene set enrichment. The most common way is to apply the Fisher's exact test or the so-called cumulative hypergeometric test on the set of discovered genes and assess the p value of over-enrichment. More advanced methods are available, such as Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) which considers all the genes in the experiment, not just those passing a pre-specified significance threshold.

In this thesis, pathway analysis is conducted on output module gene sets from MetaDiffNetwork to provide insight on potential biological mechanisms that contribute to differential regulatory patterns (Chapter 4).

1.3.4 Machine learning methods

Machine learning methods are essential for biomarker discovery and translational validation process. Due to the availability of many high-throughput experimental data, the applications of machine learning techniques are growing. The two most common classes of machine learning algorithms are supervised learning and unsupervised learning. Supervised learning starts with a known training data and known classification labels (such as case and control labels) and its goal is to construct a prediction model to predict target output of future test data with unknown class label. This problem is also known as classification when the output values are discrete. Unsupervised learning starts with training data with the goal of discovering the clusters so that the data in each cluster are similar to each other without the need to know the labels of the data. One commonly encountered problem of biomedical applications is the small sample size but large number of features (small-n-large-p), which could dilute the performance of machine learning techniques, therefore, feature selection is also a key component in the classifier construction.

In this thesis, we applied many (up to hundreds) supervised learning methods on a 2-D toy model and two large breast cancer gene expression datasets to illustrate the machine learning bias problem and demonstrate the performances of IPL and several other existing methods (Chapter 3).

1.4 OVERVIEW OF THE THESIS

This thesis composes of three parts, one application study and two novel methodological approaches, which are tightly related to various aspects of biomarker detection and subsequent translational research as shown in Figure 1-1. Below we present a brief overview of this thesis.

In Chapter 2, from a collaborative study, we presented a set of eight fusion transcripts, detected from 19 RNA-Seq data with deep coverage and validated by experiments, are of predictive value for prostate cancer recurrence and aggressiveness. After analyzing these fusion transcripts of more than 200 patients from three different centers across Pittsburgh, Stanford and Wisconsin, prediction models constructed from Pittsburgh training cohort show consistent prediction performance on Pittsburgh test cohort, Stanford cohort and Wisconsin cohort.

In Chapter 3, we present a new method, inverse power law (IPL) method, for correcting the bias of reporting the minimal cross validation error rate when multiple (can be as large as several hundred) machine learning methods are used. We first illustrate this downward bias problem and the relationship among sample size, bias and number of classifiers with a simple 2-D toy model. Then we show that IPL gives more robust estimates of true minimal error rates compared to three existing bias correction methods on both simulation datasets as well as two large breast cancer microarray datasets: TCGA and METABRIC. IPL is a useful tool to provide corrected classification error rate when multiple machine learning methods are applied, especially for clinical study with a small sample size with no external validation datasets. The bias correction is critical to avoid overly optimistic accuracy estimate that cannot be validated in subsequent translational research and clinical evaluation.

In Chapter 4, we propose a method MetaDiffNetwork, for detecting differentially coexpressed modules between different disease conditions across multiple studies. MetaDiffNetwork outputs modules with consistent density differences between two conditions by optimizing a weighted target function. It contains steps to control the false discovery rate of detected modules by permutation as well as to tune the weights of the target function. By incorporating the step of pathway guided module assembly, it can assembly small modules into larger modules with respected to a potential disease related pathway. We demonstrate this method on a simulation dataset and show good performance on four breast cancer datasets (ER+ vs ER-) and eight major depressive disorder datasets (MDD vs control). The identified modules are validated by significant enrichment in specific disease relevant pathways.

In summary, this thesis presents topics studying disease mechanisms and biomarker classifier development which may ultimately improve disease classification and management.

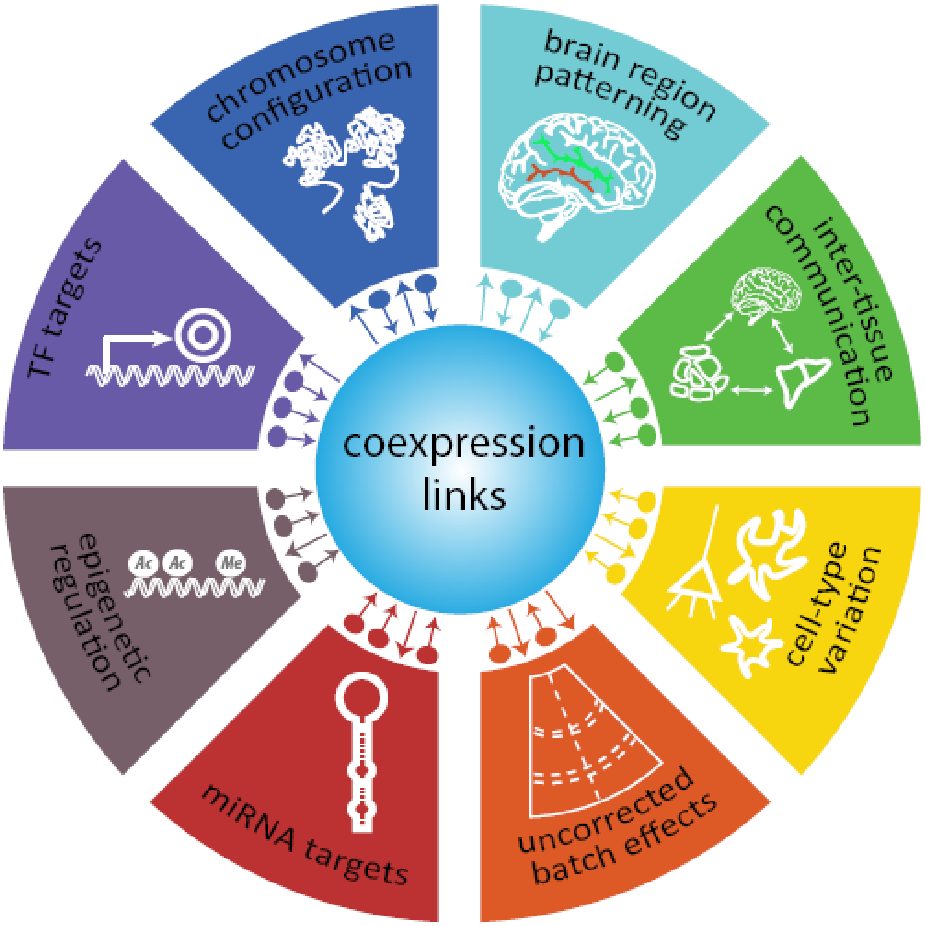

Figure 1-2 Summary of molecular, cellular, tissue and technical regulatory sources of observed gene-gene correlations/coexpression links.

This figure and caption are used with permission from a previous publication (Gaiteri et al., 2014).

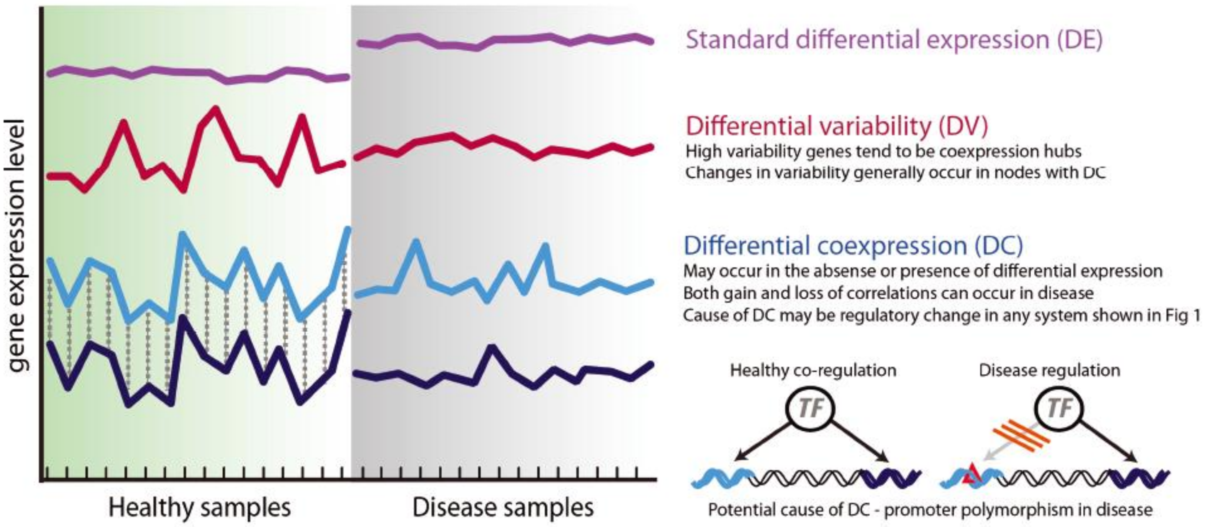
gene expression level
Healthy samples
Disease samples
Standard differential expression (DE)
Differential variability (DV)
High variability genes tend to be coexpression hubs
Changes in variability generally occur in nodes with DC
Differential coexpression (DC)
May occur in the absense or presence of differential expression
Both gain and loss of correlations can occur in disease
Cause of DC may be regulatory change in any system shown in Fig 1
Healthy co-regulation
Disease regulation
TF
TF
Potential cause of DC - promoter polymorphism in disease

Figure 1-3 Gene expression patterns translate regulatory changes into networks links.

This figure and caption are used with permission from a previous publication (Gaiteri et al., 2014).

2.0 TRANSLOATION AND FUSION TRANSCRIPTS IN PROGRESSIVE PROSTATE CANCER

Yan P. Yu, M.D., Ph.D.[+,*], Ying Ding, B.S.[+,‡], Zhanghui Chen, Ph.D.[+], Amantha Michalopoulos, B.S., Rui Chen, B.S. [‡], Zulfiqar G. Gulzar, Ph.D.[Ψ], Bing Yang, Ph.D. [Π], Kathleen Cieply, B.S., Alyssa Luvison, B.S., Bao-Guo Ren, M.D., James D. Brooks[Ψ], M.D., David Jarrard, M.D. [Π], Joel B. Nelson, M.D.[ϕ,*], George Michalopoulos, M.D., Ph.D, George C. Tseng, Ph.D.[*,‡] and Jian-Hua Luo, M.D., Ph.D.[*,ξ]

Department of Pathology, Biostatistics[‡], and Urology[ϕ], University of Pittsburgh School of Medicine, 3550 Terrace Street, Pittsburgh, PA 15261; Department of Urology[Ψ], Stanford University School of Medicine, 875 Blake Wilbur Dr, Stanford, CA 94305; Department of Urology[Π], University of Wisconsin School of Medicine at Madison, 1685 Highland Ave, Madison, WI 53705

+-Authors make equal contribution to the work

*-Authors direct the study equally

ξ-To whom correspondence should be addressed: Jian-Hua Luo, M.D., Ph.D., Scaife S-760, Department of Pathology, University of Pittsburgh School of Medicine, 3550 Terrace Street, Pittsburgh, PA 16251

Short title: Signature fusion genes for prostate cancer

Key words: fusion transcript, prostate cancer, field effect, whole genome sequencing

Acknowledgement: This work is supported by National Institutes of Health (RO1 CA098249 to JHL and 1U01CA152737-01 to JDB), American Cancer Society (RSG-08-137-01-CNE to YPY) and a grant from University of Pittsburgh Cancer Institute. All authors have no conflict of interest for the study. Dr. Luo had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Preface

Note on the contribution of YD to this paper. He is the sole author for all the statistical analysis, NGS analysis and methods and text related to fusion gene detection as well as model construction. He also read, commented on and edited the remainder of the document. This work is currently under review in JAMA.

Abstract

Importance: Predicting the clinical progression of prostate cancer remains a major challenge. An accurate and reproducible test to assess the potential behavior of prostate cancer is urgently needed.

Objective: To identify biomarkers that are predictive of prostate cancer recurrence or prostate cancer specific death.

Design: Genomic DNA and/or total RNA from 19 specimens of prostate cancer (T), matched adjacent benign prostate tissues (AT), matched blood samples (B) and organ donor prostates (OD) were sequenced. Eight novel fusion transcripts were discovered and validated. These eight novel fusion transcripts were then analyzed on 289 prostate samples from University of Pittsburgh Medical Center, Stanford University Medical Center and University of Wisconsin Madison Medical Center, including 279 prostate cancer and 10 prostate organ donor samples.

Setting: University of Pittsburgh Medical Center, Stanford University Medical Center, and University of Wisconsin Madison Medical Center.

Participants: Two hundred seventy-nine prostate cancer patients who underwent radical prostatectomy from 1992-2013 were selected for fusion transcript expression analysis. The definitive clinical outcomes (recurrence or non-recurrence for more than 5 years after surgery) were known for 83.5% (233/279) of the cohort.

Main measure: To identify the presence of any of the following fusion transcripts in prostate cancer samples: TMEM135-CCDC67, KDM4B-AC011523.2, MAN2A1-FER, TRMT11-GRIK2, CCNH-C5orf30, SLC45A2-AMACR, MTOR-TP53BP1 and LRRC59-FLJ60017.

Results: Prostate cancer recurrence, metastases and/or prostate cancer-specific death after radical prostatectomy occurred in 91% (69/76) of men carrying at least one of eight of these fusion transcripts (TRMT11-GRIK2, SLC45A2-AMACR, MTOR-TP53BP1, LRRC59-FLJ60017, TMEM135 –CCDC67, KDM4-AC011523.2, MAN2A1-FER and CCNH-C5orf30), while these outcomes occurred in only 37% (58/157) of men not carrying those fusion transcripts. Three fusion transcripts occurred exclusively in prostate cancer samples from patients who experienced recurrence or prostate cancer related death. The formation of these fusion transcripts is the result of genome recombination events.

Conclusion and relevance: Specific transcript fusion events are associated with prostate cancer recurrence and might underlie biological aggressiveness.

2.1 INTRODUCTION

Despite a high incidence (Jemal et al., 2012; Siegel, Naishadham and Jemal, 2012), only a fraction of men diagnosed with prostate cancer (PCa) develop metastases and even fewer die from the disease. The majority of prostate cancers remain asymptomatic and clinically indolent. The precise mechanisms for the development of progressive, clinically relevant PCa remain elusive. Furthermore, the inability to predict PCa's potential aggressiveness has resulted in significant overtreatment of the disease. The dichotomous nature of PCa —a subset of life-threatening malignancies in the larger background of histological alterations lacking the clinical features implicit with that label —is a fundamental challenge in disease management.

2.1.1 Fusion genes

Fusion genes (also called chimeric transcripts), formed by two different regions in the chromosomes, are found from cancer. They result from post-transcription and chromosomal rearrangements such as large segment deletion (the well-known fusion *TMPRSS2-ERG* in prostate cancer (Tomlins et al., 2005), chromosome translocation (the well-known fusion *BCR-ABL1* in chronic myeloid leukemia), trans-splicing (Gingeras, 2009) or read-through (two adjacent genes) (Kaye, 2009). Two of the most well-known fusion genes are *BCR-ABL1* found in Chronic myeloid leukemia CML (Fernandez-Luna, 2000; Tkachuk et al., 1990), and *EML4-ALK* in non-small-cell lung cancer (Soda et al., 2007). To date, many fusion genes have been found and collected in fusion gene databases. There are 9054 fusion genes in COSMIC (release 66) (Bamford et al., 2004; Forbes et al., 2011), 1374 fusions genes in TICdb (release 3.3) (Novo, de

Mendibil and Vizmanos, 2007), 1603 fusion genes in the Mitelman database (updated on August 2013) (Mitelman, Johansson and Mertens, 2007; Mitelman, Mertens and Johansson, 2005) and 16261 fusion genes in ChiTaRS (Frenkel-Morgenstern et al., 2013). These databases also gathered related literature regarding these fusion genes. Some of the databases, such as COSMIC, TICdb and ChiTaRS collected fusion gene sequences as well. COSMIC and ChiTaRS offer further summaries of the original tissue type of the fusion genes.

2.1.2 Existing tools on fusion gene/transcript detection with RNA-Seq

Since 2010, many computational tools have been developed for detecting fusion genes using RNA-Seq data especially with paired-end reads. Different properties of eight detection tools were summarized in Table 1-1. These included the well-known Tophat-Fusion (Kim and Salzberg, 2011), ChimeraScan (Iyer, Chinnaiyan and Maher, 2011), and the latest tools, including Fusioncatcher (Asmann et al., 2011), SOAPfuse (Jia et al., 2013), and FusionQ (Liu et al., 2013), which claim their own advantages over several existing tools they chose in their study for performance comparison. Some detection tools utilized machine learning techniques to improve performance, such as deFuse (McPherson et al., 2011) and EricScript (Benelli et al., 2012) and Breakfusion, which assembles reads around the fusion breakpoint (Chen et al., 2012).

These fusion gene detection tools differ from the others from a variety of aspects, from alignment tools to different advanced filtering criteria to final output information. Different tools may apply different alignment tools or a combination of alignment tools. Most tools align all reads to the reference sequence with bowtie (Langmead et al., 2009) such as Tophat-Fusion, ChimeraScan, deFuse, Fusioncatcher, and FusionQ. Other alignment tools such as EricScript and

Breakfusion use bwa (Li and Durbin, 2009) and SOAPfuse uses soap2 (Li et al., 2009). Some also use more than one alignment tool trying to align more reads and locate the fusion breakpoints more accurately. For example, SOAPfuse also uses bwa and blat, while ChimeraScan, deFuse, EricScript, Fusioncatcher, and Breakfusion also use blat. In addition, some detection tools also use assembly tools to construct new references with the alignment results. FusionQ uses cufflinks, Breakfusion uses TIGRA-SV and Fusioncatcher uses velvet which could improve the true positive rate but requires more time and memory.

As shown in Table 2-1, most of the eight fusion detection tools filter away fusion events with the number of supported spanning/split reads below a certain threshold, with one exception: Breakfusion. Table 2-1 also indicates that some of those tools further use the anchor length filter, where the minimum number of nucleotides in one end of the split reads aligning to the left or right of fusion breakpoints must be greater than a threshold. The other important filtering criterion, which can rule out many false positives due to multiple mapping problems, is the homology based filter. Utilized by Tophat-Fusion, deFuse, EricScript, Fusioncatcher, and FusionQ, it removes fusion events with reads mapped to homologous or repetitive regions, or pseudo genes. In addition, Defuse and EricScript further quantify several important characteristics of fusion genes, such as information regarding supported spanning/split reads, anchor length, fusion boundary di-nucleotide entropy and so on, as input features fed into the AdaBoost classifier to evaluate and improve prediction performance by simultaneously considering true positives and false positives.

There are also other filters considered in those detection tools. For example, the read-through filter removes the molecule formed by adjacent genes and most of the detection tools use this filter except for Breakfusion and FusionQ. The PCR artifact filter only applied by EricScript, removes the duplicated reads from PCR artifacts. Some detection tools, such as Tophat-Fusion and Fusioncatcher, also provide annotation information about known fusion genes to verify the novelty of the identified fusion genes by checking against the existing fusion gene databases.

Table 2-1 A summary of eight fusion detection tools

	Tophat-Fusion (v.2.0.8)	ChimeraScan (v.0.4.5)	deFuse (v.0.6.1)	SOAPfuse (v.1.25)	EricScript (v.0.4.1)	Fusioncatcher (v.0.97)	Breakfusion (v.1.0.1)	FusionQ (v.1)
Anchor length filter	O (20)	O (4)	O (4)	O (5)	X	O (17)	X	X
Read-through transcript filter	O	O	O	O	O	O	X	X
Supported reads filter (spanning/split = 1)	O (2/3)	O (2)	O (5/3)	O (1/1)	O (3/)	O (3/2)	X	O (1/1)
PCR artifact filter	X	X	X	X	O	X	X	X
Homology based filter	O	X	O	X	O	O	X	O
Alignment tools	bowtie	Bowtie/blat	Bowtie/blat	soap2/bwa/blat	bwa/blat	bowtie/blat	bwa/blat	bowtie
Assembly (#) / Machine learning (o)	X	X	O	X	O	#velvet	#TIGRA-SV	#cufflinks
Fusion gene database information	O	X	X	X	X	O	X	X

2.1.3 Existing measures for predicting prostate cancer recurrence

Prostate cancer diagnosis and treatment could be improved with new and effective biomarkers. For example, the ability to distinguish between the indolent prostate cancer and aggressive prostate cancer could avoid overtreatment, especially unnecessary biopsies. A number of biomarkers have been studied; however, only prostate specific antigen (PSA) is routinely used by urologists, which unfortunately often led to overdiagnosis and overtreatment. Some new biomarkers shown to be statistically significant in one study are often not validated by others

(Check, 2004). The only food and drug administration approved tests are PSA 3 and an isoform of proPSA (Romero Otero et al., 2014).

Whether the most prevalent fusion TMPRSS2-ERG could predict prostate cancer recurrence or has risk associated with diseases progression is controversial. One study of 165 patients shows that the expression of TMPRSS2-ERG fusion is a strong prognostic factor and is independent of grade, PSA level and stage (Nam et al., 2007). Another report shows this fusion is significantly associated with prostate cancer specific death with a cohort study of 111 patients (Demichelis et al., 2007) . Some authors show this fusion has no association with prostate cancer outcome (Gopalan et al., 2009).

There are other studies which claim findings of recurrent fusion genes in prostate cancer.

A study on a Chinese population showed two novel frequently occurred fusion genes in their patient cohort (Ren et al., 2012). Other fusion genes with the TMPRSS2 are also reported but they are of individually low frequency (Salagierski and Schalken, 2012). Other types of high-dimensional features such as CpG island methylation could also be used to predict prostate cancer recurrence (Luo et al., 2013; Yu et al., 2013).

There are many challenges of developing biomarkers for prostate cancer such as standardization of the diagnostic test, sample size of the test cohort, quality control, cost and benefit tradeoff and so on. Although the road of biomarker detection is filled with obstacles, the benefits greatly outweigh the costs in terms of patient care and cost to the health care system.

2.2 DESIGN AND HYPOTHESIS

To identify genome markers for PCa, tumor (T), adjacent normal prostate tissue (AT) and peripheral blood samples (B) were obtained from 5 men having prostate cancer with an aggressive clinical course. In one patient, normal adjacent prostate tissue was not available. For whole genome sequencing, an average of 200 GB was sequenced per sample to achieve 33 fold coverage of the entire genome. Total RNA from all T and AT samples was sequenced to achieve >1333 (average 400 million reads/sample) fold coverage per gene. Total RNA from four age-matched, entirely histological benign prostate tissues harvested from organ donors was similarly sequenced as a tissue control. The sequencing data were aligned to human reference genome HG19(Li and Durbin, 2009). Fusion transcripts were then identified and validated. It was our hypothesis that the presence of these fusion transcripts in the primary tumor would be associated with disease recurrence, development of metastatic disease or prostate cancer-specific death. Therefore, the fusion transcripts were analyzed on 90 PCa samples from men with known clinical outcomes and 10 benign prostates harvested at the time of organ donation. A prediction model for PCa recurrence and short post-operative prostate specific antigen doubling time (PSADT) was built. This model was then applied to 89 additional PCa samples from University of Pittsburgh Medical Center, 30 samples from Stanford University Medical Center, and 36 samples from University of Wisconsin Madison Medical Center with follow-up ranging from 1 to 15 years. One hundred twenty-seven of these samples are from patients who experienced PCa recurrence after radical prostatectomy, and 106 are from patients with no evidence of recurrence for at least 5 years after the surgery. The remaining 46 samples are from patients who had less than 5 years of follow-up and had not yet experienced biochemical recurrence. The association of fusion transcript expression with PCa recurrence was analyzed.

2.3 METHODS

2.3.1 Tissue samples

Two hundred and eighteen specimens of PCa, 4 matched AT, 5 matched B and 14 organ donor prostates (OD) were obtained from University of Pittsburgh Tissue Bank in compliance with institutional regulatory guidelines. To ensure high purity (≥80%) of tumor cells, needle-microdissection was performed to isolate the tumor cells from adjacent normal tissues (≥3 mm distance from the tumor). For AT and OD samples, similar needle-microdissections were performed to achieve 80% epithelial purity. Genomic DNA of these tissues was extracted using a commercially available tissue and blood DNA extraction kit (Qiagen, Valencia, CA). The protocols of tissue procurement and procedure were approved by Institution Board of Review of University of Pittsburgh. PCa samples of Stanford University Medical Center cohort and University of Wisconsin Madison Medical Center cohort were obtained from corresponding institutional tissue banks and approved by Institutional Review Boards. Information about PSADT and time to recurrence were not available for Wisconsin cohort.

2.3.2 Whole genome and transcriptome sequencing library preparation and sequencing

To prepare the genomic DNA libraries, 50 ng DNA was subjected to the tagmentation reactions using the NEXTERA DNA sample prep kit (Madison, WI) for 5 min at 55^{o}C. The DNA was then amplified with adaptor and sequencing primers for 9 cycles of the following procedure: 95°C for 10s, 62°C for 30s and 72°C for 3 min. The PCR products were purified with Ampure beads. The quality of genomic DNA libraries was then analyzed with qPCR using Illumina sequencing primers and quantified with Agilent 2000 bioanalyzer. For transcriptome sequencing, total RNA was extracted from prostate samples using Trizol, and treated with DNAse1. Ribosomal RNA was then removed from the samples using RIBO-Zero^{TM} Magnetic kit (Epicentre, Madison, WI). The RNA was reverse-transcribed to cDNA and amplified using TruSeq™ RNA Sample Prep Kit v2 from Illumina, Inc (San Diego, CA). The library preparation process such as adenylation, ligation and amplification was performed following the manual provided by the manufacturer. The quantity and quality of the libraries were assessed as those described in genome DNA library preparation. The procedure of 200 cycle paired-end sequencing in Illumina HiSeq2000 followed the manufacturer's manual.

2.3.3 Read alignment

Whole genome DNA-seq reads from 5 Ts, 4 ATs and 5 Bs were aligned by BWA (Li and Durbin, 2009) version 1.4.1 against the UCSC hg19 human reference genome allowing maximal 2 base mismatches per (100 nucleotide) read. After alignment, the average coverage of whole genome was above 30X for all 14 samples. Picard tool (http://picard.sourceforge.net) was applied to remove duplicate reads after the alignment. RNA-seq reads (from 5 T, 4 matched AT and 4 OD samples) were at an average of 1333X coverage. A maximum of 2 mismatches per read was allowed.

2.3.4 Fusion transcript detection

To identify fusion transcript events, we applied the Fusioncatcher (v0.97) algorithm (Edgren et al.) to the RNA sequencing samples. Embedded in fusioncatcher, BOWTIE and BLAT were used to align sequences to the reference genome. Fusion genes were visualized with CIRCOS software (Wei Zeng, 2013) as shown in Figure S1 in .

2.3.5 Machine learning classifier to predict recurrence status

8 fusion genes from 5 tumor samples validated by RT-PCR, Sanger sequencing and Fluorescence In-situ Hybridization **(**FISH) analyses were used as features to predict non-recurrence vs recurrence and the nature of the recurrence (PSADT<4 months vs PSADT≥15 months or non-recurrent). A prediction model based on the presence of any of these 8 fusion transcripts in the sample was constructed to predict recurrent status, while linear discriminant analysis (LDA) algorithm based on calculated prediction weight of each fusion transcript was used to predict PSADT. Leave-one-out cross validation (LOOCV) was used to assess prediction

performance in both models (UPMC training cohort, Step I in Figure 2-4A). A prediction model of prostate cancer recurrence was first built on a subset of Pittsburgh cohort of 90 samples. The performance of the selected best model was then applied to a separate cohort of 89 samples to assess the accuracy and applicability (Step II in Fig 2-4A). The same model was also applied to predict the Stanford and Wisconsin cohorts (Step III and IV in Fig 2-4A). A similar strategy was used to predict PSADT≤4 months.

2.3.6 RT-PCR and FISH

Double strand cDNA was synthesized as described previously (Luo et al., 2002; Yu et al., 2004). PCR assays were conducted using the following conditions: 94^{o}C for 5 min, followed by 30 cycles of 94^{o}C for 30 seconds, 61^{o}C for 1 min and 72^{o}C for 2 min. The procedure of probe preparation and FISH were described previously (Ren et al., 2006; Yu et al., 2007).

2.4 RESULTS

2.4.1 Fusion transcripts discovered by RNA and whole genome sequencing

To identify fusion transcripts, analysis of RNA sequencing was performed on 5 PCa samples. A total of 76 RNA fusion events were identified using the Fusioncatcher (Edgren et al.) program. Thirteen of these fusion events were confirmed by genome sequencing. To control for tissue-based normal fusion transcript events, fusion transcripts present in any of the four age-matched organ donor prostate tissues were eliminated. Further, fusion transcripts with less than 20 kb between each element and read in the cis direction were also eliminated. As a result of this filtering, 28 of 76 fusion transcript events were identified as PCa specific (Figure S1). Among these fusion events, TMPRSS2-ERG, the most common PCa fusion transcript (Baca et al.; Berger et al.; Tomlins et al., 2005), was found in two PCa samples. The majority of the fusion events identified were novel. No fusion transcripts were identified in any of the AT samples. To validate these fusion transcripts, RT-PCR was performed using primers specific for fusion transcript regions encompassing the fusion breakpoints, and the PCR products were sequenced. The experimental validation was conducted on all novel fusion transcripts and eight out of them were validated through sequencing with four shown in Figure 2-1.

Five of the eight fusion events resulted in truncation of a head gene and frameshift in translation of a tail gene. One of the fusion transcripts produced a truncated cyclin H and an independent open reading frame of a novel protein whose function is not known. Two fusion events, however, are predicted to produce chimeras that possibly retain at least partial function of both genes. For example, a fusion transcript between the N-terminus 703 amino acids of α-Manosidase 2A (MAN2A1) and the C-terminus 250 amino acids of FER, a Feline tyrosine kinase retains the glycoside hydrolase domain of MAN2A1 but replaces the manosidase domain with the tyrosine kinase domain from FER. Another fusion transcript couples 5 of 10 transmembrane domains of the membrane transporter protein SLC45A2 with the methyl-acyl CoA transferase domain from AMACR. Interestingly, both MAN2A1-FER and SLC45A2-AMACR fusions are in the trans-direction, eliminating the possibility of a fusion event from simple chromosome deletion or collapse of extremely large RNA transcript.

The most frequent fusion events observed in PCa were TRMT11-GRIK2 (7.9%, or 22/279) and SLC45A2-AMACR (7.2%, or 20/279) (Figure 2-3). TRMT11-GRIK2 fusion represents a giant truncation of TRMT11, a tRNA methyltransferase, and elimination of GRIK2, a glutamate receptor but reported to possess tumor suppressor activity(Sinclair et al., 2004). Indeed, when GRIK2 expression was examined in 14 TRMT11-GRIK2 positive PCa samples, it was undetectable, while it was detected in organ donor prostate samples. Only 4 of 14 samples with TRMT11-GRIK2 expressed full length TRMT11 transcripts. Thus, the fusion event of TRMT11-GRIK2 likely produces a loss of function.

2.4.2 Fluorescence in situ hybridization suggests genome recombination underlying fusion transcript formation

To investigate the mechanism of these fusion events, fluorescence in situ hybridization (FISH) was performed on PCa tissues where the fusion transcripts were present. Using the probes surrounding the MAN2A1 breakpoint, a physical separation of signals between 5' and 3' MAN2A1 in cancer cells containing the fusion was observed, while the wild type alleles in normal prostate epithelial cells showed overlapping fluorescent signals (Figure 2-2). Similar "break-apart" hybridization occurred in SLC45A2-AMACR positive PCa samples (Figure 2-2 (B)). These findings indicate that MAN2A1-FER and SLC45A2-AMACR fusions are the result of chromosomal recombination events and validate the fusion transcripts found by RNA-seq. Interestingly, in PCa cells containing "break-apart" signals of MAN2A1, only 31% of the cells retained the 3' end signal, suggesting that the recombination event results in truncation of the C-terminus of MAN2A1 in most PCa cells. A similar "collateral loss" of the N-terminus of

AMACR was found in PCa cells expressing the SLC45A2-AMACR fusion transcript (29% retaining the N-terminus signal of AMACR). Other FISH analyses confirmed that genome translocations occur in cancer cells expressing TRMT11-GRIK2, MTOR-TP53BP1, LRRC59-FLJ60017, TMEM137- CCDC67, CCNH-C5orf30 and KDM4B- AC011523.2 fusion transcripts. These fusion transcripts are either separated widely on a single chromosome (TRMT11-GRIK2, TMEM135-CCDC67, CCNH-C5orf30 and KDM4B-AC011523.2) or located on separate chromosomes (MTOR-TP53BP1 and LRRC59-FLJ60017). The overlapping signals of hybridizations in PCa cells offered additional validation of these fusion events. Finally, the genomic breakpoints were identified in 3 fusion pair through Sanger sequencing of the cancer genomic DNA (CCNH-C5orf30, TMEM135-CCDC67 and LRRC59-FLJ60017) (Figure S2).

2.4.3 Fusion transcripts associate with PCa recurrence

To investigate the clinical and biological significance of the fusion transcripts, their presence was assessed in PCa specimens obtained from 213 men and in histologically confirmed benign prostate tissues obtained from 10 organ donors free of urological disease aged 20 to 70. For 179 of the 213 PCa samples, clinical outcome data after radical prostatectomy were available, and 81 had no detectable prostate specific antigen (PSA) recurrence after a minimum of five years of follow-up, while 98 developed biochemical recurrence (defined as a measurable PSA ≥0.2 ng/ml). In the patients without recurrence, only 7.4% (6/81) primary prostate cancers expressed one of the fusion transcripts. In contrast, 52% (51/98) primary prostate cancers who developed biochemical recurrence expressed at least one fusion (Figure 2-3). No fusion transcripts were detected in benign prostate tissues obtained from healthy organ donors. Three fusion events were observed exclusively in recurrent PCa after radical prostatectomy (TRMT11-GRIK2, MTOR-TP53BP1 and LRRC59-FLJ60017).

Fisher's exact test showed a significant difference in recurrence status among patients with at least one of the 8 fusion transcripts and those without ($p=2.9\ x10^{-11}$). A simple classification rule was used so that a sample with the presence of top (N) fusion genes will be predicted to have PCa recurrence. N from 1 to 8 was attempted and feature selection was based on p values from Fisher's exact test. The model selection was conducted on Pittsburgh training cohort, a randomly selected 90-sample training cohort from UPMC. N=8 was found to perform best in this training cohort, which yielded an accuracy of PCa recurrence prediction of 71% with 89% specificity and 58% sensitivity ($p<0.005$). When this model was applied to a separate cohort of 89 samples (test set), the model correctly predicted recurrence in 70% of patients. To further validate this model, we tested its performance in a 30-patient cohort from Stanford University Medical Center and a 36-patient cohort from University of Wisconsin Madison Medical Center. Once again, the model correctly predicted recurrence with 76.2% accuracy and with 89% specificity and 67% sensitivity on the PCa cohort from Stanford, and 79% accuracy and with 100% specificity and 59% sensitivity on the cohort from Wisconsin. Interestingly, when fusion transcript status was combined with Gleason Grade$\geq$8, a mild improvement of prediction was found for all 4 cohorts: 72% for the training cohort, 73% for the test cohort, 81% for the Stanford cohort and 85% for the Wisconsin cohort. Prostate cancer recurrence, metastases and/or prostate cancer-specific death after radical prostatectomy occurred in 91% (69/76) of men carrying at least one of the eight fusion genes, while these outcomes occurred in only 37% (58/157) of men not carrying those fusion transcripts.

In itself, recurrence does not signal an aggressive prostate cancer, since many patients with PSA recurrence do not develop metastases or die from their disease. A PSA doubling time (PSADT) less than four months after radical prostatectomy is strongly associated with the early development of metastatic disease and prostate cancer-specific death, whereas these events are rare and remote in men with a PSADT of greater than 15 months (Antonarakis et al.; Freedland et al., 2007). The presence of one or more fusion transcript in the PCa tissue showed a strong association with PSADT less than four months ($p=6 \times 10^{-9}$). To examine whether these fusion transcripts have prognostic value for PCa clinical outcome, a prediction model was built with linear discriminant analysis (LDA). A combination of feature selection from 1 feature to 8 features and the prior probability of LDA from 0.05 to 0.95 with 0.05 spacing were attempted. Feature selection was conducted based on p value ranking as before from Fisher's exact test and the best model was with eight features with prior probability 0.6. The panel of eight fusion transcripts correctly predicted 74.4% for PSA doubling time in the 90-sample training cohort (blue dot). When the same algorithm was applied to a separate 89-sample test set from University of Pittsburgh Medical Center and 21-sample cohort from Stanford University Medical Center, the prediction rate for PSADT$\leq$4 months was found to be 78% and 71%, respectively. To examine the impact of fusion transcripts on patients' PSA free survival, a Kaplan-Meier analysis was performed on the PCa cohort from University of Pittsburgh. As shown in figure 2-4 (C), 84.2% of patients had an observed disease recurrence within five years of radical prostatectomy if they carried any of the 8 fusion transcripts. No patient survived five years without recurrence if their primary PCa contained a TRMT11-GRIK2, or MTOR-TP53BP1 transcript fusion. In contrast, 68% patients were free of disease recurrence if none of the fusion transcripts were detected in their primary PCa. Similar findings were also identified in the Stanford cohort: 88.9%

patients experienced recurrence of PCa if they carried any fusion transcript, while 75% patients were free of the disease recurrence if they are negative.

2.5 DISCUSSION

Transcriptome sequencing revealed numerous fusion RNA transcripts occurring not just in PCa but also in benign adjacent tissues and histologically normal organ donor prostate samples. Whether these transcripts encode functional fusion proteins is not known. Deep sequencing of the whole genome and RNA revealed a significant number of prostate cancer-specific fusion transcripts. These fusions are not detectable in either benign organ donor prostate or benign prostate tissues from PCa patients. Most of these fusion transcripts appear to express in low abundance, with only an average 6.6 reads of these fusion transcripts detected in >1333x sequencing. Indeed, when the coverage was reduced to 600x in simulation studies, only MTOR-TP53BP1 was detected consistently. One general characteristic of these fusion transcripts is that they either have a large distance between the gene targets or they are oriented in trans, features that could only occur as a result of chromosome recombination events. In either scenario, the fusions must be the product of significant structural DNA rearrangements.

Although the association between the eight novel fusion transcripts and prostate cancer recurrence is striking, the biological roles of these fusion transcripts are not yet elucidated. Given the known function of the genes contributing to the fusion transcripts, their formation may have impact on several cell pathways such as RNA stability (Towns and Begley, 2012) (TRMT11-GRIK2), protein glycosylation (Misago et al., 1995) (MAN2A1-FER), cell cycle progression(Fisher and Morgan, 1994) (Wang et al., 2012; Yang et al., 2013) (CCNH-C5orf50and MTOR-TP53BP1), fibroblast growth factor nuclear import (Zhen et al., 2012) (LRRC59-FLJ50017), histone demethylation (Yang et al.) (KDM4B-AC011523.2), and fatty acid metabolism (Savolainen et al., 2004) (SLC45A2-AMACR). Many of these pathways appear to be fundamental to cell growth and survival. Even though the prevalence of each fusion transcript in prostate cancer samples is low (ranging from 2.9% to 7.9%), up to 60% of prostate cancers that later recurred and had short PSADT were positive for at least one of these fusion transcripts. The specificity of these fusion transcripts in predicting prostate cancer recurrence appears remarkably high, ranging from 89-100% among 4 separate prediction cohorts. There were no long term recurrence-free survivors if the primary tumor contained either TRMT11-GRIK2, MTOR-TP53BP1 or LRRC59-FLJ60017 fusion transcripts.

To our knowledge, it is the first study showing that a set of fusion transcripts is strongly associated with PCa prognosis. If these fusion transcripts are found to have a biological role in prostate cancer progression, it may provide new targets for therapeutic intervention.

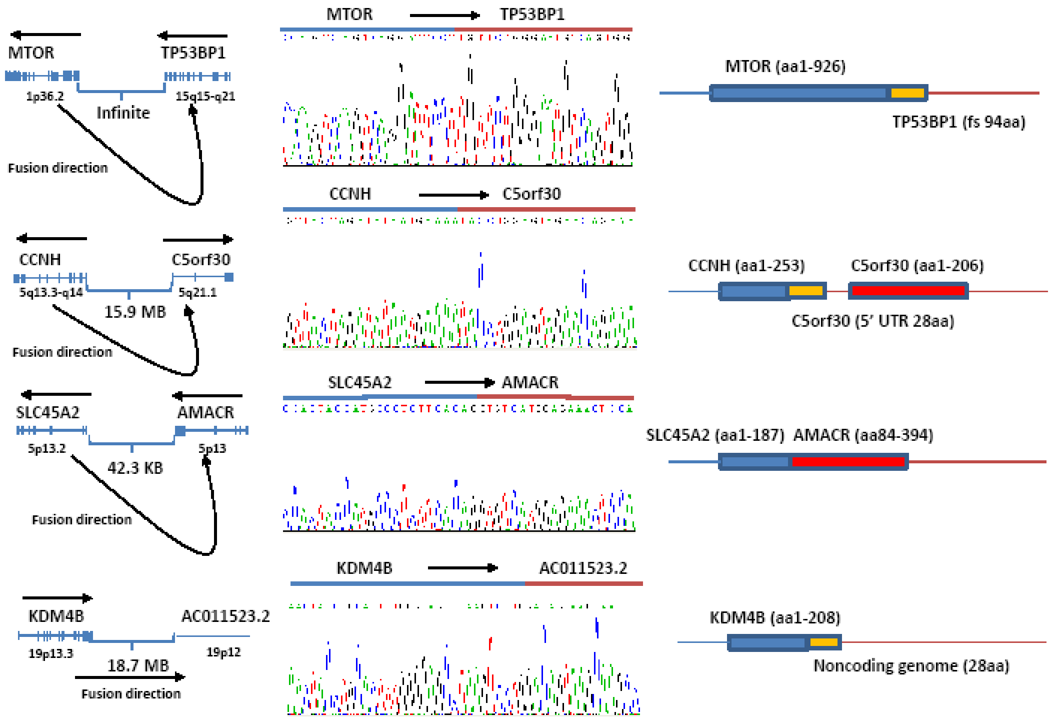

Figure 2-1 Unique fusion gene events.

Left panel: Miniature diagrams of genome of the fusion genes, the transcription directions, the distances between the joining genes and directions of the fusions. Middle panel: Representative sequencing chromograms of fusion transcripts. The joining gene sequences were indicated. Right panel: Diagrams of translation products of fusion transcripts. Blue-head gene translation product; Red-tail gene translation product; Orange-novel translation products due to framshift or translation products from a non-gene region.

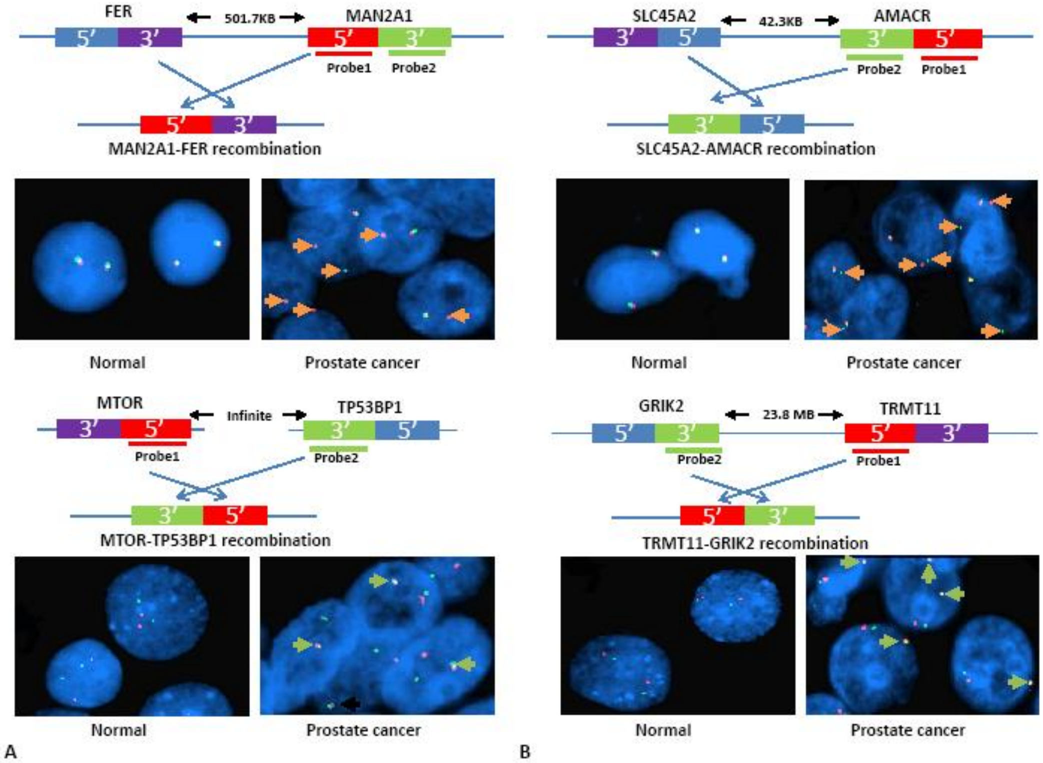

Figure 2-2 Fluorescence in situ hybridization suggests genome recombination in prostate cancer cells.

(A) Schematic diagram of MAN2A1 and FER genome recombination and FISH probe positions. Representative FISH images were shown for normal prostate epithelial cells and cancer cells positive for MAN2A1-FER fusion. Orange denotes probe 1; Green denotes probe 2. Break-apart signals are indicated by orange arrows. (B) Schematic diagram of SLC45A2 and AMACR genome recombination and FISH probe positions. Representative FISH images were shown for normal prostate epithelial cells and cancer cells positive for SLC45A2-AMACR fusion. Orange denotes probe 1; Green denotes probe 2. Break-apart signals are indicated by orange arrows. (C) Schematic diagram of MTOR and TP53BP1 genome recombination and FISH probe positions.

Representative FISH images were shown for normal prostate epithelial cells and cancer cells positive for MTOR-TP53BP1 fusion. Orange denotes probe 1; Green denotes probe 2. Fusion joining signals are indicated by green arrows. (D) Schematic diagram of TRMT11 and GRIK2 genome recombination and FISH probe positions. Representative FISH images were shown for normal prostate epithelial cells and cancer cells positive for TRMT11-GRIK2 fusion. Orange denotes probe 1; Green denotes probe 2. Fusion joining signals are indicated by green arrows.

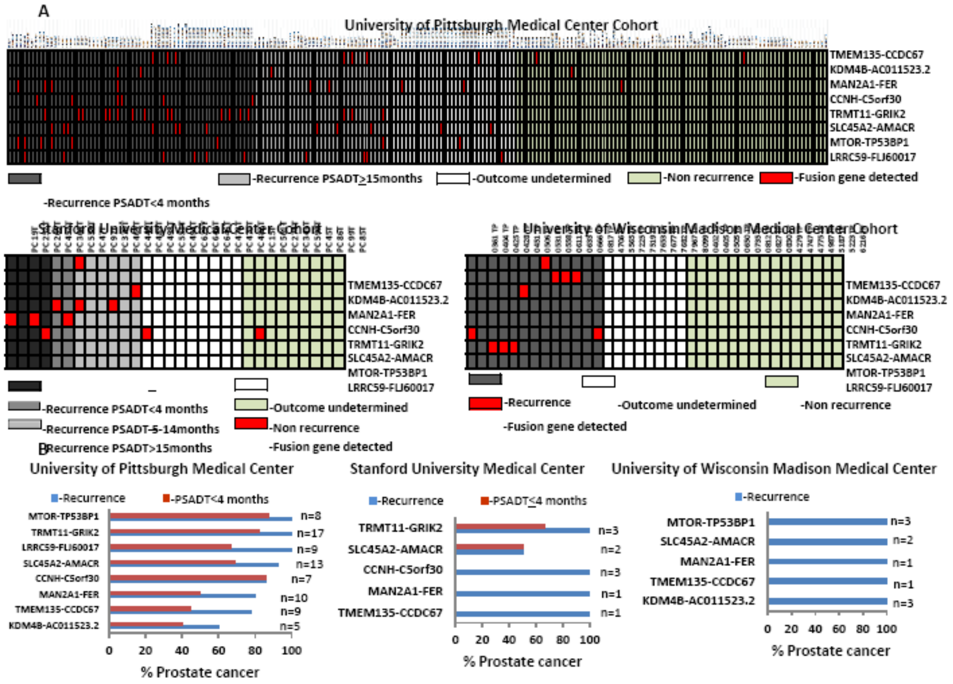

Figure 2-3 Fusion transcripts expressed in prostate cancer.

(A) Distribution of 8 indicated fusion transcripts in 213 prostate cancer samples from University of Pittsburgh Medical Center, in 30 samples from Stanford University Medical Center and in 36 samples from University of Wisconsin Madison Medical Center. Samples from patients who experienced recurrence were indicated with light grey (PSADT≥15 months) or dark grey

(PSADT<4 months) or intermediate grey (PSADT=5-14 months), samples from patients who have no recurrence for at least 5 years with green, and samples from patients whose clinical follow-up is ongoing but less than 5 years with white (undetermined). (B) Correlation of fusion transcript events with prostate cancer recurrence. Percentage of prostate cancer experiencing recurrence from samples positive for fusion transcripts was plotted for each fusion transcript. Left, University of Pittsburgh Medical Center cohort; Middle, Stanford University Medical Center cohort; Right, University of Wisconsin Madison Medical Center cohort.

Figure 4

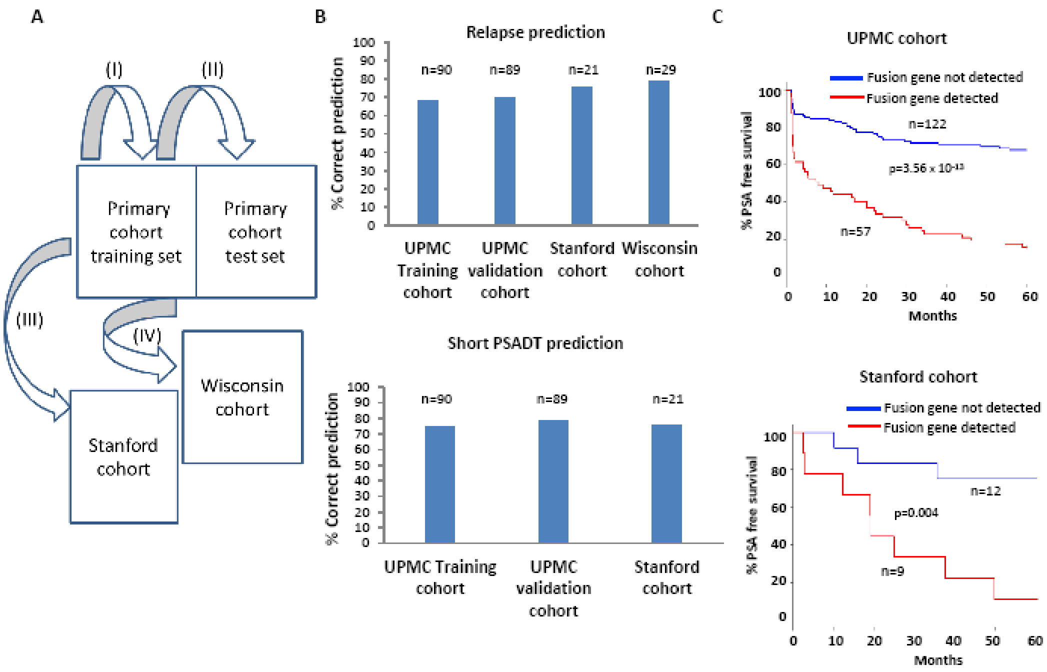

Figure 2-4 Fusion genes predict recurrence of prostate cancer.

(A) Schema of training and validation steps in building fusion gene prediction models for prostate cancer recurrence and short PSADT. The algorithm of fusion gene prediction of prostate cancer recurrence and PSADT<4 months was obtained from 90 randomly-assigned prostate cancer samples from the University of Pittsburgh Medical Center. The algorithm was then applied to 89 samples from the University of Pittsburgh Medical Center, 21 samples from Stanford University Medical center and 33 samples from the University of Wisconsin Madison Medical Center. (B) Sample sizes of cohorts of University of Pittsburgh Medical Center, Stanford Medical Center, and University of Wisconsin Madison Medical Center involved in the prediction of prostate cancer recurrence as well as PSADT < 4 months. (C) Kaplan-Meier analysis of patients who were positive for any of TRMT11-GRIK2, SLC45A2-AMACR, MTOR-TP53BP1, LRRC59-FLJ60017, TMEM135 –CCDC67 and CCNH-C5orf30 versus those who were negative for these fusion events. Top, Kaplan-Meier analysis of prostate cancer sample cohort from University of Pittsburgh; Bottom, Kaplan-Meier analysis of prostate cancer sample cohort from Stanford University Medical Center.

3.0 BIAS CORRECTION FOR SELECTING THE MINIMAL-ERROR CLASSIFIERS FROM MANY MACHINE LEARNING MODELS

Ying Ding[1,2], Shaowu Tang[2], Serena G. Liao[2], Jia Jia[2], Steffi Oesterreich[3], Yan Lin[2], George C. Tseng[1,2*]

[1] Joint Carnegie Mellon-University of Pittsburgh Ph.D. program in Computational Biology, Pittsburgh, PA, USA

[2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

[3]Magee-Womens Research Institute, Pittsburgh, PA, USA

[*] Corresponding author

Acknowledgement

Funding: This study is supported by NIH R21MH094862.

Preface

Note on the contribution of YD to this paper: he is the sole author of all statistical analysis, simulations, methods, figures and text. This work is under revision in Bioinformatics.

Abstract

Motivation: Supervised machine learning is commonly applied in genomic research to construct a classifier from the training data that is generalizable to predict independent testing data. When test datasets are not available, cross-validation is commonly used to estimate the error rate. Many machine learning methods are available and it is well-known that no universally best method exists in general. It has been a common practice to apply many machine learning methods and report the method that produces the smallest cross-validation error rate. Theoretically, such a procedure produces a negative selection bias. As a result, many clinical studies with moderate sample sizes (e.g. N=30~60) risk to report a falsely small cross-validation error rate, which could not be validated later in independent cohorts.

Results: In this study, we illustrated the probabilistic framework of the problem and explored the statistical and asymptotic properties. We examined three existing methods, nested cross validation (nestedCV) weight mean correction (WMC and WMCS), and Tibshirani' procedure. We propose a new bias correction method based on learning curve fitting by inverse power law (IPL). All the methods were applied to simulation datasets, five real datasets as well as two large breast cancer datasets. The result showed IPL gave a competitive estimate with smaller variance for estimating the unconditional error rate and its additional advantage to extrapolate error estimates for larger sample sizes, a practical feature to recommend whether more samples should be recruited to improve the prediction model and accuracy. A R package "MLbias" and all source files are publicly available.

Availability: tsenglab.biostat.pitt.edu/software.htm

3.1 INTRODUCTION

In the past two decades, fast development in bioinformatics was accompanied by the rapid production of high-throughput genomic data, such as gene expression, genotyping and various types of next generation sequencing data. Such high dimensional data usually come with small sample sizes and a large number of genes/features (a.k.a. "large p, small n" problem) and pose many new challenges in statistical learning and data mining. In the content below, we focus on machine learning of gene expression profile data, but the concept and theoretical issues also apply to other high-throughput genomic (e.g. copy number variation, DNA methylation) or proteomic data. In gene expression profile analysis, it is of great interest to predict or diagnose a disease status (e.g. classify cases versus controls or treatment responders versus non-responders). Since no universally best machine learning method exists in general (Allison et al., 2006), to fulfill this task, multiple models are often constructed with different combinations of features (genes), different machine learning methods as well as different tuning parameters in the methods. To choose among such large number of classifiers (models), it is a common practice to select the model with the smallest cross validation error rate, named as the minimal error classifier (MEC), and report its associated error rate.

The MEC error rate is, however, generally downward biased and an overly optimistic estimator of the true optimal classification error rate. This is because taking the minimum of cross validation error rates, where the estimates are random variables, will inevitably yield a downward bias. Such a selection bias has great adverse impact in many biomedical pilot studies with moderate sample sizes (e.g. N=30~60). The problem, however, has often been overlooked in applications. For example, one can examine the small pilot data using ~10 popular machine learning methods and simultaneously choose among many different numbers of features and tuning parameters in each method. This easily increases the number of tested classifiers to several hundreds and selects the MEC classifier with a falsely small error rate due to the selection bias. When the model proceeds to a large cohort validation for translational research, it will likely fail. In the METABRIC example that will be demonstrated in Section 3.3., we will show that the MEC bias can mistakenly reduce the error rate from 28.2% to an overly optimistic 19.1% in early and late stage classification (i.e. a -9.11 % error rate bias). Many researchers have recognized this problem (Bernau, Augustin and Boulesteix, 2013; Berrar, Bradbury and Dubitzky, 2006; Efron, 2009; Fu, Carroll and Wang, 2005; Tibshirani, 2009; Varma and Simon, 2006; Wood, Visscher and Mengersen, 2007). Dupuy and Simon recommended to "report the estimates for all the classification algorithms not just the minimal error rate" (Dupuy and Simon, 2007). Some proposed to compare the minimal error rate with the median error rate from the original data sets with permuted class labels (Boulesteix and Strobl, 2009). These suggestions, however, did not provide a real solution. Yousefi et al. provides a careful probabilistic analysis of this "multiple-rule bias" and quantitative demonstrate the large degree of overoptimism when sample size is small and recommends the problem could be mitigated by averaging performance of multiple datasets (Yousefi, Hua and Dougherty, 2011). Three existing methods have been proposed in the literature, the Tibshrani's procedure (Tibshirani, 2009) the nested cross validation (nestedCV) (Varma and Simon, 2006) procedure and the weighted mean correction (WMC/WMCS) (Bernau et al., 2013). Tibshirani proposed a simple bias estimation method which is computationally efficient and could be calculated through a traditional K-fold cross validation. They claimed that the bias is only an issue when $p \gg n$ where p is the number of genes and n is the number of samples. The nestedCV, proposed by Varma and Simon, introduced another outer loop of cross validation so that the model selection stage is wrapped in the training samples of the outer loop. This double loop procedure, which amounts to nested double leave-one-out cross validation, is computationally expensive with complexity of $O(N^2)$. WMC/WMCS was proposed as a smooth analytical alternative to nestedCV based on subsampling, which yields a competitive estimates compared to nestedCV at a much lower computational price. Theorectically Tibshirani's method estimates the conditional error rate while nestedCV and WMC/WMCS are estimating unconditional error rate (see detailed discussion in (Bernau et al., 2013). We are putting these methods together because of the purpose of this study is to show what should be reported given a fixed dataset with multiple machine learning methods in hand.

In this paper, we first illustrate the MEC bias by a 2-D toy example and discuss its asymptotic theory and statistical properties. The performance of the nestedCV and , WMC/WMCS and Tibshirani's procedure will be examined. A subsampling-based inverse power law (IPL) method will be proposed for the bias correction and compared to the other three existing methods in both simulated and real datasets. In real data evaluation, we will use five GEO datasets and two large breast cancer datasets (TCGA and METABRIC).

3.2 METHODS

3.2.1 Problem setting and formulation under simulation scheme

Assume that an observed dataset D with sample size n and number of features p is to be analyzed for machine learning. Assume that D is generated from an underlying data distribution Δ_n. $M \geq 2$ classification methods are used to learn a good classifier for future prediction. M can be very large (e.g. several hundred) since different feature selection or different parameter settings under a machine learning method are considered different classifiers. Suppose the unknown true error rate of classification method m for data distribution Δ_n is $P_{n,m}$. The theoretical best machine learning method for data distribution Δ_n is $m^* = argmin_m P_{n,m}$ and the resulting error rate is $P_{n,M}^* = min_{1\leq m\leq M} P_{n,m}$ (Box A of Figure 3-1).

We will illustrate the problem in a simulation framework in the remaining of Figure 3-1 since the underlying truth and error rates can be estimated well from repeated simulations. Suppose B datasets $D_b (1 \leq b \leq B)$ are independently generated from Δ_n. We use cross-validated error rate (from leave-one-out cross validation) to approximate the error rate for D_b of sample size n using method m, denoted as $\hat{P}_{n,m,b}$. (i.e. $E(\hat{P}_{n,m,b}) = P_{n,m}$), we have $P_{n,m} = \lim_{B\to\infty} \hat{P}_{n,m}(B)$, where $\hat{P}_{n,m}(B) = (\sum_{b=1}^{B} \hat{P}_{n,m,b})/B$. Denote by $\hat{P}_{n,M}^*(B) = min_{1\leq m\leq M} \hat{P}_{n,m}(B)$. We will show later that $P_{n,M}^* = min_{1\leq m\leq M} P_{n,m} = \lim_{B\to\infty} \hat{P}_{n,M}^*(B)$ (see Box B of Figure 1). In other words, the true best classifier error rate $P_{n,M}^*$ can be estimated by $\hat{P}_{n,M}^*(B)$ when many data sets (i.e. B is large) can be repeatedly simulated from Δ_n. In real data analysis, such repeated simulation is, however, not possible. When a single simulated dataset Db is given, the MEC is chosen by the minimal error rate: $\widetilde{m}_b^{(MEC)} = argmin_{1\le m\le M}\hat{P}_{n,m,b}$ and $\tilde{P}_{n,M,b} = min_{1\le m\le M}\hat{P}_{n,m,b}$. The expected value of the MEC error rate can be estimated as $E\left(\tilde{P}_{n,M,b}\right) = \lim_{B\to\infty} \tilde{P}_{n,M}(B)$, where $\tilde{P}_{n,M}(\mathrm{B}) = (\sum_{b=1}^{B} \tilde{P}_{n,M,b})/B$ (Box C in Figure 3-1). Finally, the bias of MEC classifier error rate can be estimated as $\mathrm{b_{n,M}} = E\left(\tilde{P}_{n,M,b}\right) - P_{n,M}^{*} = \lim_{B\to\infty} \tilde{P}_{n,M}(B) - \hat{P}_{n,M}^{*}(\mathrm{B})$, where the first term is the estimated expectation of MEC error rate and the second term is the estimated true best classifier error rate. In Section 3.2.2, a 2D toy example will be used to demonstrate the issue and properties of the MEC bias $\mathrm{b_{n,M}}$. In Section 3.2.3, we will show that $\mathrm{E}\left(\widetilde{\mathrm{P}}_{\mathrm{n,M,b}}^{*}\right) < \mathrm{P}_{\mathrm{n,M}}^{*}$ is always true and the MEC error rate is always downward (optimistically) biased (i.e. $\mathrm{b_{n,M}} < 0$).

3.2.2 Illustration by a 2-D toy model

Below we present the problem in a 2D toy simulation model. Although the simple simulation model is not intended to mimic a real gene expression profile setting, it illustrates the MEC bias issue with known underlying truth. We simulate $\mathrm{B} = 1000$ training sets D_1 ,…, D_{1000}, from Δ_n where Δ_n contains n = (20, 30, 40, 60, 80, 120, 160, 320, 640, 1280, 2560) data points in a two-dimensional Euclidean space. Data points from two equal-size classes are simulated, one (with n/2 data points) from $\mathrm{N}(\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}3 & 1\\1 & 1\end{pmatrix})$, and the other from $\mathrm{N}(\begin{pmatrix}0\\-2\end{pmatrix},\begin{pmatrix}1 & 1\\1 & 3\end{pmatrix})$. M=10 classifiers are applied to each dataset: k-nearest-neighbors (KNN) with k = 1, 3, 5, diagonal linear discriminant analysis (DLDA), quadratic discriminant analysis (QDA), shrunken centroids discriminant analysis (SCDA), support vector machines (SVM-linear) with linear kernel, support vector machines (SVM-nonlinear) with radial basis kernel, random forest (RF), neural networks with one hidden layer (NNET). The R package CMA (Slawski, Daumer and Boulesteix, 2008) is applied to implement all the classification methods. Given the Gaussian assumptions and non-identical covariance matrixes in two classes, QDA is expected to be the optimal Bayes classifier and the best decision boundary is of quadratic form. Figure 2 left panel shows the averaged error rate for classification method m at sample size n (i.e. $\hat{P}_{n,m}(B)$) estimated from B=1000 independently simulated datasets. As expected, QDA (solid line) is the optimal Bayes classifier and has the lowest error rates for all different n. However, in real data analysis we are only given one observed data set. Figure 3-2 right panel shows the probability of the methods chosen as the minimal-error classifier (MEC) (i.e. $\tilde{m}_b^{(MEC)}$) for different n. When the sample size is small, MEC may not necessarily select the best method QDA. Particularly, when n=20 and 30, KNN methods (K=1 with dashed line and K=3 with dotted line) generate a smaller error rate than QDA with higher probability. When the sample size becomes large (n>160), the data set contains enough information for QDA to be dominantly (greater than 50% probability) selected by MEC. In the Figure 3-3 (a), the true minimal error rate $\hat{P}^*_{n,M}(B) \cong P^*_{n,M}$ from QDA (cross) and expectation of MEC error rate $\tilde{P}_{n,M}(\mathrm{B}) \cong E(\tilde{P}_{n,M,b})$ (circle) are shown for different n. In Figure 3-3 (b), the estimated MEC biases $\hat{\mathrm{b}}_{\mathrm{n,M}}(\mathrm{B}) = \tilde{P}_{n,M}(\mathrm{B}) - \hat{P}^*_{n,M}(B)$ are shown for different n. The result clearly demonstrates a downward bias of MEC error rate and the bias is greater for small sample sizes and diminishes to 0 when sample size is large. It is notable that the bias can be up to 3-5% for n=20-30. Figure 3-3(c) shows the MEC bias for different n when the number of classifiers examined reduce from ten methods (circle) to four methods (QDA, LDA, SVM-linear and SVM-nonlinear) (triangle) or two methods (QDA and LDA) (cross). The result shows that the bias increases as more methods were compared. To our knowledge, few studies have recognized the increasing trend of MEC bias magnitude when sample size is small or when the searching space of machine learning methods is large. For example, Figure 2 in (Boulesteix and Strobl, 2009) indirectly showed that bias increases with decreasing sample size and it showed a counter-balance between the optimization bias and generally decrease of accuracies among all classifiers when sample size is small, which results in a fluctuating minimal error rate when sample size is small as confirmed in Figure 3-3(a) here.

To further study the relationship between the bias and number of classifiers employed, we fix n=20, started with employing the classifier of KNN with k=1 and keep adding one classifier at a time until all the ten classifiers were used. The true optimal classification error rates $\widehat{P}^*_{n,M}(B)$ (cross) and MEC error rate $\widetilde{P}_{n,M}(B)$(circle) (Figure 3-3(d)) and the MEC error rate biases (Figure 3-3(e)) for different number of classifiers are shown. The result shows that the bias increased from ~2% for 2 classifiers (KNN K=1 and QDA) to ~5% for 10 classifiers, while the true optimal error rate $\widehat{P}^*_{n,M}(B)$ (cross) no longer decreases once QDA is added as the 2nd classifier. This result indicates that including additional low-performing classifiers inflates the bias to the extent which may not be compensated by the decrease of the true best classification error rate. In other words, examining too many classifiers and choosing the best is not a good practice, especially if the added classifiers are likely not the top performers. Therefore, caution is called in the small sample size regime when reporting minimal error rate from multiple classifiers in practice and it is advantageous if the optimal classifiers can be applied as early as possible without adding more low-performing classifiers. The theorems in the next subsection show that the increasing magnitude of bias for small sample sizes or large numbers of examined classifiers are common statistical properties in data analysis.

3.2.3 Properties and asymptotic theorems of MEC bias

The proofs of the following theorems are included in Appendix A:

Theorem 1: Given a smaller set of classifiers, adding more classifiers will decrease the true best classification error rate.

Theorem 2: For a given observed dataset $D^{(n)}$ from Δ_n, $\mathrm{E}\left(\widetilde{\mathrm{P}}_{\mathrm{n,M}}^*\right) < \mathrm{P}_{\mathrm{n,M}}^*$, , where $\widetilde{\mathrm{P}}_{\mathrm{n,M}}^* = \min_{1\leq m\leq M}\widehat{\mathrm{P}}_{\mathrm{n,m}}$ and $\widehat{\mathrm{P}}_{\mathrm{n,m}}$ is the cross-validation error rate of $D^{(n)}$ using classifier m. In other words, the bias of MEC error rate $\mathrm{b}_{\mathrm{n,M}} = \mathrm{E}\left(\widetilde{\mathrm{P}}_{\mathrm{n,M}}^*\right) - \mathrm{P}_{\mathrm{n,M}}^*$ is strictly less than 0.

Theorem 3: For observed datasets $D^{(n)}$ from Δ_n of varying n and a fixed number of classifiers$\mathrm{M} \geq 2$, it holds that $\lim_{n\to\infty} \widetilde{\mathrm{P}}_{\mathrm{n,M}}^* = \lim_{n\to\infty}\mathrm{P}_{\mathrm{n,M}}^*$. In other words, $\mathrm{b}_{\mathrm{n,M}} \to 0$ as $\mathrm{n} \to \infty$ for fixed M. Theorem 1 shows that when we have two sets of classifiers and the smaller set is a subset of the larger set, the larger set classifiers will yield a smaller true best classification error rate. Theorem 2 shows that the expected MEC error rate $\mathrm{E}\left(\widetilde{\mathrm{P}}_{\mathrm{n,M}}^*\right)$ always underestimates the true minimal error rate $\mathrm{P}_{\mathrm{n,M}}^*$ and the negative bias always strictly holds. This is consistent with the result in Figure 3-3. In Theorem 3, when the number of classifiers M is fixed, the bias diminishes to 0 as the sample size n increases to infinity. This is also consistent with the 10-classifier result in Figure 3-3(b) where the bias diminishes to around 0 when n is beyond 320. According to Figure 3-3(d), we observe that although the true best classification error rate does not decrease after the QDA is applied; the MEC bias estimate continued to decrease as more classifiers are included. This theoretical result brings clear caution to use MEC without bias correction. In other words, if a researcher runs n=20-30 samples of pilot study and examines M=300 classifiers via conventional cross-validation to choose the best, the minimal error rate from the 300 classifiers will likely generate low (or almost zero) error rate while the underlying true error rate may stay high. The researcher may be misled to expand the study to a larger cohort or a prospective clinical trial and eventually find it difficult to validate the model and cannot translate into a clinically useful diagnostic tool.

3.2.4 Three existing bias correction methods

In the literature, several methods have been developed to correct the downward bias of MEC error rate. Below we describe three (Bernau et al., 2013; Tibshirani, 2009; Varma and Simon, 2006). Bernau, et.al did assess the condition with multiple machine learning methods while the rest two focused on correcting the bias of parameter tuning via cross validation (e.g. estimate K for KNN) for a given machine learning method (Bernau et al., 2013). In practice, if one considers many machine learning models along with feature selection and parameter tuning, the number of classification methods (M) examined can easily reach several hundreds. All the three methods considered here can be generalized to this condition.

Below we describe existing methods that can be applied for MEC error rate bias correction.

Nested cross validation (nestedCV): Instead of using a single loop cross validation to find the minimal error estimate for a particular classifier, nestedCV utilizes two CV loops. The dataset is initially divided into training and testing sets. Then leave-one-out cross validation

(LOOCV) is applied on the training set using all the classifiers and the classifier with the smallest error rate is selected and used to build the model based on the training set and then evaluate the error rate on the testing set in the end. Therefore, the testing set is independent is independent of the model selection stage, including the selection of MEC. Finally, the process is repeated until each sample acts as testing set once, thus it is a double LOOCV with two CV loops. The computation therefore scales with the square of the sample size. Instead of LOOCV, it is possible to use 5 fold or 10 fold cross validation to accelerate the computing when sample size is large.

Weighted mean correction (WMC/WMCS): It is proposed to be a smooth analytical alternative to nestedCV, which is a weighted mean of the resampling error rates, obtained using the different machine learning models/parameter values. It estimates the unconditional error rate as $\widehat{Err}_{WMC} = \sum_{k=1}^{K} \hat{P}\,(k^*(S) = k)e(k||S)$ where $e(k||S)$ stands for the average error rate for method k from subsampling and $P(k^*(S) = k)$ is the weight for classifier k. It was shown in a prior study to give comparable estimate, more stable and with a much lower computational price. WMC and WMCS are two variants in terms of estimating $P(k^*(S) = k)$.

Tibshirani's procedure (BC): Tibshirani's procedure applies the idea of estimating the bias and adding back to the minimal error rate estimate to correct for the bias in the setting of K-fold cross validation. It estimates the true best classification error rate as $2\mathrm{CV}(\hat{\theta}) - \frac{1}{\mathrm{K}}\sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{e}_{\mathrm{k}}(\hat{\theta}_{\mathrm{k}})$ where $\mathrm{CV}(\hat{\theta})$ is the biased MEC error rate and $\mathrm{e}_{\mathrm{k}}(\hat{\theta}_{\mathrm{k}})$ is the minimal error rate in the kth fold among all classifiers (Tibshirani, 2009). It does not require a significant amount of additional computation as in nestedCV and scales linearly with the number of cross validation folds. Due to the calculation of $\mathrm{e}_{\mathrm{k}}(\hat{\theta}_{\mathrm{k}})$, Tibshirani's method is not suitable for leave-one-out cross validation (LOOCV) or when the size of left-out test set is too small and this estimate was shown to over-estimate the bias in some settings (Christoph Bernau, 2011).

3.2.5 The resampling-based inverse power law (IPL) method

All the above three methods estimate the error rates based on a smaller sample size other than n. The Tibshirani's method generally uses a small number of folds K. NestedCV relies on inner loop with a smaller sample size.WMC/WMCS estimates based on subsamples. In order to estimate at the sample size n as well as give insights on estimates beyond n, we propose a new resampling-based inverse power law (IPL) method to correct the MEC error rate bias and estimate the true optimal classification error rate $\mathrm{P}_{\mathrm{n,M}}^{*}$. By constructing learning curves for each individual classifier from repeated resampling of the original dataset at different subsample sizes (Mukherjee et al., 2003), we could estimate the error rate of each classifier by fitting a learning curve. Consider sample sizes $1 \leq n_1 < n_2 < \ldots < n_L < n$. For a given machine learning method m, assume that the true error rate equals $\mathrm{P}_{\mathrm{n_l,m}}$ and these true error rates follow an inverse power law function: $\mathrm{P}_{\mathrm{n_l,m}} = \mathrm{a_m n_l^{-\alpha_m}} + \mathrm{b_m}$. Normally we assume $\mathrm{a_m}, \mathrm{b_m}, \alpha_\mathrm{m} > 0$ since theoretically larger sample size contains more information to produce lower prediction error rate. To estimate$\mathrm{a_m}$, $\mathrm{b_m}$ and α_m, we first estimate the underlying $\mathrm{P}_{n_l,\mathrm{m}}$ from subsampling n_l samples from the whole data and repeat for B times. The resulting observed cross-validated error rate of each of the sub-sampled data is denoted by $\widehat{\mathrm{P}}_{\mathrm{n_l,m,b}}^{\mathrm{sub}}$ $(1 \leq \mathrm{b} \leq \mathrm{B})$ and the averaged error rate is $\widehat{\mathrm{P}}_{\mathrm{n_l,m}}^{\mathrm{sub}} = (\sum_{\mathrm{b}=1}^{\mathrm{B}} \widehat{\mathrm{P}}_{\mathrm{n_l,m,b}}^{\mathrm{sub}})/\mathrm{B}$. The least squared error (LSE) method is then used to estimate $\mathrm{a_m}$, $\mathrm{b_m}$ and α_m.

$$(\hat{a}_m, \hat{b}_m, \hat{\alpha}_m) = \arg\ \min_{a_m, b_m, \alpha_m} \sum_{l=1}^{L} (\hat{P}_{n_l,m}^{sub} - a_m n_l^{-\alpha_m} - b_m)^2$$
$$\text{s.t. } a_m, b_m, \alpha_m \geq 0$$

The inverse power law has been found to fit well in simulation and many real data sets (Mukherjee et al., 2003). It has the advantage to obtain an accurate estimate of $\hat{P}_{n,m}^{IPL} = \hat{a}_m n^{-\hat{\alpha}_m} + \hat{b}_m$ for $P_{n,m}$ for any sample size n.

The bias of MEC error rate can then be estimated by $b_n^{IPL}(B) = \tilde{P}_{n,M} - \hat{P}_n^{IPL}$, where $\tilde{P}_{n,M}$ denotes the MEC error rate for a fixed dataset and $\hat{P}_n^{IPL} = min_m\ \hat{P}_{n,m}^{IPL}$.

The IPL approach has two advantages. Firstly, through subsampling and fitting by constructing learning curves, the IPL method borrows information from neighboring estimates at different sample sizes which has the potential to reduce the random noise of the true best classification error rate estimate and the estimator will be more stable and accurate. The second advantage for IPL is its potential to extrapolate the learning curves to estimate the true best classification error rate beyond the current sample size so that it can provide a prediction on how the error rate will further decline if more samples are included in future studies. For example, from an existing observed data of n=40 samples, IPL can estimate the expected accuracy at n=100 or n=250 samples and inform researchers whether it is worthwhile to extend the study to a larger cohort.

3.3 RESULTS

3.3.1 Bias correction of the simulated 2D example

We evaluated the performances of all four bias correction methods, Tibshirani's algorithm, nestedCV, WMC/WMCS and IPL, on the 2D toy model. We calculated both bias corrected estimates with M=2 classifiers (DLDA and QDA) and all M=10 classifiers and compared them with true best classification error rate at sample size n=20, 40, 80, 160, 640, 1280 with B=100 simulated datasets. As shown in the left panel of Figure 4, nestedCV method generally overestimated the true best classification error rate and the overestimated bias was larger when using 10 classifiers compared with 2 classifiers. Considering each inner cross validation model selection stage of nested cross validation, if the true best classifier (QDA in this example) was not selected frequently (i.e. $\tilde{m}_b^{(MEC)} \neq m^*$ using the notation in Section 3.2.1), the final minimal error rate was estimated with other suboptimal mix of classifiers. In this respect, on average, without being able to select the true best classifier to estimate the error rate on the test dataset, nestedCV resulted in an overestimate of the true best classification error rate. The upward bias problem could get more severe with more classifiers included which make the true best classifier less likely to be selected. Tibshirani's procedure sometimes overcorrects the bias as we can see the estimates at 40 and 80 when using 2 classifiers and 80 and 160 when using 10 classifiers, which are consistent with a prior technical report (Christoph Bernau, 2011). IPL methods under-estimate the true optimal error rate at sample size 20, but perform well for larger sample size. For WMC/WMCS, especially when using 10 classifiers, it can sometimes be conservative and sometimes anti-conservative. Overall, WMC/WMCS and IPL yield less variable estimates compared with nestedCV and BC.

To illustrate the advantage of IPL of extrapolating to predict performance in larger sample size, Figure 3-5 shows the estimated best error rates for sample sizes n=20, 30, 40, 60 and 80 using IPL (i.e. $\widehat{\mathrm{P}}_{\mathrm{n}}^{\mathrm{IPL}}$) from an observed dataset of n=40. The true best error rates ($\mathrm{P}_{\mathrm{n,M}}^{*}$) are marked by "cross" for reference. The IPL extrapolations generally estimated the truth pretty well. The result shows that increasing sample size from n=40 to n=80 only slightly improved the prediction accuracy and it is probably not worthwhile to collect an additional 40 samples.

3.3.2 Application on five GEO datasets

We applied all four methods (BC, nestedCV, WMC/WMCS and IPL) on 5 randomly selected GEO real datasets (see Appendix A for details on selection criteria). Since BC method is not applicable for leave-one-out cross validation, we applied it to 5-fold cross validation and 10-fold cross validation. For WMC/WMCS, B=30 subsampling times were applied with 80% portion .Ten machine learning methods were applied: KNN (K nearest neighbor with k=1, 3, and 5), DLDA (diagonal linear discriminant analysis), QDA (quadratic discriminant analysis), Neural network (NNET), SVM (support vector machine) with linear kernel, SVM with nonlinear kernel (radial), RF (random forest) and SCDA (shrunken centroids discriminant analysis). Feature selection was done by simple t-test with 2 to 30 top features selected by p-value in each cross validation to construct the classifiers separately, therefore, a total of 290 classifiers were utilized.

Table 3-1 shows the best error rate after bias correction by the all four methods.

The results show that MEC have a significant downward bias compared to the estimates from those correction methods especially for datasets GDS2190 and GDS2415. IPL generally gives smaller estimates compared to the rest methods which may be a result extrapolation at sample size n. NestedCV sometimes gives larger estimates compared to WMC/WMCS and sometimes smaller.

3.3.3 TCGA and METABRIC breast cancer data

Due to lack of the underlying truth in the real data in Section 3.3.2, the results could not be conclusive. To circumvent this shortcoming, we applied the three methods to two large breast cancer gene expression profiles, one from TCGA and one from METABRIC. The TCGA breast cancer dataset was downloaded from the Cancer Genome Atlas (TCGA) website (http://tcga-data.nci.nih.gov/tcga) in October 2012. Level 3 RNA-Seq data were extracted from the Illumina HiSeq 2000 platform. We selected the TCGA breast cancer data set that contained expression data of n=406 tumor samples. We defined two classification problems: one is to classify between ER positive (n=391) and ER negative (n=89) and the second is to classify between early stage (stage I and II, n=292) and late stage (stage III and IV, n=114) tumors. The METABRIC gene expression and clinical data are retrieved from Synapse (https://www.synapse.org/#!Synapse:syn2133309). We obtained 1897 samples, consisting of 945 early stage (stage I and II) and 952 late stage (stage III and IV) tumors (Curtis et al., 2012). We applied the same set of 290 classifiers in Section 3.2 to both datasets. Since the datasets contained a large sample size, we mimicked the simulation scheme described in Section 2.1. and randomly split the data into equal parts of ~40 samples. Under this setting, we pretended that we obtained B=10 independent data sets of n=40-41 from an unknown underlying distribution Δ_n in the TCGA dataset. Similarly, we have B=47 and n=40-41 for the METABRIC dataset. The random partition was also constrained such that sample sizes in two classes are as balanced as possible. By following the workflow of Figure 1, we generate the estimated best error rate $\widehat{P}^*_{n,M}(B)$ (denoted as "truth") and MEC error rates $\widetilde{P}_{n,M,b}$ $(1 \leq b \leq B)$ (denoted as "MEC") in Figure 3-6. Five-fold cross validation was used for Tibshirani's procedure,leave-one-out was used for nestedCV and IPL.

30 subsampling were used for both IPL and WMC/WMCS. For WMC/WMCS, subsampling was at proportion 80%. The results showed the nestedCV gives the most variable error estimates spanning a large range, which is confirmed also in (Bernau et al., 2013). WMC/WMCS yields more stable estimates compared with nestedCV. BC gives relatively stable estimates and can either overcorrect or under-correct the bias in terms of unconditional error rate, although in theorem it estimates the conditional error rates(Bernau et al., 2013). IPL gives the most stable and accurate error rates in terms of range of estimates as well as the closeness to the true minimal error rate especially in the case of classifying tumor stage in TCGA. The averaged MEC bias is as large as 1.69% in TCGA ER status classification, 6.15% in the TCGA tumor stage classification and 9.11% in the METABRIC tumor stage classification respectively. Without bias correction, an overly optimistic conclusion will be drawn.

3.4 CONCLUSION AND DISCUSSION

In this paper, we illustrated the downward bias of minimal-error classifier (MEC) error rate when selecting from many machine learning models in biomedical classification problems. In the application of high-throughput genomic data, this problem is especially magnified since the addition of feature selection easily increases the number of classification models to several hundreds. We firstly demonstrated the problem using a 2D toy example where QDA is known to be the best classifier. The simulation results and asymptotic theoretical results both illustrated the need of bias correction for MEC especially when the sample size (n) is limited and the number of classifiers examined (M) is large. We discussed two existing methods (the nested cross validation and Tibshirani's procedure) and developed a new inverse power law (IPL) method from the concept of learning curve fitting. Application of the three methods to the 2-D toy example, five selected GEO data sets and two large breast cancer data sets concluded that the NestedCV consistently overestimated the error rate and Tibshirani's procedure produced unstable bias correction. IPL provided a stable and accurate solution. The method has an additional advantage to extrapolate and predict the optimal error rate for larger sample sizes, a useful feature to help decide whether it is worthwhile to expand the study to recruit more samples.

With the advance of high-throughput genomic and proteomic techniques, data are generated at an unprecedentedly increasing pace. Machine learning methods have become a powerful tool in almost all biomedical research of complex diseases to seek new diagnostic or treatment selection tools. In most studies, small sample sizes are encountered (n=30-60) and researchers are tempted to test many classifiers and select the best to report (i.e. applying the minimal-error classifier; MEC). Our paper provides a careful framework and theoretical investigation of the problem and our result shows that severe bias can be generated for MEC with small sample size (e.g. n=30-60) and a large number of classifiers (e.g. M=300). Without bias correction, one runs the risk to obtain an overly optimistic error estimate of the classification model, excitedly expand the investigation to larger independent cohorts and eventually fail to validate and translate into a useful clinical tool. The IPL method we proposed in this paper not only generates more accurate bias correction, but also provide extrapolation estimates to determine whether larger cohorts might warrant improved accuracy. In the era of pursuing translational research and personalized (or precision) medicine, rigorous evaluation and interpretation of the machine learning results are essential to evaluate the clinical potential of a research finding.

There are a few limitations of our study. Firstly, The IPL method has the modeling assumption that learning curves of each classifier could be fitted by an inverse power law, although, as shown in the simulations as well as real datasets, this shows to be a reasonable assumption. Secondly, the error rates estimated by the correction methods included in this study are different which makes them less comparable, i.e. conditional error rate by Tibshirani's method, unconditional error rate by WMC/WMCS, nestedCV, IPL. Thirdly, the IPL methods are more costly compared to WMC/WMCS due to its subsampling at different points, however, since all the classifiers are fitted independently, it is very suitable for parallelization which will greatly reduce the cost and the feature is now included in the version of the package. Lastly, we sum up all different feature selection, machine learning methods and their associated parameter setting into M classifiers in the investigation. Theoretically different sources of classifiers have different correlated performance. Understanding their correlations may elucidate the contribution of bias from different sources and develop a better solution. In practice, one may determine high- and low-performance methods for empirical studies (e.g. comparison of performance in similar studies in large databases, such as GEO). How to systematically integrate the information to decide the set of classifiers for investigation is still an open question. All code and source files are available at http://tsenglab.biostat.pitt.edu/publication.htm to reproduce the results in the paper. A R package "MLbias" is also available.

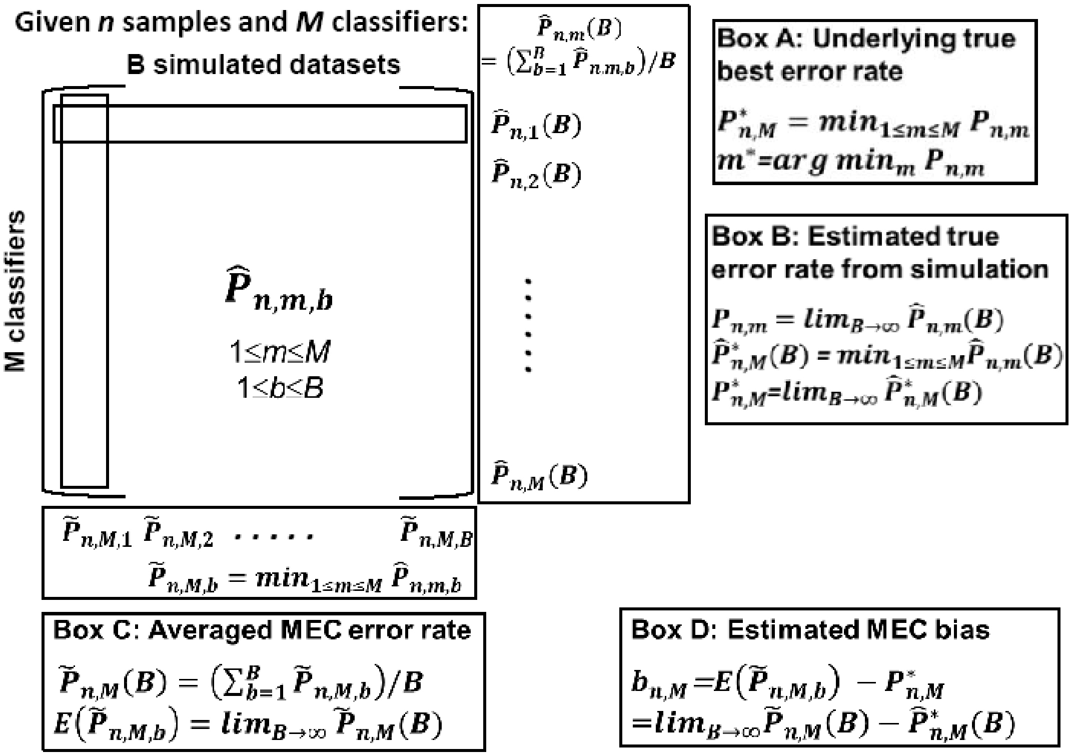
Given n samples and M classifiers:
B simulated datasets
M classifiers
$\widehat{P}_{n,m,b}$
$1\le m\le M$
$1<b<B$
$\widehat{P}_{n,m}(B) = (\sum_{b=1}^{B}\widehat{P}_{n,m,b})/B$
$\widehat{P}_{n,1}(B)$
$\widehat{P}_{n,2}(B)$
$\widehat{P}_{n,M}(B)$
$\widetilde{P}_{n,M,1}\ \widetilde{P}_{n,M,2}\ \cdots\cdots\ \widetilde{P}_{n,M,B}$
$\widetilde{P}_{n,M,b} = min_{1\le m\le M}\ \widehat{P}_{n,m,b}$
Box A: Underlying true best error rate
$P^*_{n,M} = min_{1\le m\le M}\ P_{n,m}$
$m^* = arg\ min_m\ P_{n,m}$
Box B: Estimated true error rate from simulation
$P_{n,m} = lim_{B\to\infty}\ \widehat{P}_{n,m}(B)$
$\widehat{P}^*_{n,M}(B) = min_{1\le m\le M}\widehat{P}_{n,m}(B)$
$P^*_{n,M} = lim_{B\to\infty}\ \widehat{P}^*_{n,M}(B)$
Box C: Averaged MEC error rate
$\widetilde{P}_{n,M}(B) = (\sum_{b=1}^{B}\widetilde{P}_{n,M,b})/B$
$E(\widetilde{P}_{n,M,b}) = lim_{B\to\infty}\ \widetilde{P}_{n,M}(B)$
Box D: Estimated MEC bias
$b_{n,M} = E(\widetilde{P}_{n,M,b})\ - P^*_{n,M}$
$= lim_{B\to\infty}\widetilde{P}_{n,M}(B) - \widehat{P}^*_{n,M}(B)$

Figure 3-1 The framework for estimating the true optimal classification error rate, MEC error rate and its bias from independent simulated data sets.

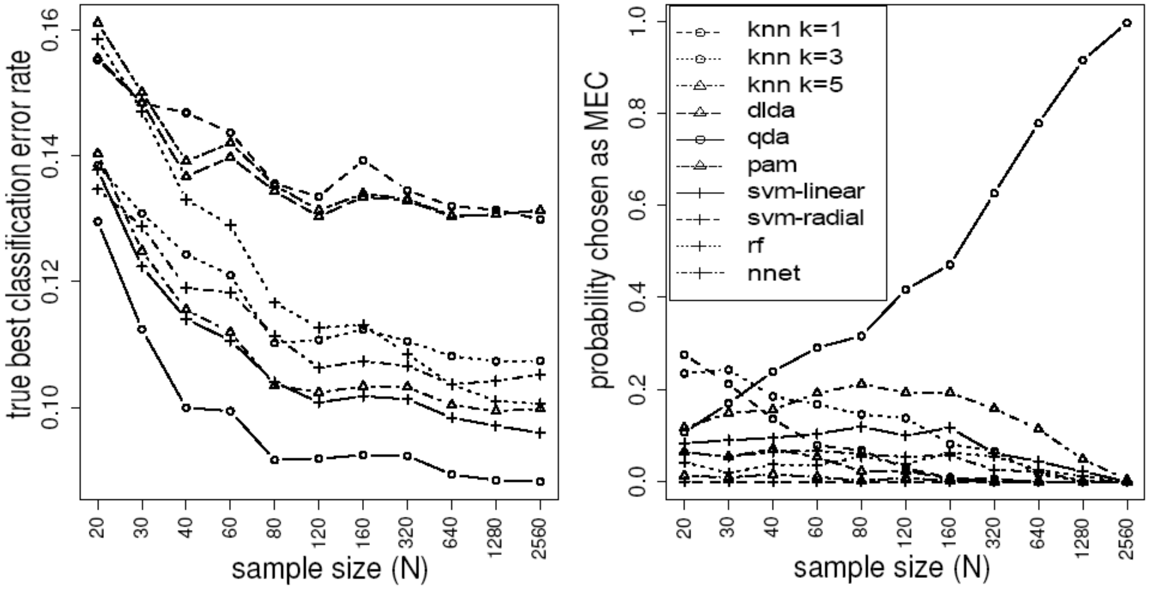

Figure 3-2 Error rate estimates in two dimensional toy model.

Left panel: The estimated classification error rate of each method at different sample sizes from 1000 independent simulations of the 2-D toy model in which QDA was the top classifier in this context. Right panel: Probability for each classifier to be chosen as the minimal error rate classifier in the 1000 simulations.

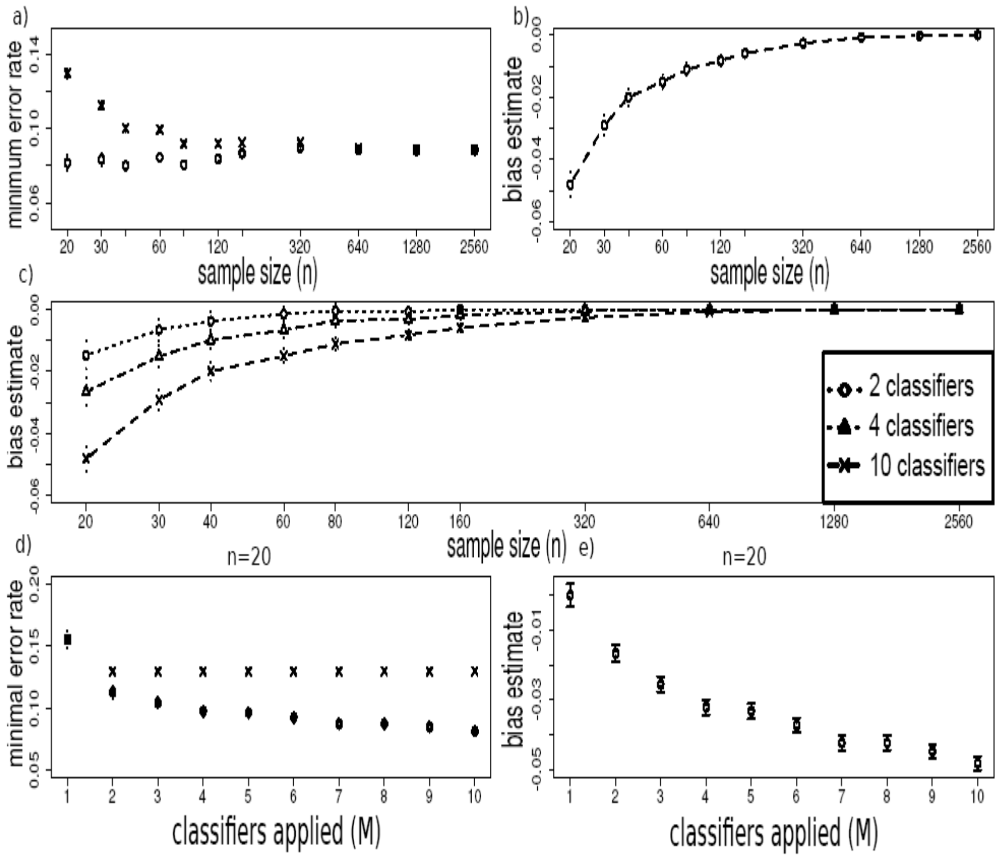

Figure 3-3 Illustration of the downward bias of MEC error rate

(a) trend of true optimal classification error rate (cross) and MEC error rate (circle) as sample size increased. (b) bias estimate of the MEC error rate diminishes as the sample size increased. (c) bias estimate when using all 10 classifiers (cross), 2 classifiers QDA and DLDA (circle) and 4 classifiers QDA, DLDA, SVM-linear and SVM-nonlinear (triangle). (d) For a fixed sample size n=20, the true minimal error rate (cross) and bias of MEC error rates (circle) as more classifiers were added in the sequence of KNN k=1, QDA, DLDA, KNN k=5, PAM, KNN K=3, SVM-linear, SVM-radial basis, RF and NNET. (e) The corresponding bias as the number of classifiers used increased from 2 to 10.

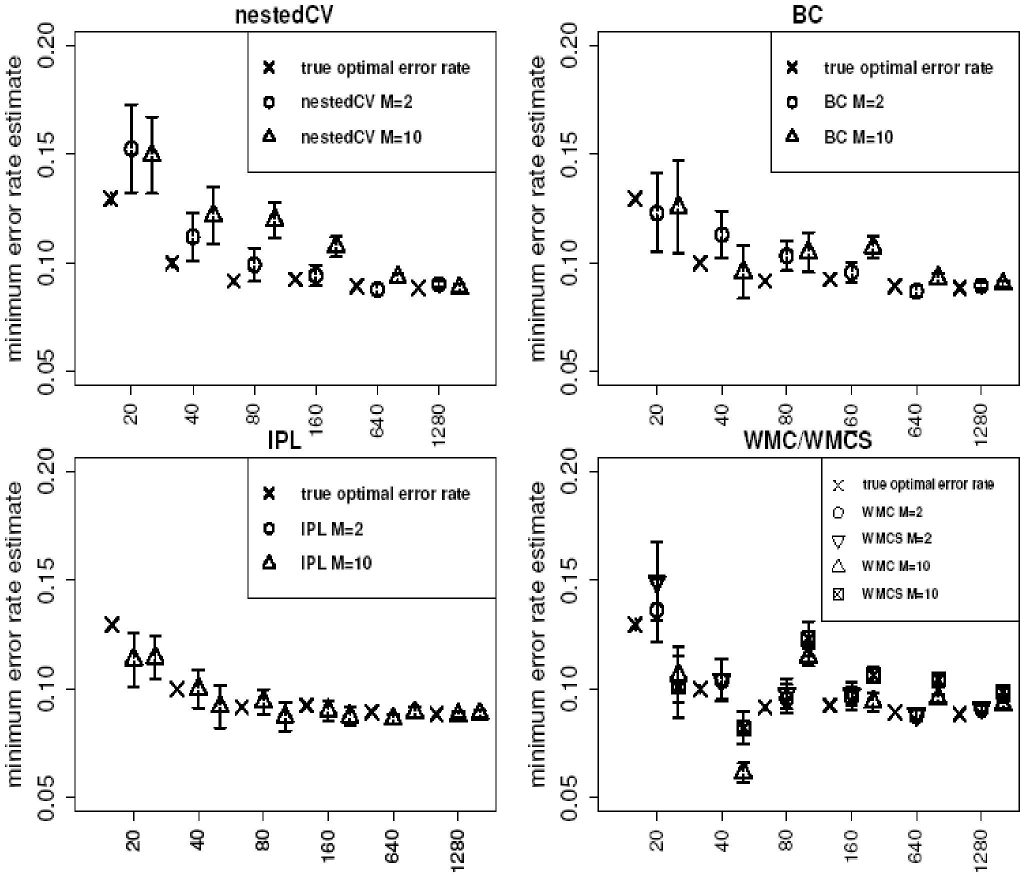

Figure 3-4 Comparison of minimal error rate estimates from all four bias correction methods.

Comparison of minimal error rate estimates from all four bias correction methods with 2 classifiers (QDA and DLDA), as well as with all 10 classifiers.

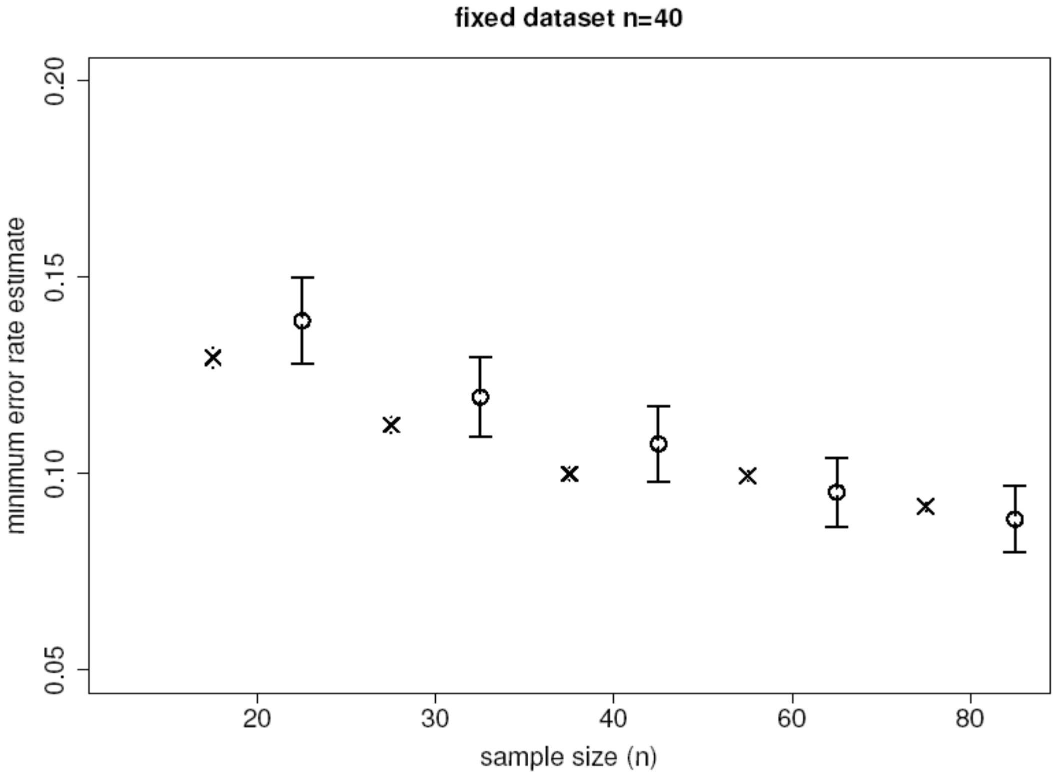

Figure 3-5 IPL estimates along with extrapolation.

Estimated minimum error rate (circle) using IPL extrapolation ($\hat{P}_n^{IPL}$) for a fixed dataset with n=40. The bars reflect 95% coverage interval from 100 simulations. The true minimum error rate ($P_{n,M}^{*}$) is shown (cross) as a reference.

Table 3-1 Error rate estimates on 5 GEO datasets

GEO datasets	n	MEC	BC (5)	BC (10)	nestedCV	IPL	WMC	WMCS
GDS1627	22	0.00	0.00	0.00	0.02	0.00	0.04	0.04
GDS2190	61	0.32	0.45	0.44	0.39	0.38	0.45	0.42
GDS2362	49	0.00	0.00	0.00	0.06	0.00	0.04	0.03
GDS2415	59	0.32	0.47	0.45	0.49	0.38	0.44	0.42
GDS2520	44	0.02	0.04	0.05	0.05	0.03	0.08	0.08

MEC error rate and corrected error rates by Tibshirani's procedure (5 fold and 10 fold cross validation), nested CV,IPL WMC, WMCS.

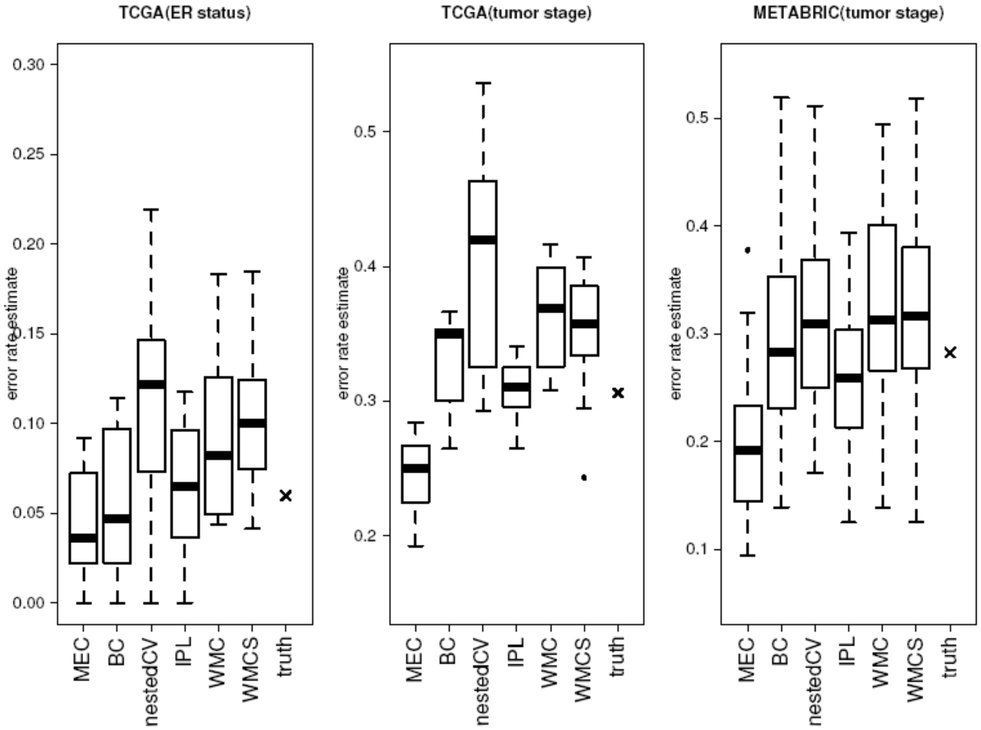

Figure 3-6 Error ratas estimates on real datasets.

MEC error rate estimates from all four correction methods along with the true best optimal classification error rate and MEC error rates on the TCGA and METABRIC datasets. Left panel: classification between ER positive and ER negative in TCGA. Middle and right panel: classification between early and late tumor stage in TCGA and METABRIC.

4.0 DIFFERENTIAL COEXPRESSED MODULE DETECTION

Ying Ding[1, 2], Sunghwan Kim[2], Etienne Sibille[3], Steffi Oesterreich[4], George C. Tseng[1, 2*]

[1] Joint Carnegie Mellon-University of Pittsburgh Ph.D. program in Computational Biology, Pittsburgh, PA, USA

[2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA

[3] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

[4]Magee-Womens Research Institute, Pittsburgh, PA, USA

[*] Corresponding author

Preface

Note on the contribution of YD to this paper: he is the sole author of all statistical analysis, simulations, methods, figures and text.

4.1 INTRODUCTION

Differential coexpression refers to changes in gene-gene correlations between two conditions (e.g. cases and controls). Changes in gene-gene correlation may occur in the absence of differential expression, meaning that a gene may undergo radical changes in regulatory pattern that would be undetected by traditional differential expression (DE) analyses. A specific phenotype could be contributed by differential coexpression without altering the expression levels of genes. This phenomenon has been found in aging (Southworth, Owen and Kim, 2009) as well as other biological conditions. Disease-associated changes in the regulatory systems that create coexpression relationships may be revealed through comparing gene-gene correlations that are computed separately from control and disease populations. Therefore, differential co-expression (DC) analysis can provide complementary information to standard differential expression (DE) analyses. Differential coexpression in two conditions could shy light on the potential biological mechanism. For example, a group of genes may be under the control of a common regulator (e.g. transcription factor or epigenetic modifications) which is active in one condition but disrupted in the other.

To detect differential coexpression between different conditions, some studies have aimed at detecting pairs (Lai et al., 2004) while others aim at detecting gene modules (Amar et al., 2013; Bhattacharyya and Bandyopadhyay, 2013). Some other method detect differential coexpression using predefined gene sets such as those based on known gene annotations, such as GO categories (Choi and Kendziorski, 2009). Although this approach incorporate prior information, it lacks the ability to detect novel differential modules. Another class of methods detect differential modules based on detecting modules in one reference condition but test the difference in the other condition. This type of methods rely on a clustering algorithm on one reference condition which may induce a bias towards one of the conditions (Ihmels et al., 2005; Watson, 2006). To circumvent this problem, methods have been developped for detecting differential coexpression modules independent of choosing a reference condition (Tesson, Breitling and Jansen, 2010).

The differential correlation relationship could arise from several biological and non-biological sources (Gaiteri et al., 2014). Any mechanism that synchronously regulates transcription of multiple genes, unwanted batch effect, mixture of tissues could potentially generate coexpression relationships. Therefore, instead of looking for differential coexpression between two conditions in a single study with classical application, differential coexpression may be confirmed across multiple datasets via meta-analyses to increase the detection power. Differential modules that are significant in one dataset may become more convincing if the differential patterns are preserved across multiple datasets. For instance, MDD (major depressive disorder) networks and control networks can be constructed separately in multiple studies and differential modules can be detected across multiple studies simultaneously so that the pattern occurs frequently in MDD networks but not in the control networks. Differences between modules can be assessed by different choices of measures; for example, differential modules with a predominant measure such as density (Li W, 2011) , other network measures (Kugler KG, 2011) or a weighted measure incorporating multiple network properties (Langfelder et al., 2011).

Few studies attempted to detect differential coexpressed modules with respect to certain phenotype across multiple studies. (Michael R. Mehan and Zhou, 2009) first proposed a simulated annealing based methods to detect differential modules which has density higher in one group but not the other. However, their method built pathway enrichment in the object function while searching, that is, the optimization phase heavily depends on the prior knowledge and also the output module size is small. In this study, we extend their idea and search the modules without any prior information. Moreover we propose ways to tune the weight parameters and control the false discovery rate (FDR) of detected modules. We evaluate the method on simulated data and two real datasets on breast cancer (ER+ vs ER-) and MDD (cases vs controls). The identified differential network modules are significantly enriched in many cancer and depression related pathways and shed light on the underlying disease mechanisms.

4.2 METHODS

4.2.1 Network construction

Starting with N datasets from N studies, e.g. gene expression arrays, we divided the samples in each dataset into two groups (referred as group1 and group2) with respect to the phenotype of interest such as case and control. For each study, we built two gene coexpression networks with 2N networks in total.

Gene coexpression networks could be generated with weighted or unweighted edges, we demonstrated our method based on unweighted networks, but the approach can easily be extended to weighted ones. To build unweighted networks and normalized across different studies, we first calculated all pair-wise gene-gene Pearson correlations, and then decide the cutoff for edge connections so that only the top 0.4% of possible connections in each network are kept (Lee et al., 2003). This is because different studies may have different sample sizes with different experiment platforms, which could result in relatively high or low correlation distributions.

The objective is to search for the module (network from a subset of genes) in one group that has consistently higher paired density difference than the other group. We examine the paired difference which can potentially cancel out potential differences between studies.

4.2.2 Target function

We propose to minimize the following energy function (target function) for identification of differential networks. :

$$\mathbf{E_{tot} = w_1 E_{size} + w_2 E_{density,min} + w_3 E_{diff,mean} + w_4 E_{diff,var}}$$

This target function searches modules with larger densities in group 1 than those in group 2 across N studies. The proposed target function comprises of the following four components: (1) $\mathbf{E_{size}}$ for the size of the module. (2) $\mathbf{E_{density,min}}$ for the minimum density in networks of group 1. (3) $\mathbf{E_{diff,mean}}$ for the average of the density differences between the two groups across N studies. (4) $\mathbf{E_{diff,var}}$ for the consistency of the density difference between the two groups across N studies. The first three terms have been employed by (Michael R. Mehan and Zhou, 2009). Note that, group1 and group2 can be switched in order to find higher density modules in group 2 than in group 1.

Let us first denote by x the genes in a module, |x| as the size of the module and $\delta_i(x)$ as the density of the module in study i. The first component $E_{size} = e^{(-\alpha(\frac{|x|}{\gamma} - o_s)}$ is related to the size of the modules which encourages the algorithm to output larger modules while penalizing very small modules (controlled by parameters α, γ, o_s). Without this term, dimmers and triplets which have density values 1 could easily dominate the output. The second component $E_{density,min} = e^{(-\alpha(\min_{i \in group1}(\delta_i(x)-o_\delta))}$ penalizes the module with small density in any one of the studies in group1 with parametersα, o_δ. The third component $E_{diff,mean} = \frac{\sum_i(\delta_{i,group1}-\delta_{i,group2})}{N}$ gives small values to the module with large average density differences between two groups. And the fourth component is the variance of the paired difference of density under two different conditions in each study which encourages consistent finding across studies with term $E_{diff,var} = \sqrt{\frac{(\delta_{i,group1}-\delta_{i,group2}-E_{diff,mean})^2}{N}}$. To incorporate all these terms, we weighted each term to get the total energy term as the final target function.

4.2.3 Optimization with simulated annealing

To minimize E_{tot}, due to its nonconvex nature, we applied simulated annealing, a stochastic optimization algorithm for non-convex optimization (Kirkpatrick, Gelatt and Vecchi, 1983). In each Monte Carlo step with simulated annealing, a new state is proposed denoted as X_{new}, which is either adding or removing a node (gene). If the resulting energy is smaller, then the state is accepted, if not, then the state is accepted with an acceptance probability as follows :

$$P_{acc} = \min\left(1, \frac{\pi(x_{new})p(x_{new \to old})}{\pi(x_{old})p(x_{old \to new})}\right)$$

where P_{acc} is the acceptance probability and is $p(x_{old \to new})$ is the transition probability from old state to the new state, if the probability of removing and adding a node is the same, both terms cancel. $\pi(x_{new})$ is the Boltzmann distribution of the energy function to be minimized:

$\pi(\mathrm{x_{new}}) = \exp\,(-\frac{\mathrm{E_{tot}(x_{new})}}{\mathrm{T}})$. T is a temperature parameter. When temperature is high, new trial moves will be accepted easily and thus more freely cross the local minimum, while when temperature gets lower, it tends to converge to a local minimum. We apply the temperature schedule $T_{(k+1)} = 0.95 * T_k$ and stop the annealing run if the acceptance ratio is smaller than 2%, where the acceptance ratio is calculated as the ratio of steps accepted in every 400 MC steps.

Due to large searching space, we bounded the module size between 3 and 30. If module size is 3, a node will be added for a new state while if module size is 30 then one node will be removed. A trial move gene set is updated in each step so that once a node is added, that corresponding gene is removed from the trial set while when removing a node, the node is added to the trial set. At the beginning, the trial set is determined as the genes which have at least one edge connected with the seed module genes in any of the networks in group 1, which enables local searching.

4.2.4 Seed module generation before optimization

A good starting point is critical for optimization in high dimensional space. Instead of randomly selecting a subset of genes from the genome, an edge-study matrix was constructed where rows represent all edges and columns represent all studies in two conditions of size 2N (Walley et al., 2012). A simple statistical paired t test is used to get initial set of differential coexpressed edges with a preset p value threshold (p<0.1). Based on these initial differential coexpressed edges, an initial network is constructed and connected components are generated. If the size of the connected component is larger than 30, then we randomly subsample 10 genes from it as the seed module for optimization starting points, otherwise, the optimization starts from the connected components directly. Other community detection algorithm may also be applied in this stage in order to generate better seed modules (Fortunato, 2010).

4.2.5 Control of False Discovery Rate (FDR)

In each simulated annealing run, a resulting module which is a suboptimal solution from the target function will be obtained. Suppose starting from K seed modules, M simulated annealing repeats, we will generate KM candidate differential modules. To control FDR of each module, we permute the labels of the case and control samples at the beginning with same procedures to obtain null modules. The null hypothesis here is that the case and control networks have no differences. Let $\mathrm{E_{ij}}$ denotes the optimized energy value of module u_{ij} from the i-th seed module and j-th simulated annealing repeat where $1 \le \mathrm{i} \le \mathrm{K}$ and $1 \le \mathrm{j} \le \mathrm{M}$. Suppose the permutation is conducted B times and the energy value after same procedure of optimization is denoted as $\mathrm{E_{i,j}^{(b)}}$ where $1 \le \mathrm{b} \le \mathrm{B}$, $1 \le \mathrm{i} \le \mathrm{K^{(b)}}$, $1 \le \mathrm{j} \le \mathrm{M}$ .Therefore, the p value for each module u_{ij} could be estimated as

$$\mathrm{p(E_{ij})} = \frac{\sum_{\mathrm{b=1}}^{\mathrm{B}} \sum_{\mathrm{i=1}}^{\mathrm{K^{(b)}}} \sum_{\mathrm{j=1}}^{\mathrm{M}} \mathrm{I\{E_{ij}^{(b)} \le E_{ij}\}+1}}{\mathrm{M} * \sum_{\mathrm{b=1}}^{\mathrm{B}} \mathrm{K^{(b)}}+1}.$$

Pseudo-count one is added to both the denominator and the numerator to avoid the zero p-values. FDR is derived from Benjamini & Hochberg correction to account for multiple comparison.

4.2.6 Assembly of small modules

Since the current approach limits the size of the module between 3 and 30, small modules may not yield significant pathway enrichment to inspire further hypothesis generation. Therefore, in order to obtain larger modules, we proposed to use statistical significance of pathway enrichment to guide module assembly. We applied pathway enrichment analysis on all detected modules under certain FDR cutoff (e.g. FDR<0.3). The enrichment was checked against 2,141 pathways downloaded from MSigDB (http://www.broadinstitute.org/gsea/msigdb/index.jsp) that contained Biocarta, KEGG, Reactome and GeneOntology databases (excluding large pathways with more than 200 genes). One sided Fisher's exact test (overrepresentation) was applied to calculate the statistical significance of overrepresentation.

Pathway enrichment was applied on each of the modules passing FDR<0.3. Then the enrichment p value for each pathway for those modules were combined with the sum of log2 enrichment p and ranked. The small negative combined value indicates several modules were all enriched in the same pathway, which may hint the potential relevance of the module to that pathway. We then examined the modules with enrichment p value <0.05 and assembled all those modules with respect to the specific pathway so that the new enrichment p value is the smallest. In this study, we limit the maximum modules to combine to be 3 due to computational issue. Assembling the modules can yield larger modules with more genes involved in specific pathway which might be of more interest to biologists.

4.2.7 Weight parameter tuning

To tune the parameters w_1, w_2, w_3, w_4 in the target function, we first constrain the sum of the four parameters to be 1000, i.e. $w_1 + w_2 + w_3 + w_4 = 1000$. In our experience, we found that the weight w_1, w_2 are only to prevent singular solutions (very tiny modules) and their weights do not contribute much to the quality of differential network identification. As a result, we fix the $w_1 = 100$ and $w_2 = 100$ and there is only one parameter w_3 remained ($w_4 = 800 - w_3$). We search w_3 from 100 to 700 with spacing 100. Then the problem is reduced as the tradeoff between the size of the density differences and the consistence of the density differences. Under each set of parameters, same permutation procedures described in Section 4.2.5 will be used to calculate the FDR for each identified module. To decide which parameters to choose, we output a table of all assembled modules, their associated energy function p values and pathway enrichment p values for users to decide. In general, we select w_3 such that the number of identified modules under FDR cutoff 0.3 is largest.

4.3 RESULTS

4.3.1 Simulation datasets

To illustrate the method, we first applied to simulated datasets of 4 studies. We simulated 100 artificial genes named from 1 to 100 and generated a module involving genes 1 to 15 with probability $\mathrm{P_{connect}} = 0.4$ of having an edge between any pair of these 15 genes. This module was treated as underlying true differential coexpressed module. We then simulated 4 pairs of networks with two conditions (group 1 and group 2), group 1 with the module densely connected while group 2 not. To model the noise in different studies, among the 4 networks in group 1, the module has a probability $\mathrm{p_{add}} \sim \mathrm{beta}(\mathrm{S_1}, 1)$ of adding an edge and probability $\mathrm{p_{delete}} \sim \mathrm{beta}(1, \mathrm{S_2})$ of deleting an edge. Here we chose a beta distribution simply to treat the noise on each edge differently and the probability of adding edges larger than missing edges. For the rest of genes 16-100 as well as the networks under condition 2, the edges are generated randomly with probability $\mathrm{p_{null}}$. In this simulation, we chose the parameters as $\mathrm{S_1} = 5, \mathrm{S_2} = 9, \mathrm{p_{null}} = 0.1$.

With these four pairs of simulated datasets, we constructed the edge-study matrix, obtained the differential coexpressed edges and calculated the connected components. The initial configuration for simulated annealing with an energy value ~2052 is shown in Figure 4-1 (A). After ~1500 Monte Carlo steps with trace plot shown in Figure 4-2, the energy is converged at ~1105. In this final configuration, 13 of the 15 genes in the true module were recovered and it has a better density contrast between the networks of two groups compared with the initial configuration as shown in Figure 4-1 (B).

4.3.2 Breast cancer datasets

We then applied our method to four large breast cancer datasets, including two GEO datasets, TCGA breast cancer datasets and METABRIC. The TCGA breast cancer dataset was downloaded from the Cancer Genome Atlas (TCGA) website (http://tcga-data.nci.nih.gov/tcga) in October 2012. Level 3 RNA-Seq data were extracted from the Illumina HiSeq 2000 platform. We selected the TCGA breast cancer data set that contained expression data of n=406 tumor samples. The METABRIC gene expression and clinical data are retrieved from Synapse (https://www.synapse.org/#!Synapse:syn2133309) where we obtained 1981 samples, (Curtis et al., 2012). The four breast cancer datasets are described in Table 4-1. Here we attempted to identify differentially coexpressed modules between networks from ER positive patients and networks from ER negative patients.

Microarrays were scanned and summarized by manufacturers' defaults. Data from Affymetrix arrays were processed by robust multi-array (RMA) method and data from Illumina arrays by manufacturer's BeadArray software for probe analysis. Oligonucleotide probes (or probesets) were matched to gene symbols using hgu133plus2.db and illuminaHumanv4.db Bioconductor packages. If multiple probes match to the same gene, then the gene with the largest inter-quantile range (IQR) is used. After matching all the genes across the four studies and filter away genes with an average standard deviation smaller than 0.2 across all studies, which left 10674 genes for following analysis.

4 pairs of gene coexpression networks were constructed for ER+ patients and ER- patients in the 4 studies. The cutoff for each unweighted network is chosen such that 0.4% possible edges are maintained among all possible edges. Edge-study matrices were calculated

and connected components were obtained as starting seeds for simulated annealing algorithm with repeat number M=20. FDR was calculated for each of the modules with permutation time B=20. The same procedure was conducted with different parameter sets of w_3 from 100 to 700 with spacing 100. w_3 was then selected to be 700 for breast cancer datasets by the criterion of largest number of output modules passing the FDR <0.3 threshold.

With $w_3 = 700$, two example modules, one densely connected in ER+ networks with 22 genes and one densely connected in ER- networks with 16 genes are illustrated in Figure 4-3(A), and Figure 4-3(B) respectively. Both modules achieved FDR 0.02. Although we found some modules with striking average paired differences, small modules (small sets of genes) usually did not yield significant pathway enrichment results. Therefore, module assembly was applied according to the procedures outlined before. Top list of pathways enriched by those assembled modules with q value < 0.005 are listed in Table 4-2 and Table 4-3.

For top pathways associated with assembled modules densely connected in ER+ status, many are cell cycle related pathways such as cell cycle pathway from KEGG, cell cycle checkpoint, S phase pathway and M_GI transition pathway from REACTOME, which are known to play an important role in cancer (Ouhtit et al., 2013). Activation of calpain is shown to induce anti-tumor activity in breast cancer and melanoma (Colunga et al., 2014). Also MTOR and EIF4 pathways are tightly related, inhibitor of mTOR haven been used as cancer therapy and EIF4E activity predicts sensitivity to mTOR inhibition in breast tumours (Satheesha et al., 2011). Also some other cancer pathway such as prostate cancer pathway is also enriched.

By examining the top pathways associated with assembled modules densely connected in ER- status, some immune related pathways are in the top list such as immune response pathway from GO, immunoregulatory interactions between a lymphoid and a non-lymphoid cell pathway and humoral immune response from REACTOME, B cell receptor signaling pathway from KEGG (Schmidt et al., 2008). It is shown that PD-L1 expression has association with tumor cells and objective response (Topalian et al., 2012). JAK-STAT signaling pathway is also highly related to different cancers.

4.3.3 Major depressive disorder (MDD) example

Then we applied our procedures to eight MDD microarray studies which contain a total of 51 MDD subjects and 50 control subjects. A description of the eight studies is listed in Table 4-4. Major depressive disorder is widely known for its weak signal provided by the evidence of genome-wide association studies (GWAS) (Major Depressive Disorder Working Group of the Psychiatric et al., 2013).

Microarrays were scanned and summarized by manufacturers' defaults. Data from Affymetrix arrays were processed by robust multi-array (RMA) method and data from Illumina arrays by manufacturer's BeadArray software for probe analysis. Batch effects were evaluated and normalized. Oligonucleotide probes (or probesets) were matched to gene symbols using hgu133plus2.db and illuminaHumanv4.db Bioconductor packages. The potential bias caused by this selection was addressed by permutation analysis in the following analytical steps. 16,689 unique genes were matched across the eight studies. Two sequential steps of gene filtering were then performed. First, genes with very low expression across studies were filtered out. Specifically, mean intensities of each gene across all samples in each study were calculated and the corresponding ranks were obtained. The sum of such ranks across eight studies of each gene was calculated and genes with the lowest 20% rank sum were considered non-expressed and were filtered out. Secondly, genes displaying very small variation in expression (i.e., lowest 20% rank sum of standard deviations) were filtered out, together leaving 10,680 unique genes. Then we further filtered away genes with average standard deviation smaller than 0.2 across all studies, which left 8121 genes for following analysis.

8 pairs of gene coexpression networks were constructed for MDD subjects and control subjects for the 8 studies. The cutoff for each unweighted network is chosen such that 0.4% possible edges are maintained among all possible edges. Simulation annealing repeat was chosen to be 20 (M=20). FDR was calculated on each output module based on permutation time B=25. The same procedure was conducted with different parameter sets of w_3 from 100 to 700 with spacing 100 as before. w_3 was then selected to be 200 since it has the largest number of detected modules with FDR<0.3 which contain 20 modules densely connected in MDD group and 37 modules densely connected in control group.

We showed two example modules when $w_3 = 200$, one densely connected in MDD networks and one densely connected in control networks as illustrated in Figure 4-4 (A), and Figure 4-4 (B) respectively. The former module achieved FDR 0.07 with 6 out of 8 studies have paired density difference more than 0.1 and the latter module achieved FDR 0.28 with 5 out of 8 studies have more than 0.1 difference. Although it is not consistent across all eight studies, the density difference relationship is preserved across the majority of the studies. This situation could come from either the heterogeneity of the studies (different brain region, gender) or could come from the cutoff selection when selecting the edges, if the cutoff for a specific pair of studies both have too high correlation cutoff, then the module will have low density in both networks. Output modules densely connected in control networks and densely connected in MDD networks were assembled separately as shown in 4.2.6. The assembled modules with their associated pathways enriched with q value <0.005 and other informations are shown in Table 4-5 (module densely connected in control group) and Table 4-6 (module densely connected in MDD group).

For assembled modules densely connected in control group shown in Table 4-5, the enriched pathways included three neurological diseases: Parkinson's disease with 13 genes in that pathway, as well as Huntington's disease with 11 genes and Alzheimer's disease with 10 genes in the pathway. These neurological diseases were shown to be associated with major depressive disorder (Folstein et al., 1983; Nilsson et al., 2002; Zubenko et al., 2003) Also, it is highly enriched in mitochondria where mitochondrial abnormalities and deficiencies in oxidative phosphorylation have been reported in individuals with schizophrenia (SZ), bipolar disorder (BD), and major depressive disorder (MDD) in transcriptomic (Rollins et al., 2009). A decrease of electron transport chain have been reported in bipolar disorder (Benes et al., 2006).

For assembly modules densely connected in MDD group shown in Table 4-6, there are also enriched pathways related to the mitochondria. Also there are cell cycle, apoptosis related pathways in the top list. It has been reported that depression could contribute to atrophy and cell loss in hippocampus and antidepressant treatment could increases neurogenesis (Duman, 2004).

As a whole, the identified differentially coexpressed modules could help elucidate the underlying disease mechanism and help generate new hypotheses to understand the molecular network basis of the disease.

4.4 CONLUSION

In this study, we proposed a method to simultaneously detect differential coexpression modules across multiple studies with respect to certain phenotype of interest. The method utilized prior free target function which is an extension of pathway guided target function proposed in (Michael R. Mehan and Zhou, 2009). We also presented the procedures to control the FDR as well as tuning the weights in the target function. The applicability of the optimization algorithm is illustrated in the simulation datasets. Then we applied the procedures on 4 breast cancer datasets as well as 8 MDD datasets. We showed the algorithm could output modules with large average paired density differences as well as consistent density differences. With module assembly guided by pathway analysis, we showed it is able to generate assembled modules highly enriched in cancer related pathways in the breast cancer example as well as depression related pathways in the MDD example.

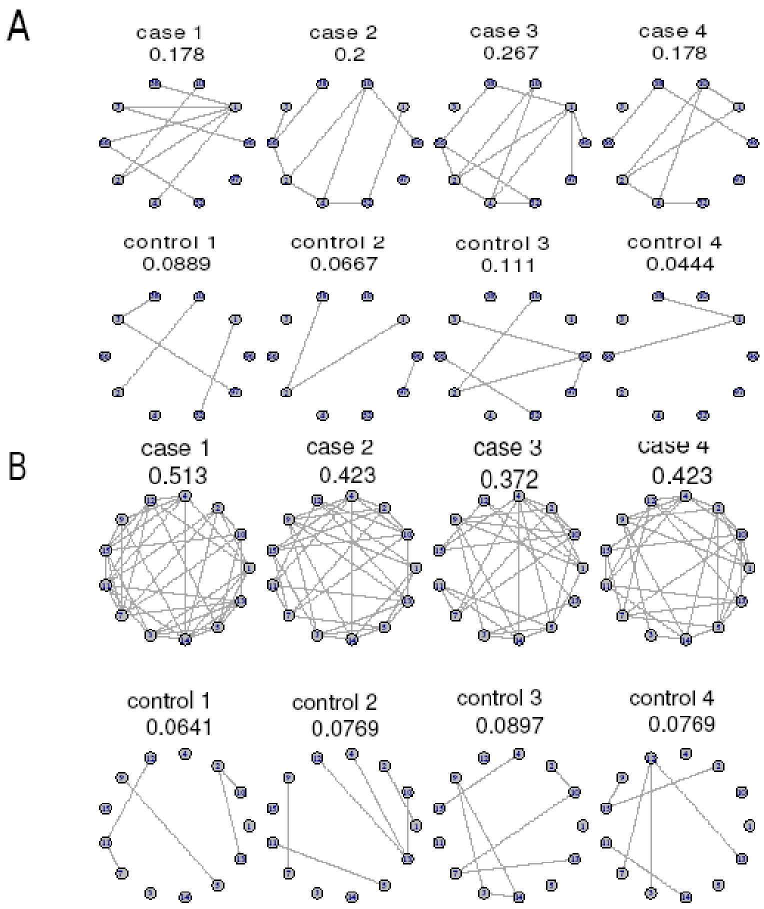

Figure 4-1 Module output before and after optimization in simulation datasets.

(A) Before optimization (B) After optimization.

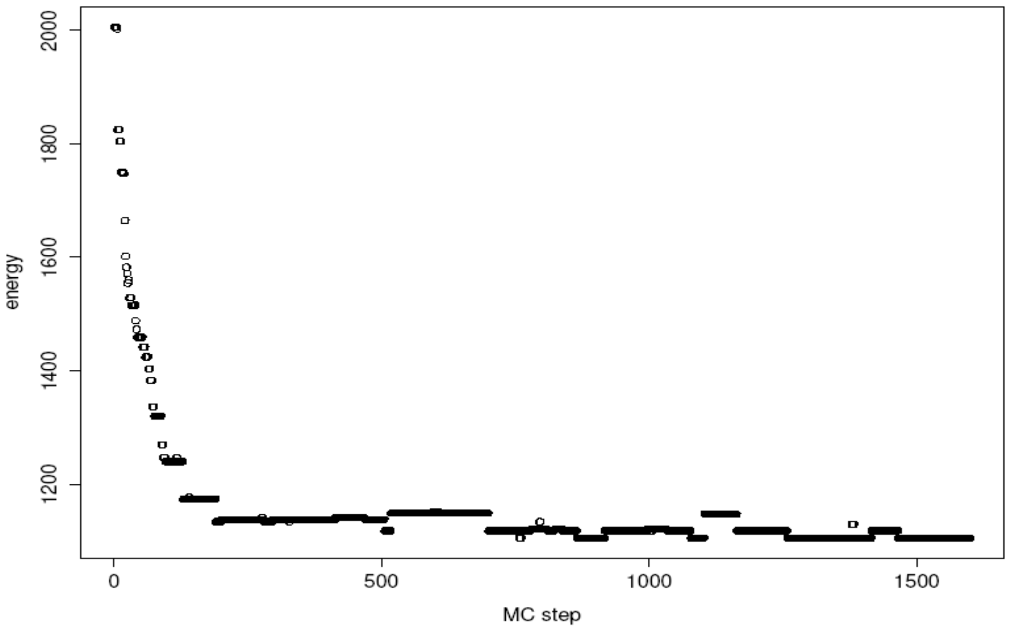

Figure 4-2 Energy trace plot in simulated annealing.

Table 4-1 Data description of four breast cancer datasets

Data sets	Sample size	Array platform
GSE7390	198 (134 vs 64)	Affymetrix Human Genome U133A Array
GSE2034	286 (209 vs 77)	Affymetrix Human Genome U133A Array
TCGA-breast cancer	406 (319 vs 87)	RNA-Seq
METABRIC	1981 (1512 vs 469)	Illumina

Table 4-2 Top pathways enriched after module assembly for breast cancer datasets (densely connected in ER+ group).

Pathway name	FDR (q value)	size	Genes in set	Mean density difference	Module assembled
KEGG_CELL_CYCLE	4.28e-10	46	9	0.29	M1,M3,M4
KEGG_HUNTINGTONS_DISEASE	4.81e-09	46	9	0.14	M1,M3,M33
KEGG_OXIDATIVE_PHOSPHORYLATION	5.46e-08	41	7	0.19	M1,M4,M8
REACTOME_CELL_CYCLE_CHECKPOINTS	1.7e-07	49	7	0.17	M2,M3,M17
REACTOME_ELECTRON_TRANSPORT_CHAIN	3.2e-07	31	5	0.20	M1,M9
REACTOME_PLATELET_ACTIVATION	5.11e-07	54	8	0.23	M2,M5,M8
BIOCARTA_MCALPAIN_PATHWAY	8.3e-07	35	4	0.26	M1,M3
REACTOME_S_PHASE	9.41e-07	40	6	0.33	M2,M3,M4
REACTOME_HIV_INFECTION	4.33e-06	29	6	0.24	M1,M3
KEGG_PROGESTERONE_MEDIATED_OOCYTE_MATURATION	3.94e-06	36	5	0.24	M1,M2
BIOCARTA_INTEGRIN_PATHWAY	4.21e-06	33	4	0.22	M1,M2
REACTOME_DNA_REPLICATION_PRE_INITIATION	6.31e-06	43	5	0.34	M1,M2,M3
BIOCARTA_UCALPAIN_PATHWAY	8.93e-06	28	3	0.19	M1,M11
GO_MF_ENZYME_INHIBITOR_ACTIVITY	1.05e-05	50	6	0.28	M1,M4,M6
REACTOME_M_G1_TRANSITION	1.65e-05	32	4	0.30	M1,M2
BIOCARTA_MTOR_PATHWAY	2.23e-05	30	3	0.18	M1,M22
BIOCARTA_EIF4_PATHWAY	2.23e-05	30	3	0.18	M1,M22

BIOCARTA_PTDINS_PATHWAY	2.99e-05	33	3	0.22	M2,M3
REACTOME_SYNTHESIS_OF_DNA	7.81e-05	32	4	0.30	M2,M3
GO_MF_ENDOPEPTIDASE_ACTIVITY	0.0001	54	5	0.30	M1,M2,M3
KEGG_PROSTATE_CANCER	0.0002	36	4	0.24	M1,M2

Table 4-3 Top pathways enriched after module assembly for breast cancer datasets (densely connected in ER-group).

Pathway name	FDR (q value)	size	Genes in set	Mean density difference	Module assembled
GO_BP_IMMUNE_RESPONSE	8.5e-12	31	12	0.47	M1,M5,M7
REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY	1.2e-06	38	6	0.42	M3,M8,M10
REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	3.5e-06	26	5	0.47	M3,M4
KEGG_JAK_STAT_SIGNALING_PATHWAY	4.3e-06	37	7	0.41	M1,M5,M6
REACTOME_PD1_SIGNALING	7.3e-06	37	4	0.43	M3,M5,M10
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY	9.4e-06	29	5	0.48	M5,M6
KEGG_LEISHMANIA_INFECTION	1.2e-05	37	5	0.43	M1,M2,M4
GO_BP_HUMORAL_IMMUNE_RESPONSE	1.2e-05	32	4	0.45	M3,M6,M8
GO_MF_CYTOKINE_BINDING	1.3e-05	24	4	0.51	M4,M5
GO_CC_EXTERNAL_SIDE_OF_PLASMA_MEMBRANE	1.8e-05	26	3	0.47	M5,M7
KEGG_CELL_ADHESION_MOLECULES_CAMS	2.1e-05	23	5	0.49	M2,M4
GO_BP_CALCIUM_MEDIATED_SIGNALING	2.4e-05	28	3	0.49	M3,M6

BIOCARTA_NKCELLS_PATHWAY	3.0e-05	24	3	0.51	M4,M6
GO_BP_CELLULAR_DEFENSE_RESPONSE	3.0e-05	28	4	0.48	M1,M4
GO_MF_INTERLEUKIN_BINDING	5.5e-05	24	3	0.51	M4,M5
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	0.0001	36	5	0.45	M2,M3,M4
KEGG_FOCAL_ADHESION	0.0005	29	5	0.25	M3,M4
GO_BP_ANTI_APOPTOSIS	0.0005	29	4	0.47	M1,M5,M7

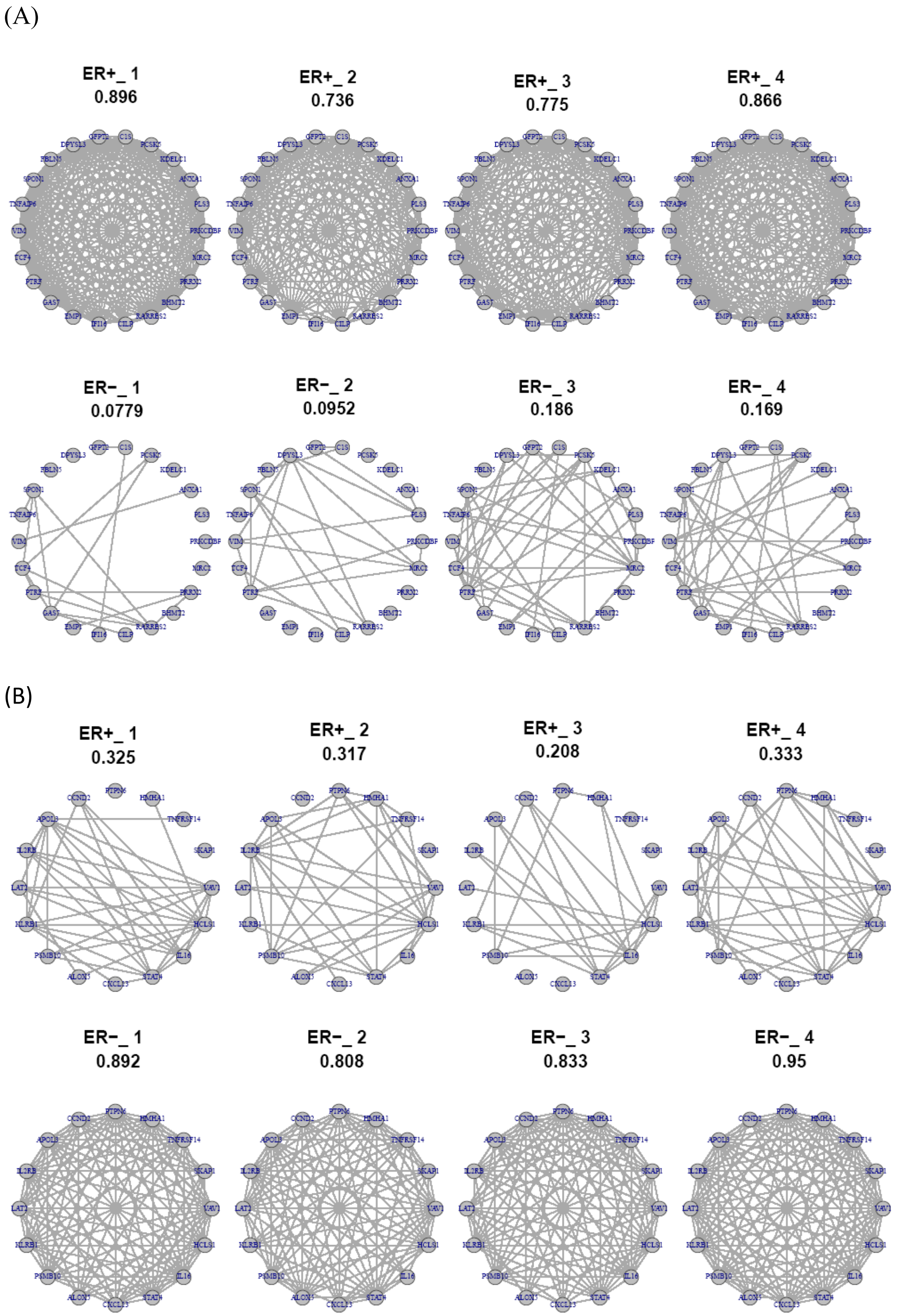
(A)
ER+_ 1
0.896
ER+_ 2
0.736
ER+_ 3
0.775
ER+_ 4
0.866
ER−_ 1
0.0779
ER−_ 2
0.0952
ER−_ 3
0.186
ER−_ 4
0.169
GFPT2
C1S
PCSK5
KDELC1
ANXA1
PLS3
PRKCDBP
MRC2
PRRX2
BHMT2
RARRES2
CILP
IFI16
EMP1
GAS7
PTRF
TCF4
VIM
TNFAIP6
SPON1
FBLN5
DPYSL3
(B)
ER+_ 1
0.325
ER+_ 2
0.317
ER+_ 3
0.208
ER+_ 4
0.333
ER−_ 1
0.892
ER−_ 2
0.808
ER−_ 3
0.833
ER−_ 4
0.95
PTPN6
HMHA1
TNFRSF14
SKAP1
VAV1
HCLS1
IL16
STAT4
CXCL13
ALOX5
PSMB10
KLRB1
LAT2
IL2RB
APOL3
CCND2

Figure 4-3 Two example modules in breast cancer datasets. Example module densely connected in ER+ group but not in ER- group. (B) example module densely connected in ER- group but not in ER+ group.

Table 4-4 Data description of eight MDD microarray studies.

Study name	Gender	Brain region	Sample size	Array platform
1-MD_ACC_M	Male	ACC	30 (15 pairs)	Affy. HG-U133 Plus 2
2-MD_ACC_M	Male	ACC	18 (9 pairs)	Affy. HG-U133 Plus 2
3-MD_ACC_F	Female	ACC	28 (14 pairs)	IlluminaHumanHT-12
4-MD_ACC_F	Female	ACC	22 (11 pairs)	IlluminaHumanHT-12
5-MD_AMY_M	Male	AMY	28 (14 pairs)	Affy. HG-U133 Plus 2
6-MD_AMY_F	Female	AMY	42 (21 pairs)	IlluminaHumanHT-12
7-MD_DLPFC_F	Female	DLPFC	28 (14 pairs)	Affy. HG-U133 Plus 2
8-MD_DLPFC_M	Male	DLPFC	32 (16 pairs)	Affy. HG-U133 Plus 2

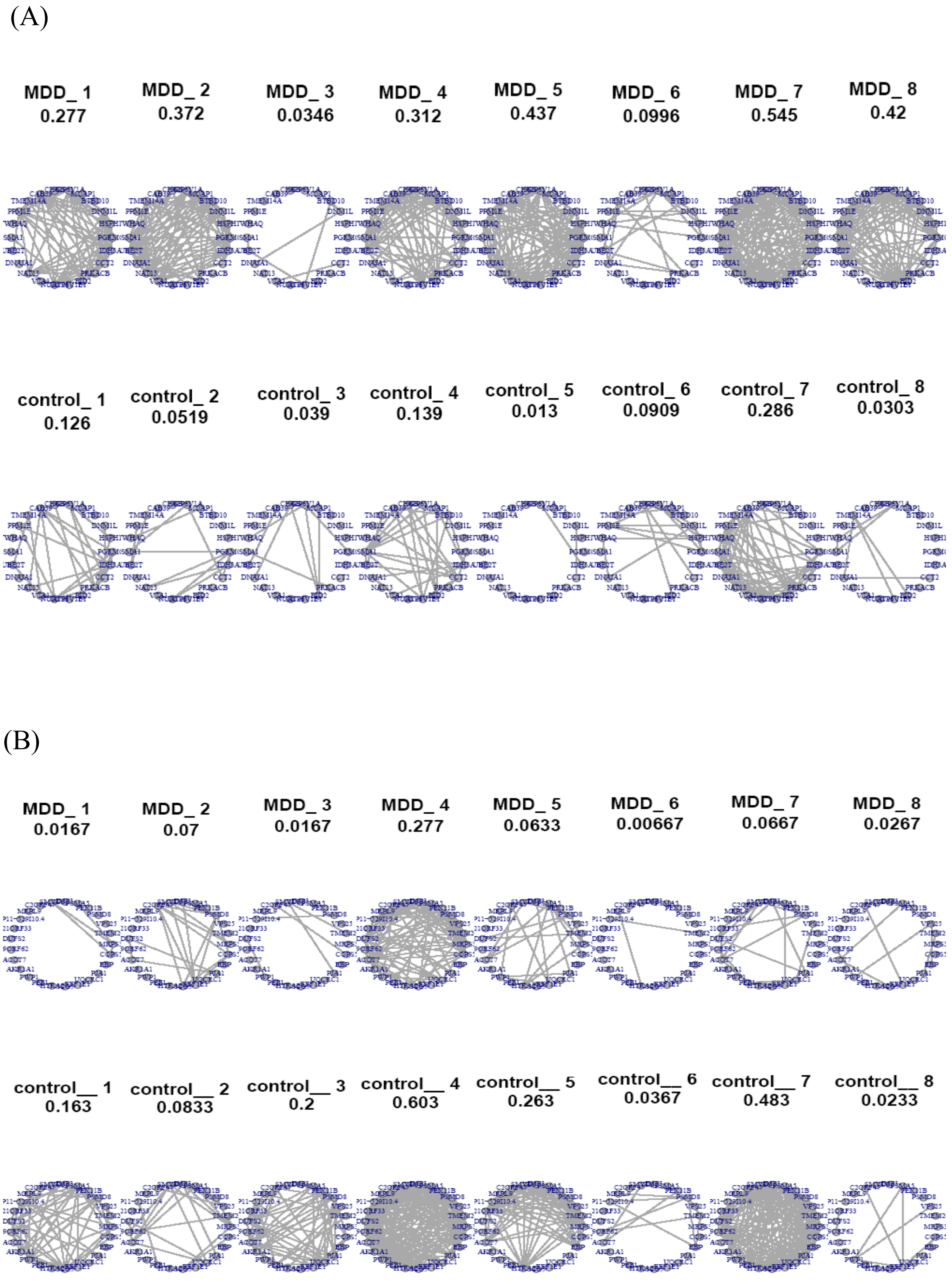
(A)
MDD_ 1
0.277
MDD_ 2
0.372
MDD_ 3
0.0346
MDD_ 4
0.312
MDD_ 5
0.437
MDD_ 6
0.0996
MDD_ 7
0.545
MDD_ 8
0.42
control_ 1
0.126
control_ 2
0.0519
control_ 3
0.039
control_ 4
0.139
control_ 5
0.013
control_ 6
0.0909
control_ 7
0.286
control_ 8
0.0303
CAB39
TMEM14A
BTBD10
DNM1L
HSPH1
YWHAQ
CCT2
PRKACB
NAT13
DNAJA1
(B)
MDD_ 1
0.0167
MDD_ 2
0.07
MDD_ 3
0.0167
MDD_ 4
0.277
MDD_ 5
0.0633
MDD_ 6
0.00667
MDD_ 7
0.0667
MDD_ 8
0.0267
control__ 1
0.163
control__ 2
0.0833
control__ 3
0.2
control__ 4
0.603
control__ 5
0.263
control__ 6
0.0367
control__ 7
0.483
control__ 8
0.0233
PEX11B
VPS25
AKR1A1
PWP1
PPA1
EBP

Figure 4-4 Two example modules in MDD datasets. (A) Example module densely connected in MDD group but not in control group. (B) example module densely connected in control group but not in MDD group.

Table 4-5 Top pathways enriched after module assembly for MDD datasets (densely connected in control).

Pathway name	FDR (q value)	size	Genes in set	Mean density difference	Module assembled
KEGG_PARKINSONS_DISEASE	5.2e-11	72	13	0.10	M2,M5,M8
REACTOME_GLUCOSE_REGULATION_OF_INSULIN_SECRETION	3.3e-09	72	13	0.10	M8,M13
KEGG_OXIDATIVE_PHOSPHORYLATION	3.3e-09	72	12	0.10	M6,M9
REACTOME_REGULATION_OF_INSULIN_SECRETION	3.1e-08	72	13	0.10	M7,M12
KEGG_HUNTINGTONS_DISEASE	7.3e-08	72	11	0.10	M1,M3
GO_CC_MITOCHONDRIAL_ENVELOPE	3.4e-07	69	9	0.10	M1,M6
REACTOME_ELECTRON_TRANSPORT_CHAIN	4.7-07	70	8	0.10	M1,M2,M3
GO_CC_MITOCHONDRIAL_RESPIRATORY_CHAIN	5.2e-07	69	6	0.09	M4,M5
KEGG_ALZHEIMERS_DISEASE	5.3e-07	72	10	0.10	M4,M9
GO_CC_NADH_DEHYDROGENASE_COMPLEX	1.3e-05	52	4	0.11	M1,M2
GO_MF_ELECTRON_CARRIER_ACTIVITY	0.001	69	5	0.11	M4,M9
GO_MF_ISOMERASE_ACTIVITY	0.001	48	3	0.12	M1,M3
GO_MF_OXIDOREDUCTASE_ACTIVITY_ACTING_ON_NADH_OR_NADPH	0.002	53	3	0.11	M1,M3
KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.005	41	3	0.14	M1,M2

Table 4-6 Top pathways enriched after module assembly for MDD datasets (densely connected in MDD).

Pathway name	FDR (q value)	size	Genes in set	Mean density difference	Module assembled
GO_BP_RESPONSE_TO_BIOTIC_STIMULUS	0.002	58	5	-0.13	M2,M5,M8
GO_BP_APOPTOTIC_PROGRAM	0.005	45	3	-0.14	M6,M9
GO_BP_MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	0.005	46	3	-0.13	M1,M3
KEGG_EPITHELIAL_CELL_SIGNALING_IN_HELICOBACTER_PYLORI_INFECTION	0.005	62	4	-0.14	M1,M2,M3
KEGG_OXIDATIVE_PHOSPHORYLATION	0.005	47	5	-0.12	M4,M5
GO_BP_MEMBRANE_ORGANIZATION_AND_BIOGENESIS	0.005	61	5	-0.14	M1,M3,M4
GO_BP_CELLULAR_COMPONENT_DISASSEMBLY	0.005	50	3	-0.13	M1,M2
REACTOME_CELL_CYCLE_CHECKPOINTS	0.005	47	4	-0.15	M1,M2
GO_MF_SMALL_CONJUGATING_PROTEIN_LIGASE_ACTIVITY	0.005	46	3	-0.16	M1,M2
KEGG_CITRATE_CYCLE_TCA_CYCLE	0.005	45	3	-0.17	M1,M2

5.0 CONCLUSIONS AND FUTURE DIRECTIONS

5.1 SUMMARY OF CONTRIBUTIONS

Disease biomarkers are increasingly important in this era of personalized medicine, as they can both help predict the progress of a disease as well as its response to a treatment. How to discover biomarkers and how to construct models with biomarkers and evaluate their prediction performances are important methodological questions that could facilitate the progress of this field. In this thesis, we have presented one application study and two computational methods that contribute to the biomarker research field.

Firstly, in a collaborative study, we constructed prediction models based on eight novel fusion genes detected from RNA sequencing with consistent prediction accuracies on prostate cancer recurrence (defined by relapse within 5 years) and aggressiveness (defined by PSA doubling time < 4 months) across three centers in Pittsburgh, Stanford and Wisconsin. Our prediction models could potentially help reduce the overtreatment of the disease. This finding was the first study combining fusion genes alone to predict prostate cancer prognosis. We showed that fusion genes, although individually rare, can predict prostate cancer recurrence and aggressiveness.

We next developed a new method, IPL, which corrects the bias of reporting the minimal cross validation error rate when multiple machine learning methods are applied. The IPL method estimates the true minimal error rate based on fitting the individual learning curve of each classifier independently. The error rate of each classifier is then extrapolated to the sample size we want to estimate. IPL was able to provide more robust estimates compared to existing bias correction methods as shown in the 2-D toy model and two large breast cancer datasets. Prior to the development of IPL, existing bias correction methods depended on evaluating performances based on cross validation or subsampling; therefore, error rates were estimated on part of the dataset. These methods thus may give estimates that either overestimate the bias or are more variable (Bernau et al., 2013; Tibshirani, 2009; Varma and Simon, 2006). The use of extraction in the IPL method avoids these problems. IPL is of particular value to researchers who conduct small clinical studies with the objective of building prediction models to predict disease status or prognosis. Moreover, the ability to extrapolate beyond the existing sample size can aid study design by exploring how much additional prediction value can be gained if more patients are recruited.

Finally, we presented a computational algorithm (MetaDiffNetwork) to identify coexpression modules that are consistently differentially coexpressed across disease conditions in multiple transcriptomic studies. By modifying a previous target function (Michael R. Mehan and Zhou, 2009), our method does not rely on prior information at the initial module searching stage. Our method also includes steps to control the false discovery rate of output modules and to tune the weight in the target function. In order to generate larger modules, our method uses a pathway guided approach to assemble statistically significant modules to maximize pathway enrichment with respect to specific biological pathways. We demonstrated good performance of this algorithm via simulation as well as testing on two real datasets including a set of eight major depressive disorder transcriptome studies (postmortem brain samples from MDD patients versus matched control subjects) and four breast cancer studies (ER+ vs ER-). The identified modules were validated using knowledge of existing disease pathways for each disease. This method is useful for researchers with multiple expression datasets of similar designs at hand who wish to find a set of genes which give consistent differential coexpressed patterns across disease conditions. This enables researchers to generate biological hypotheses on the potential mechanisms that could contribute to the differential patterns.

In conclusion, the thesis first demonstrated in a collaborative study that eight novel fusion genes detected from transcriptome sequencing under deep coverage could predict prostate cancer recurrence as well as prostate cancer aggressiveness with consistent results across three sites with >200 patients. The thesis then presented two new methods intended to aid biomarker research: IPL and MetaDiffNetwork. The first of these methods corrects the bias when reporting minimal cross validation error rates when multiple machine learning methods are applied, which is a common problem in small clinical studies. The latter provides a method to detect network modules which are consistently differentially coexpressed between disease conditions across multiple studies. Taken together, this thesis provides new contributions to biomarker research which can eventually be used to benefit clinical decision-making, including evaluating the prediction performance of biomarker models and revealing potential novel biomarkers from a network perspective.

5.2 FUTURE DIRECTIONS

Below I briefly discuss possible future directions from this thesis.

5.2.1 Combining fusion gene with existing measures to improve prediction

In Chapter 2, we only considered using fusion genes to predict cancer recurrence and aggressiveness. Future studies may combine fusion genes with other measures, such as Gleason score, tumor stage and PSA (prostate-specific antigen) level. It would also be useful to compare the current model to the existing nomogram by Memorial Sloan-Kettering Cancer Center (Kattan, Wheeler and Scardino, 1999), which used a simple regression model incorporating common diagnoses measures such as the patient's PSA value prior to surgery or other treatment, the primary Gleason grade at surgery, secondary Gleason grade at surgery, year of prostatectomy, number of months the patient has been disease-free, whether surgical margins were positive, and so on. It would be helpful to evaluate how much additional value can be added by combining both the fusion gene and nomogram models. Also the current design of this study did not consider the most prevalent TMPRSS2-ERG2 fusion; it would be of interest to assess the value of the model if that fusion is included.

5.2.2 Power analysis on fusion gene analysis using RNA-Seq

The findings in Chapter 2 suggest that down-sampling the original RNA-seq datasets of 1300X coverage would miss some of the previously detected fusion genes. This could result from two factors: 1) tumor tissues are often contaminated by normal tissues; and 2) even within tumor tissues, not all cells would harbor the fusion genes. Therefore, good quality control of the samples used to detect features is of utmost importance. Missing features reduces the number of possible predictors, resulting in a low sensitivity. Future work may involve modeling the power of fusion gene detection. For example, a patient carrying a low-allelic-fraction fusion gene may not be detected if sequencing depth R is low and will be falsely diagnosed. The power calculation framework can involve three major factors of the specific fusion gene, prevalence φ, allelic fraction λ and abundance of expression β. Prevalence φ is the proportion of patients among the population that carry the fusion gene. Low prevalence will require a large sample size (N) to detect fusion genes with high sensitivity. Allelic fraction λ is the proportion of cells carrying the fusion gene and abundance β is the proportion of fusion transcripts relative to the entire transcriptome. A fusion gene with small λ and/or small β will require high sequencing depth (R) to guarantee detection. As a result, the power function Pow(N,R) of a target fusion transcript would provide useful information to investigators given pre-specified φ, λ and β.

5.2.3 Increasing the speed of IPL and taking into consideration the dependence structure of classifiers in bias correction

Although IPL gives a more robust estimate compared to existing methods, as shown in Chapter 3, it can be more computationally extensive due to subsampling multiple times at smaller sample sizes. Future work could involve improving the IPL method by making it more computational efficient. Since the learning curves of all classifiers are fitted independently, it would be very suitable for parallelization which would greatly reduce the computational cost. For example, computation for different classifiers could be sent to different cores.

Another aspect of the method that could be optimized is the strategy for combining the classifiers. Currently, we sum together all the different feature selections, machine learning methods and their associated parameter setting into M classifiers without taking into consideration the correlation structure of the classifiers. Understanding their correlations may elucidate the contribution of bias from different sources and may yield a better solution. In practice, if one could determine high- and low-performance methods for empirical studies, this could aid in the selection of the set of classifiers for investigation.

5.2.4 Improving module assembly process

Currently, our method MetaDiffNetwork for assembling small modules into larger modules guided by pathway enrichment, as shown in Chapter 4, solved the problem of small sizes of output modules (<30) (Michael R. Mehan and Zhou, 2009). However, this module assembly stage does not consider the decrease of average density difference, and instead only considers the significance gain in the enrichment p value. It would be useful to assemble the modules with another round of simulated annealing by incorporating both enrichment p values with respect to specific pathways as well as average density differences into the target function. This stage could help trim the larger module to provide a balance between pathway enrichment and density differences between disease conditions across multiple studies.

5.2.5 Summarizing differential coexpressed modules to aid biological exploration

Our method MetaDiffNetwork (Chapter 4) has focused primarily on detecting modules which have consistent density differences between two conditions across multiple studies as well as associating the modules after assembly with the existing pathway knowledge. However, to let biologists better explore the results and generate biological hypotheses, it would be beneficial to summarize the differential modules across multiple studies into one module which can be easily visualized and explored using tools such as Cytoscape (Shannon et al., 2003). To summarize the differential module, a standardized score could be designed to represent the weight on each edge. Denote $u_{ij}^{(k)}$ as the correlation between gene i and j in study k in group 1, while $v_{ij}^{(k)}$ as the correlation between gene i and j in study k in group 2. Let $d_{ij}^{k} = u_{ij}^{(k)} - v_{ij}^{(k)}$ and let $\overline{d_{ij}}$ be the mean of paired correlation differences, σ_{ij} be the standard deviation of paired correlation differences and σ_o be the fudge parameter that could be estimated by the median of all standard devision of correlation difference. Then the edge weight of the summarized module could be represented by the standardized score $Z_{ij} = \frac{\overline{d_{ij}}}{\sigma_{ij}+\sigma_0}$. In visualization, edge width and node size could be adjusted to represent the magnitude of the weight as well as the degree of the node. To better understand the biology, the nodes (genes) that are enriched in transcription factors, protein-protein interactions as well as drug interactions could be highlighted and the edges which overlap with known interaction could be highlighted as well. This would provide a useful tool for biologists to explore the network and generate biological hypotheses on potential disease mechanisms.

APPENDIX A- Supplementary information for Chapter 2

TRANSLOATION AND FUSION TRANSCRIPTS IN PROGRESSIVE PROSTATE CANCER

Supplementary Figure and Tables

Supplementary Figures and Tables

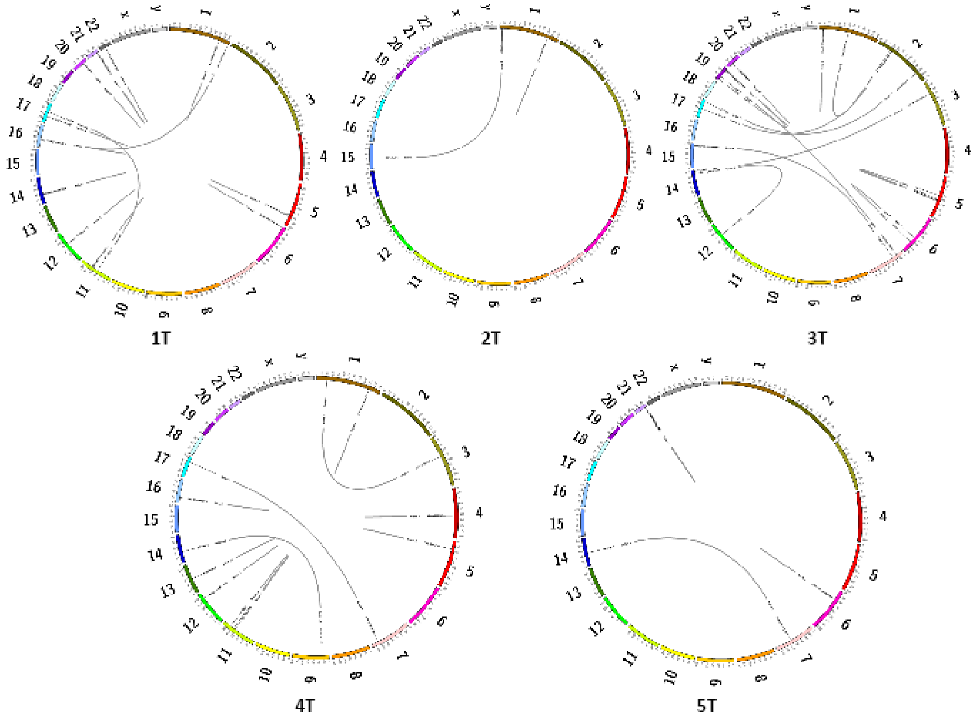

Figure S-5-1 Circus plots of prostate cancer functional genome translocation.

Five prostate cancer functional translocations were based on RNA sequencing. Fourteen of these functional translocations were supported by whole genome sequencing analysis. Functional translocation is defined as at least one transcript identified in the translocation process. Translocations in non-gene area were excluded.

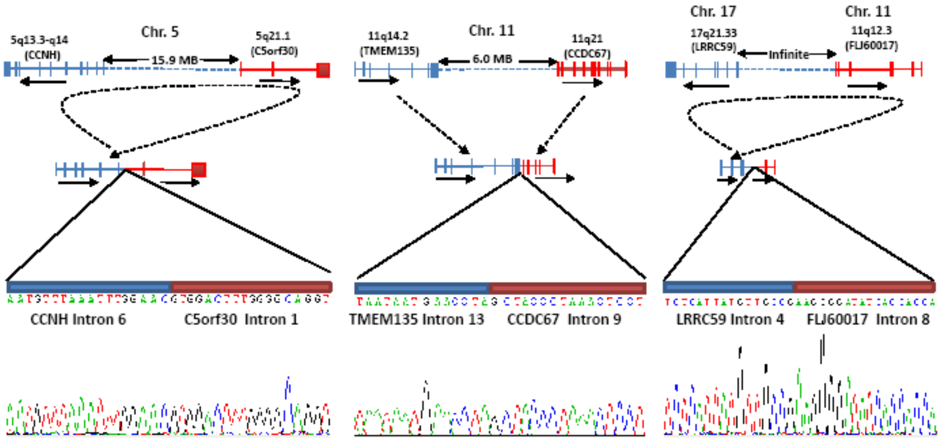

Figure S-5-2 Genome breakpoint analysis of fusion genes

Top panel: Miniature diagrams of genome of the fusion genes, the transcription directions, the distances between the joining genes and directions of the chromosome joining. Middle panel: Miniature of fusion genome and transcription direction. Bottom: Representative sequencing chromograms encompassing the joining breakpoint of chromosomes.

APPENDIX B- Supplementary information for Chapter 3

BIAS CORRECTION FOR SELECTING THE MINIMAL-ERROR CLASSIFIERS FROM MANY MACHINE LEARNING MODELS

Supplementary Proofs for theoretical results

Supplementary Method - Selection of the 5 GEO datasets.

B.1 SUPPLEMENTARY PROOFS FOR THEORETICAL RESULTS

Proof for Theorem 1

Proof. This is trivial, since $A \in B$ always implies that $min_{x \in B} B \leq min_{x \in A} A$.

Proof for Theorem 2

Proof. For $X = \{X_1,\ldots,X_m\}$ with $X_i \geq 0$, let us a function $f(\cdot)$ as

$$f(X) = \min_i \{X_i\}$$

then obviously $f(\cdot)$ is a concave function, which implies that by Jensen inequality, it holds

$$f\big(E(\hat{P}_{n,M})\big) \geq E(f(\hat{P}_{n,M}))$$

which is equivalent to

$$\min_m \{E(\hat{P}_{n,M})\} \geq E(\min_m(\hat{P}_{n,M})) \Leftrightarrow E(\tilde{P}^*_{n,M}\ \ ) \leq P^*_{n,M}$$

Recall that the equality holds only if $f(\cdot)$ is linear and or $\hat{P}_{n,M}$ is constant. Obviously this is not true for our case, therefore the strict inequality holds, i.e.,

$$E(\tilde{P}^*_{n,M}\ \ ) < P^*_{n,M}$$

Proof for Theorem 3

Proof. Let us denote by $\hat{C}_{i,m}$ and $C_{i,m}$ the estimated and the true class of sample i predicted by classifier m respectively for $i = 1, \cdots, n$ and $m = 1, \cdots, M$, then the estimated error rate of classifier m is $\hat{P}_{n,M} = \frac{\sum_{i=1}^{n} I(\hat{C}_{i,m} \neq C_{i,m})}{n}$, where $I(\cdot)$ is the indicator function. Obviously $\hat{P}_{n,M}$ follows a binomial distribution with mean P_m^* and variance $\frac{P_m^*(1-P_m^*)}{n}$.

Without loss of generality, let us assume $m_0 = arg\min_{1 \le m \le M}\{P_m^*\}$ (i.e., there is only one optimal classifier) and define

$$\delta = \min_{m \neq m_0} \{|P_m^* - P_{m_0}^*|\}$$

then by Chebyshev inequality, it holds

$P\left(\left|P_{n,m}^* - P_m^* \ \right| > \frac{\delta}{2}\right) < \frac{4P_m^*(1-P_m^*)}{n\delta^2} \to 0 \ as \ n \to \infty.$

which implies that for any $m \neq m_0$, it holds

$P\left(\hat{P}_{n,m} > P_{m_0}^* + \frac{\delta}{2}\right) \to 1 \ as \ n \to \infty.$

and therefore

$\lim_{n\to\infty} \tilde{P}_{n,M}^* = \lim_{n\to\infty} P_{n,M}^* = \lim_{n\to\infty} P_{m_0}^*$ holds with probability 1.

The proof can be easily extended to the case where more than one classifier have the same minimal error rate. Therefore this theorem states that when sample size is large enough, one always can find one classifier with minimal error rate.

B.2 SUPPLEMENTARY METHODS-SELECTION OF THE 5 GEO DATASETS

1. Search "Homo sapiens [porgn: txid9606]" at http://www.ncbi.nlm.nih.gov/geo/

2. Obtain 984 human datasets and 324,059 human samples.

3. Normalization: log transform, median-center to 0, and normalize SD

to 1. Note: expression can be null.

4. Selected 764 datasets with # samples >= 6.

5. For the gene ID with multiple probe IDs, take average expression.

After obtaining all the datasets, the five representative datasets are selected randomly among the datasets larger than 40.

BIBLIOGRAPHY

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55-65.

Amar, D., Safer, H., and Shamir, R. (2013). Dissection of regulatory networks that are altered in disease via differential co-expression. *PLoS Comput Biol* **9**, e1002955.

Antonarakis, E. S., Zahurak, M. L., Lin, J., Keizman, D., Carducci, M. A., and Eisenberger, M. A. Changes in PSA kinetics predict metastasis- free survival in men with PSA-recurrent prostate cancer treated with nonhormonal agents: combined analysis of 4 phase II trials. *Cancer* **118**, 1533-1542.

Asmann, Y. W., Hossain, A., Necela, B. M., *et al.* (2011). A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res* **39**, e100.

Atkinson, A. J., Colburn, W. A., DeGruttola, V. G., *et al.* (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*. *Clin Pharmacol Ther* **69**, 89-95.

Baca, S. C., Prandi, D., Lawrence, M. S., *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677.

Baek, S., Tsai, C. A., and Chen, J. J. (2009). Development of biomarker classifiers from high-dimensional data. *Brief Bioinform* **10**, 537-546.

Bamford, S., Dawson, E., Forbes, S., *et al.* (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91**, 355-358.

Baskerville, S., and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**, 241-247.

Benelli, M., Pescucci, C., Marseglia, G., Severgnini, M., Torricelli, F., and Magi, A. (2012). Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript. *Bioinformatics* **28**, 3232-3239.

Benes, F. M., Matzilevich, D., Burke, R. E., and Walsh, J. (2006). The expression of proapoptosis genes is increased in bipolar disorder, but not in schizophrenia. *Mol Psychiatry* **11**, 241-251.

Berger, M. F., Lawrence, M. S., Demichelis, F., *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220.

Bernau, C., Augustin, T., and Boulesteix, A. L. (2013). Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics* **69**, 693-702.

Berrar, D., Bradbury, I., and Dubitzky, W. (2006). Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22**, 1245-1250.

Bhattacharyya, M., and Bandyopadhyay, S. (2013). Studying the differential co-expression of microRNAs reveals significant role of white matter in early Alzheimer's progression. *Mol Biosyst* **9**, 457-466.

Boulesteix, A.-L., and Strobl, C. (2009). Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Medical Research Methodology* **9**, 85.

Check, E. (2004). Proteomics and cancer: running before we can walk? *Nature* **429**, 496-497.

Chen, K., Wallis, J. W., Kandoth, C., *et al.* (2012). BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics* **28**, 1923-1924.

Chen, L., and Zhao, H. (2005). Gene expression analysis reveals that histone deacetylation sites may serve as partitions of chromatin gene expression domains. *BMC Genomics* **6**, 44.

Choi, Y., and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics* **25**, 2780-2786.

Christoph Bernau, T. A. a. A.-L. B. (2011). Correcting the optimally selected resampling-based error rate: A smooth analytical alternative to nested cross-validation. *Technical Report.*

Colunga, A., Bollino, D., Schech, A., and Aurelian, L. (2014). Calpain-dependent clearance of the autophagy protein p62/SQSTM1 is a contributor to DeltaPK oncolytic activity in melanoma. *Gene Ther* **21**, 371-378.

Curtis, C., Shah, S. P., Chin, S. F., *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-352.

Demichelis, F., Fall, K., Perner, S., *et al.* (2007). TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene* **26**, 4596-4599.

Deng, Y., Dai, X., Xiang, Q., *et al.* (2010). Genome-wide analysis of the effect of histone modifications on the coexpression of neighboring genes in Saccharomyces cerevisiae. *BMC Genomics* **11**, 550.

Dong, H., Luo, L., Hong, S., *et al.* (2010). Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma. *BMC Systems Biology* **4**, 163.

Duman, R. S. (2004). Depression: a case of neuronal life and death? *Biol Psychiatry* **56**, 140-145.

Dupuy, A., and Simon, R. M. (2007). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* **99**, 147-157.

Edgren, H., Murumagi, A., Kangaspeska, S., *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* **12**, R6.

Efron, B. (2009). Empirical Bayes Estimates for Large-Scale Prediction Problems. *J Am Stat Assoc* **104**, 1015-1028.

Fernandez-Luna, J. L. (2000). Bcr-Abl and inhibition of apoptosis in chronic myelogenous leukemia cells. *Apoptosis* **5**, 315-318.

Fisher, R. P., and Morgan, D. O. (1994). A novel cyclin associates with MO15/CDK7 to form the CDK-activating kinase. *Cell* **78**, 713-724.

Folstein, S., Abbott, M. H., Chase, G. A., Jensen, B. A., and Folstein, M. F. (1983). The association of affective disorder with Huntington's disease in a case series and in families. *Psychol Med* **13**, 537-542.

Forbes, S. A., Bindal, N., Bamford, S., *et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-950.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**, 75-174.

Freedland, S. J., Humphreys, E. B., Mangold, L. A., *et al.* (2007). Death in patients with recurrent prostate cancer after radical prostatectomy: prostate-specific antigen doubling time subgroups and their associated contributions to all-cause mortality. *J Clin Oncol* **25**, 1765-1771.

Frenkel-Morgenstern, M., Gorohovski, A., Lacroix, V., *et al.* (2013). ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. *Nucleic Acids Res* **41**, D142-151.

Fu, W. J., Carroll, R. J., and Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics* **21**, 1979-1986.

Gaiteri, C., Ding, Y., French, B., Tseng, G. C., and Sibille, E. (2014). Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes Brain Behav* **13**, 13-24.

Gingeras, T. R. (2009). Implications of chimaeric non-co-linear transcripts. *Nature* **461**, 206-211.

Gopalan, A., Leversha, M. A., Satagopan, J. M., *et al.* (2009). TMPRSS2-ERG gene fusion is not associated with outcome in patients treated by prostatectomy. *Cancer Res* **69**, 1400-1406.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet* **11**, 476-486.

Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005). Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. *PLoS Genet* **1**, e39.

Iyer, M. K., Chinnaiyan, A. M., and Maher, C. A. (2011). ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903-2904.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2012). Global cancer statistics. *CA Cancer J Clin*.

Jia, W., Qiu, K., He, M., *et al.* (2013). SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biol* **14**, R12.

Kattan, M. W., Wheeler, T. M., and Scardino, P. T. (1999). Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *J Clin Oncol* **17**, 1499-1507.

Kaye, F. J. (2009). Mutation-associated fusion cancer genes in solid tumors. *Mol Cancer Ther* **8**, 1399-1408.

Kim, D., and Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12**, R72.

Kirkpatrick, S., Gelatt, C. D., Jr., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Kostka, D., and Spang, R. (2004). Finding disease specific alterations in the co-expression of genes. *Bioinformatics* **20 Suppl 1**, i194-199.

Kugler KG, M. L., Graber A Dehmer M (2011). Integrative Network Biology: Graph Prototyping for Co-Expression Cancer Networks. *PLoS ONE* **6(7)**, e22843.

Lai, Y., Wu, B., Chen, L., and Zhao, H. (2004). A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* **20**, 3146-3155.

Langfelder, P., Luo, R., Oldham, M. C., and Horvath, S. (2011). Is my network module preserved and reproducible? *PLoS Comput Biol* **7**, e1001057.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2003). Coexpression analysis of human genes across many microarray data sets. *Genome Research* **14**, 1085-1094.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760.

Li, R., Yu, C., Li, Y.*, et al.* (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967.

Li W, L. C.-C., Zhang T Li H Waterman MS Zhou XJ (2011). Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation. *PLoS Comput Biol* **7(6)**, e1001106.

Liu, C., Ma, J., Chang, C. J., and Zhou, X. (2013). FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. *BMC Bioinformatics* **14**, 193.

Luo, J. H., Ding, Y., Chen, R.*, et al.* (2013). Genome-wide methylation analysis of prostate tissues reveals global methylation patterns of prostate cancer. *Am J Pathol* **182**, 2028-2036.

Luo, J. H., Yu, Y. P., Cieply, K.*, et al.* (2002). Gene expression analysis of prostate cancers. *Mol Carcinog* **33**, 25-35.

Major Depressive Disorder Working Group of the Psychiatric, G. C., Ripke, S., Wray, N. R.*, et al.* (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry* **18**, 497-511.

Marco, A., Konikoff, C., Karr, T. L., and Kumar, S. (2009). Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics* **25**, 2473--2477.

McGettigan, P. A. (2013). Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol* **17**, 4-11.

McPherson, A., Hormozdiari, F., Zayed, A.*, et al.* (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol* **7**, e1001138.

Michael R. Mehan, J. N.-I., Mrinal Kalakrishnan Michael S. Waterman, and Zhou, X. J. (2009). An Integrative Network Approach to Map the Transcriptome to the Phenome. *Journal of Computational Biology* **16(8)**, 1023-1034.

Misago, M., Liao, Y. F., Kudo, S.*, et al.* (1995). Molecular cloning and expression of cDNAs encoding human alpha-mannosidase II and a previously unrecognized alpha-mannosidase IIx isozyme. *Proc Natl Acad Sci U S A* **92**, 11766-11770.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245.

Mitelman, F., Mertens, F., and Johansson, B. (2005). Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Genes Chromosomes Cancer* **43**, 350-366.

Mukherjee, S., Tamayo, P., Rogers, S.*, et al.* (2003). Estimating Dataset Size Requirements for Classifying DNA Microarray Data. *Journal of Computational Biology* **10**, 119-142.

Nam, R. K., Sugar, L., Yang, W.*, et al.* (2007). Expression of the TMPRSS2:ERG fusion gene predicts cancer recurrence after surgery for localised prostate cancer. *Br J Cancer* **97**, 1690-1695.

Nilsson, F. M., Kessing, L. V., Sorensen, T. M., Andersen, P. K., and Bolwig, T. G. (2002). Major depressive disorder in Parkinson's disease: a register-based study. *Acta Psychiatr Scand* **106**, 202-211.

Novo, F. J., de Mendibil, I. O., and Vizmanos, J. L. (2007). TICdb: a collection of gene-mapped translocation breakpoints in cancer. *BMC Genomics* **8**, 33.

Ouhtit, A., Gaur, R. L., Abdraboh, M., *et al.* (2013). Simultaneous inhibition of cell-cycle, proliferation, survival, metastatic pathways and induction of apoptosis in breast cancer cells by a phytochemical super-cocktail: genes that underpin its mode of action. *J Cancer* **4**, 703-715.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet* **2**, 418-427.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-375.

Ren, B., Yu, G., Tseng, G. C., *et al.* (2006). MCM7 amplification and overexpression are associated with prostate cancer progression. *Oncogene* **25**, 1090-1098.

Ren, S., Peng, Z., Mao, J. H., *et al.* (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res* **22**, 806-821.

Reverter, A., Ingham, A., Lehnert, S. A., *et al.* (2006). Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. *Bioinformatics* **22**, 2396-2404.

Rollins, B., Martin, M. V., Sequeira, P. A., *et al.* (2009). Mitochondrial variants in schizophrenia, bipolar disorder, and major depressive disorder. *PLoS ONE* **4**, e4913.

Romero Otero, J., Garcia Gomez, B., Campos Juanatey, F., and Touijer, K. A. (2014). Prostate cancer biomarkers: An update. *Urol Oncol* **32**, 252-260.

Salagierski, M., and Schalken, J. A. (2012). Molecular diagnosis of prostate cancer: PCA3 and TMPRSS2:ERG gene fusion. *J Urol* **187**, 795-801.

Satheesha, S., Cookson, V. J., Coleman, L. J., *et al.* (2011). Response to mTOR inhibition: activity of eIF4E predicts sensitivity in cell lines and acquired changes in eIF4E regulation in breast cancer. *Mol Cancer* **10**, 19.

Savolainen, K., Kotti, T. J., Schmitz, W., *et al.* (2004). A mouse model for alpha-methylacyl-CoA racemase deficiency: adjustment of bile acid synthesis and intolerance to dietary methyl-branched lipids. *Hum Mol Genet* **13**, 955-965.

Schmidt, M., Bohm, D., von Torne, C., *et al.* (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res* **68**, 5405-5413.

Shannon, P., Markiel, A., Ozier, O., *et al.* (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498--2504.

Siegel, R., Naishadham, D., and Jemal, A. (2012). Cancer statistics, 2012. *CA Cancer J Clin* **62**, 10-29.

Sinclair, P. B., Sorour, A., Martineau, M., *et al.* (2004). A fluorescence in situ hybridization map of 6q deletions in acute lymphocytic leukemia: identification and analysis of a candidate tumor suppressor gene. *Cancer Res* **64**, 4089-4098.

Slawski, M., Daumer, M., and Boulesteix, A.-L. (2008). CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics* **9**, 439--.

Soda, M., Choi, Y. L., Enomoto, M., *et al.* (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* **448**, 561-566.

Southworth, L. K., Owen, A. B., and Kim, S. K. (2009). Aging Mice Show a Decreasing Correlation of Gene Expression within Genetic Modules. *Plos Genetics* **5**.

Subramanian, A., Tamayo, P., Mootha, V. K., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* **11**, 497.

Tibshirani, R. J. T. a. R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat* **3**, 822-829.

Tkachuk, D. C., Westbrook, C. A., Andreeff, M., *et al.* (1990). Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. *Science* **250**, 559-562.

Tomlins, S. A., Rhodes, D. R., Perner, S., *et al.* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-648.

Topalian, S. L., Hodi, F. S., Brahmer, J. R., *et al.* (2012). Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med* **366**, 2443-2454.

Towns, W. L., and Begley, T. J. (2012). Transfer RNA methytransferases and their corresponding modifications in budding yeast and humans: activities, predications, and potential roles in human health. *DNA Cell Biol* **31**, 434-454.

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7**, 91.

Walley, A. J., Jacobson, P., Falchi, M., *et al.* (2012). Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. *Int J Obes (Lond)* **36**, 137-147.

Wang, H., Luo, K., Tan, L. Z., *et al.* (2012). p53-induced gene 3 mediates cell death induced by glutathione peroxidase 3. *J Biol Chem* **287**, 16890-16902.

Watson, M. (2006). CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics* **7**, 509.

Wei Zeng, C.-W. F., Stefan Muller Arisona, Huamin Qu (2013). Visualizing Interchange Patterns in Massive Movement Data. *Computer Graphics Forum*, 271-280.

Wood, I. A., Visscher, P. M., and Mengersen, K. L. (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics* **23**, 1363-1370.

Yang, H., Rudge, D. G., Koos, J. D., Vaidialingam, B., Yang, H. J., and Pavletich, N. P. (2013). mTOR kinase structure, mechanism and regulation. *Nature* **497**, 217-223.

Yang, J., Jubb, A. M., Pike, L., *et al.* The histone demethylase JMJD2B is regulated by estrogen receptor alpha and hypoxia, and is a key mediator of estrogen induced growth. *Cancer Res* **70**, 6456-6466.

Yousefi, M. R., Hua, J., and Dougherty, E. R. (2011). Multiple-rule bias in the comparison of classification rules. *Bioinformatics* **27**, 1675-1683.

Yu, Y. P., Ding, Y., Chen, R., *et al.* (2013). Whole-genome methylation sequencing reveals distinct impact of differential methylations on gene transcription in prostate cancer. *Am J Pathol* **183**, 1960-1970.

Yu, Y. P., Landsittel, D., Jing, L., *et al.* (2004). Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J Clin Oncol* **22**, 2790-2799.

Yu, Y. P., Yu, G., Tseng, G., *et al.* (2007). Glutathione peroxidase 3, deleted or methylated in prostate cancer, suppresses prostate cancer growth and metastasis. *Cancer Res* **67**, 8043-8050.

Zhen, Y., Sorensen, V., Skjerpen, C. S., *et al.* (2012). Nuclear import of exogenous FGF1 requires the ER-protein LRRC59 and the importins Kpnalpha1 and Kpnbeta1. *Traffic* **13**, 650-664.

Ziegler, A., Koch, A., Krockenberger, K., and Grosshennig, A. (2012). Personalized medicine using DNA biomarkers: a review. *Hum Genet* **131**, 1627-1638.

Zografos, G., Liakakos, T., and Roukos, D. H. (2013). Deep sequencing and integrative genome analysis: approaching a new class of biomarkers and therapeutic targets for breast cancer. *Pharmacogenomics* **14**, 5-8.

Zubenko, G. S., Zubenko, W. N., McPherson, S., *et al.* (2003). A collaborative study of the emergence and clinical features of the major depressive syndrome of Alzheimer's disease. *Am J Psychiatry* **160**, 857-866.