ENHANCED INTER-STUDY PREDICTION AND BIOMARKER DETECTION IN MICROARRAY WITH APPLICATION TO CANCER STUDIES

by

Chunrong Cheng

MS in Statistics, West Virginia University, 2004

BS in Pediatrics, Jiangxi Medical College, P.R. China, 1998

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Chunrong Cheng

It was defended on

June 23rd, 2008

and approved by

Dissertation Advisor:
Chien-Cheng (George) Tseng, Sc.D.
Assistant Professor
Department of Biostatistics
Department of Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Eleanor Feingold, Ph.D.
Associate Professor
Department of Human Genetics
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lan Kong, Ph.D.
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Jianhua Luo, M.D., Ph.D.
Associate Professor
Department of Pathology
School of Medicine
University of Pittsburgh

Copyright © by Chunrong Cheng

2008

ENHANCED INTER-STUDY PREDICTION AND BIOMARKER DETECTION IN MICROARRAY WITH APPLICATION TO CANCER STUDIES

Chunrong Cheng, PhD

University of Pittsburgh, 2008

Although microarray technology has been widely applied to the analysis of many malignancies, integrative analyses across multiple studies are rarely investigated, especially for studies of different platforms or studies of different diseases. Difficulties with the technology include issues such as different experimental designs between studies, gene matching, inter-study normalization and disease heterogeneity. This dissertation is motivated by these issues and investigates two aspects of inter-study analysis.

First, we aimed to enhance the inter-study prediction of microarray data from different platforms. Normalization is a critical step for direct inter-study prediction because it applies a prediction model established in one study to data in another study. We found that gene-specific discrepancies in the expression intensity levels across studies often exist even after proper sample-wise normalization, which cause major difficulties in direct inter-study prediction. We proposed a sample-wise normalization followed by a ratio-adjusted gene-wise normalization (SN+rGN) method to solve this issue. Taking into account both binary classification and survival risk predictions, simulation results, as well as applications to three lung cancer data sets and two prostate cancer data sets, showed a significant and robust improvement in our method.

Second, we performed an integrative analysis on the expression profiles of four published studies to detect the common biomarkers among them. The identified predictive biomarkers

achieved high predictive accuracy similar to using whole genome sequence in the within-cancer-type prediction. They also performed superior to the method using whole genome sequences in inter-cancer-type prediction. The results suggest that the compact lists of predictive biomarkers are important in cancer development and represent common signatures of malignancies of multiple cancer types. Pathway analysis revealed important tumorogenesis functional categories.

Our research improved predictions across clinical centers and across diseases and is a necessary step for clinical translation research.

TABLE OF CONTENTS

PREFACE XII

1.0 INTRODUCTION 1

1.1 INTRODUCTION TO MICROARRAYS 1

1.2 OVERVIEW OF PROBLEMS CONSIDERED IN THIS DISSERTATION 5

1.2.1 Normalization in inter-study and inter-platform cross prediction 5

1.2.2 Meta-analysis of different tissues in same microarray platform 7

2.0 RATIO-ADJUSTED GENE-WISE NORMALIZATION TO ENHANCE CLASSIFICATION MODELS FOR INTER-STUDY PREDICTION IN MICROARRAY 11

2.1 ABSTRACT 12

2.2 INTRODUCTION 13

2.3 METHOD 16

2.3.1 Data sets, preprocessing and gene matching 18

2.3.2 Classification methods and evaluation 20

2.3.3 Ratio-adjusted gene-wise normalization 21

2.3.4 Simulation 25

2.3.5 Calibration scheme for perspective studies 26

2.4 RESULTS 27

2.4.1 Simulations to validate ratio-adjusted procedure 27
2.4.2 Inter-study prediction regarding binary classification 29
2.4.3 Inter-study prediction regarding survival risk 32
2.5 DISCUSSION 36
2.5.1 Gene-wise Normalization (GN) 36
2.5.2 Applicability of calibration 36
2.5.3 Application in survival analysis by SuperPC 37
2.5.4 Clinical implication 38
2.5.5 Limitations 39
2.6 CONCLUSIONS 39
3.0 COMMON PREDICTIVE BIOMARKERS AND CROSS-PREDICTABILITY IN THE EXPRESSION PROFILES OF MULTIPLE CANCER TYPES 41
3.1 ABSTRACT 42
3.2 INTRODUCTION 42
3.3 MATERIALS AND METHODS 45
3.3.1 Data and preprocessing 45
3.3.2 Biomarker selection by ANOVA and t-test 46
3.3.3 Gene-specific scaling in inter-cancer-type classification 49
3.3.4 Classification method and leave-one-out cross validation 51
3.3.5 Confusion matrix and prediction index 52
3.3.6 Pathway analysis 53
3.3.7 External evaluation of batchII-PBs by independent prostate data 55
3.4 RESULT 55

3.5 DISCUSSION 62

4.0 CONCLUSION, DISCUSSION AND FUTURE WORK 65

4.1 RATIO-ADJUSTED GENE-WISE NORMALIZATION 66

4.1.1 Conclusion 66

4.1.2 Discussion 66

4.1.3 Future work 68

4.2 COMMON PREDICTIVE BIOMARKERS IN MULTIPLE CANCER TYPES 69

4.2.1 Conclusion 69

4.2.2 Discussion 69

4.2.3 Future work 70

APPENDIX A : SUPPLEMENT MATERIAL OF CHAPTER 2 71

A. 1 CROSS PREDICTION FIGURES FOR LUNG CANCER DATA 71

A. 2 CROSS PREDICTION FIGURES FOR PROSTATE CANCER DATA 74

APPENDIX B : SUPPLEMENT MATERIAL OF CHAPTER 3 76

B. 1 BOOTSTRAP PROCEDURE FOR GENE-WISE NORMALIZATION 76

BIBLIOGRAPHY 82

LIST OF TABLES

Table 2.1 Description of three lung cancer data sets 19

Table 2.2 Mean error rate of 10000 simulations with large sample size 28

Table 3.1 Overview of data sets used in batch I and batch II analyses. 45

Table 3.2 Prediction performance indexes (PPI) in batch I analysis 57

Table 3.3 Prediction performance indexes (PPI) in batch II analysis. 58

Table 3.4 PPI summary of within-cancer-type and inter-cancer-type predictions in batch II analysis. 58

Table 3.5 The 44 batchII-PBs overlapped by pair-wise comparisons of liver, prostate and lung data sets. 59

Table B.1 An example of confusion matrix. 76

Table B.2 Batch I leave-one-out cross validation analysis result (confusion matrix) 77

Table B.3 A Batch II leave-one-out cross validation analysis result (confusion matrix) 78

Table B.4 A List of 109 biomarkers identified in batch I 79

LIST OF FIGURES

Figure 2.1 Predictive biomarker with different expression intensity levels across studies.....15

Figure 2.2 A predictive biomarker with different expression intensity levels across studies: 16

Figure 2.3 A calibration scheme for prospective study.....27

Figure 2.4 Raw Prediction in raw and normalized data from simulation 28

Figure 2.5 Inter-study prediction of lung cancer data by PAM.....30

Figure 2.6 Inter-study prediction of prostate cancer data by PAM.....30

Figure 2.7 Inter-study predictions by SN_std+rGN_std with different number of calibration....32

Figure 2.8 Within-study survival prediction 33

Figure 2.9 Inter-study survival prediction.....35

Figure 2.10 PCA based on unmatched data.....37

Figure 3.1 ANOVA model for batch I analysis.....48

Figure 3.2 Expression patterns of selected representative genes in liver and prostate samples....50

Figure 3.3 Schemes of leave-one-cross-validation or external validation for batchI-PBs and batchII-PBs. 52

Figure 3.4 Pathway analysis heatmap. 54

Figure 3.5 Diagram of batchI-PBs and batchII-PBs and their intersection genes. 61

Figure 3.6 MDS plot of existing training data set and independent prostate cancer data. 62

Figure A.1 Inter-study prediction by LDA 72
Figure A.2 Inter-study prediction by KNN 73
Figure A.3 Inter-study prediction applying SN_std+rGN_std with calibration by LDA 73
Figure A.4 Inter-study prediction applying SN_std+rGN_std with calibration by KNN 74
Figure A.5 Inter-study prediction of two prostate studies by LDA 75
Figure A.6 Inter-study prediction of two prostate studies by KNN 75

PREFACE

First, I would like to give sincere thanks to Dr. George Tseng, the chair of my committee, for mentoring me through the PhD program with insightful guidance, persistent encouragement, tremendous support, and unwavering patience. Without his support academically and financially, I would not have completed this work.

Next, I would like to thank Professors Eleanor Feingold, Lan Kong, and Jianhua Luo for their valuable suggestions and thoughtful comments on the statistical and biological aspects of my research.

I would also like to thank the Department of Biostatistics and the Center for Health Equity Research and Promotion (CHERP), which is affiliated with the VA Pittsburgh Healthcare System for providing financial aid during the first three and half years of my Ph.D. studies. I learned a great deal from my supervisor Dr. Roslyn Stone.

I would like to thank the students in our group for their help and support. The presentations in our group meetings helped me prepare for my oral defense.

Last, but not least, I dedicate this work to my parents and husband.

1.0 INTRODUCTION

1.1 INTRODUCTION TO MICROARRAYS

Microarray analysis is a technology used to simultaneously monitor the mRNA expression level of thousands of genes. (Schena et. al., 1995; Chu *et al*., 1998; Golup *et al*., 1999; Garber *et al*., 2001; Huang *et al*., 2003; Potti *et al*., 2006). This technology is based on two fundamental rationales. First, according to the famous central dogma of molecular biology (Crick, 1970), expressed DNA sequences are transcribed into mRNA before proteins can be synthesized using the information in this mature mRNA as a template. Second, because of the double strand binding property of complimentary DNA sequences, complimentary hybridization can be utilized in microarray technology to measure the expression intensities of the target mRNA.

There are several microarray platforms available, including cDNA, Affemetrix GeneChip, GE (Amersham) Codelink, Illumina, and Agilent. They generally belong to two types of arrays: the two-color cDNA microarray developed in the Brown and Botstein labs at Stanford (DeRisi *et al*., 1997, Eisen *et al*., 1998) and one-color oligonucleotide chips from the Affymetrix Company (Lockhard *et al*., 1996). The cDNA microarray probes are DNA fragments usually amplified by PCR and spotted by a robot on a glass microscope slide. The two complementary DNA samples are obtained from mRNA by reverse transcription and the relative abundance of these DNA sequences is assessed by monitoring the differential hybridization of the two samples to the

probes on the array. The two cDNA samples are labeled with a spectrally distinguishable red (Cy5) or green (Cy3) fluorescent dye. After washing, only the bound strands are kept and the ratio of the relative fluorescence of two dyes will be used as the expression intensity (Schena, 1999).

The probes of Affymetrix GeneChips are short oligonucleotides that are synthesized using a photolithographic approach with a length of 25-mers (Pease *et al*., 1994). Originally, 16-20 probe pairs, each corresponding to a different part of the sequence for a gene, make up a probe set (Bolstad *et al.*, 2003). Now, the number of probe pairs is reduced from 11 to 16. The samples are labeled in one color and have their absolute intensity measured. The intensity information from the value of each of the probes in a probe set are combined to get an expression measure. Detailed Affymetrix microarray technology is described by Lipshutz *et al*. (1999) and Warrington *et al*. (2000).

A lot of comparisons have been done between cDNA array and Affymetrix GeneChip (Yuen *et al*., 2002; Li *et al*., 2002; Woo *et al*., 2004; Park *et al*., 2004; Irizarry *et al.*, 2005; Larkin *et al.*, 2005; Kuo *et al*., 2006; Shi *et al*., 2006). In general, the Affymetrix data have better accuracy and precision and are more reliable for interrogating changes in gene expression than data from long cDNA microarrays. The hybridization of a cDNA array is less sensitive and specific; also the image analysis is more difficult and the data is noisier with poorer reproducibility when compared to the Affymetrix GeneChip (Li *et al*., 2002; Woo *et al*., 2004; Irizarry *et al*., 2005). Affymetrix was proven to perform the best by the MAQC project (Shi *et al*., 2006). However, because it is much cheaper and can be custom made for special species, cDNA array is still a popular tool for researchers.

Since microarrays were first invented in 1995 (Schena et. al., 1995), they have been widely used in biomedical research such as gene expression studies (Brown *et al*., 1999), cancer diagnosis (Golub *et al*., 1999; Yu *et al*., 2004), prognosis prediction (van't Veer *et al*., 2002; Beer *et al*., 2002), and target drug treatment (Shipp, *et al*., 2002; Potti *et al*., 2006). The keyword "microarray" in the public database PubMed yields over 25,000 search results and the number is increasing quickly. We expect that microarray technology will be more useful in the future for disease prediction, diagnosis, prognosis prediction and treatment selection, especially in clinical medicine.

Unlike the traditional data structure in statistics, microarray data have a very high dimension with a relatively small number of samples. Software and packages have been developed to process and analyze microarray data. For example, dChip (Li & Wang, 2001) and MAS (Irizarry, *et al*., 2003) are specific to the data processing and analysis of Affymetrix data. SAM, K-Means and PAM are commonly used R packages for various purposes, such as detecting differentially expressed (DE) genes, clustering, and prediction. With the adoption of machine learning skills, analysis of microarray data is more feasible and powerful. Recently, the trend is to use gene modules as basic building blocks instead of individual genes and several new methods have been proposed (Lamb *et al*., 2003; Huang *et al*., 2003; Segal *et al*., 2004, Segal *et al*., 2005; Tongbai *et al*., 2008). These methods, including pathway analysis and module map methods, help to explore a high-order and more interpretable characterization of transcription changes; also, patterns too subtle to be detected by a single gene can be detected by the large module consisting of many coherent genes (Segal *et al*., 2005). For example, Ingenuity Pathways Analysis (IPA) (Ingenuity Systems, Redwood City, CA) is a popular and powerful analysis system and database for pathway analysis. However, research in microarray studies is still facing several unsolved

problems. For example, with more and more microarray data sets available, the statistical analysis of multiple microarray studies is important, but challenging. This motivated my dissertation work, which aims to solve two problems. The first problem deals with the cross-prediction of studies involving the same disease from different labs. Literature investigating the same disease with different array platforms or in different labs is often reported with similar high disease prediction accuracies without mention of direct inter-study predictions. Many technical difficulties exist when attempting to directly apply prediction models to independent studies. Our objective is to develop an improved and robust normalization method for direct inter-study prediction. Our proposed normalization method dramatically improves the cross prediction's accuracy for studies of any microarray platforms. Chapter 2 focuses on this proposal. The second problem deals with cross-prediction of studies of same platform arrays from different diseases. Multiple microarray studies for the same disease are very common. However, meta-analysis to integrate multiple studies has rarely been investigated. In this project, we performed a meta-analysis on 455 arrays collected from four microarray studies in the Affymetrix U95Av2 platform in order to detect the common predictive biomarkers in the microarray studies of four different cancer types. The identified predictive biomarkers achieved high predictive accuracy similar to using a whole genome in the within-cancer-type prediction. This project is the central concern of chapter 3. In the next section, we will introduce the background of these two projects in details.

1.2 OVERVIEW OF PROBLEMS CONSIDERED IN THIS DISSERTATION

1.2.1 Normalization in inter-study and inter-platform cross prediction

Literature investigating the same disease with different array platforms or in different labs is often reported with similar high disease prediction accuracies. These studies either compared and validated the DE genes or the biomarkers independently found in each study (Tan, *et al*. 2003; Shi *et al*.2006; Mitchell, *et al*. 2004) or evaluated inter-lab or inter-platform concordance by correlation (Parmigiani, *et al*. 2004). The direct application of prediction models to independent studies was, however, rarely investigated. For example, literature investigating lung cancer with different array platforms or in different labs is often reported with similar high disease prediction accuracies without any mention of direct inter-study predictions (i.e., establishing a prediction model from one data set and applying it to another). The major difficulties for such direct inter-study prediction may include: (1) biological differences in the sample population across studies; (2) different sample preparations and experimental protocols; and (3) different microarray platforms (Fishel, *et al*. 2007). Different data preprocessing and incorrect gene matching across studies have also been mentioned as having a great impact on such inter-study analysis (Bosotti, 2007). This creates a barrier for the progress of array technology from beyond bench work to prospective clinical use. Suppose a pilot study or a clinical trial has performed in an old Affymetrix U95 platform and an effective prediction model has been constructed. The test site of another medical center may apply another commercial system (such as Agilent or Illumina platforms) or the original medical center might even migrate to a newer U133 system. The translational research of microarray would not be successful if the prediction model cannot

predict inter-platform or inter-lab studies. This happens often because of the variety of microarray platforms.

Of the difficulties mentioned above, some can be overcome by applying consistent data preprocessing strategies or improving gene matching across studies. UniGene cluster ID has been commonly used as a gene identifier for gene matching (Parmigiani, *et al*. 2004). However, it often happens that a gene can have multiple UniGene IDs, and many genes could share the same UniGene ID. All this adds to the difficulty of gene matching. Entrez Gene ID turns out to be more unique and becomes more recognizable to researchers (Maglott, *et al.* 2005; Bosotti, *et al*. 2007). In the other extreme, if the intrinsic difference across studies is indeed large due to a differential sample population or to experimental protocols as mentioned in (1) and (2) above, the direct inter-study prediction will never be valid. Inter-study normalization is a critical step for cross prediction because it can bring the gene expression intensities from different studies to a comparable level. In section 2.2, we will investigate a commonly encountered situation in which gene-specific discrepancies in expression intensity levels across studies are found even after proper sample-wise normalization. We will compare the intensity levels based on the raw data and the data after various normalizations for a given gene across three studies. The gene-specific discrepancies often come from differential probe sequence selections and experimental protocols in different array platforms that cause different gene-specific hybridization efficiencies across studies.

In section 2.3, we will introduce and compare the current sample-wise normalization methods and how most of the time they are not sufficient for inter-study predication. To solve the gene-specific discrepancies, we proposed a sample-wise normalization followed by ratio-adjusted gene-wise normalization (SN+rGN) method. The classification information for the test

data is unknown, but is necessary for the ratio-adjustment gene-wise normalization. Thus we introduced the calibration idea to get away from this problem. The calibration scheme for a prospective study will also be detailed in this section.

In section 2.4, we will show the inter-study prediction results after applying our proposed method to three lung cancer and two prostate cancer data sets, considering both binary classification and survival risk predictions. Two of the three lung cancer studies are from the Affymetrix data and the third is from the cDNA data. One of the two prostate cancer sets is from the Affymetrix data and the other is from the cDNA data. Our prediction tools for binary classification include PAM, LDA, and KNN. SuperPC was used for survival prediction. Our proposed SN+rGN normalization method yielded significant and robust improvement for the inter-study predictions compared to the sample-wise normalization.

In the last part of chapter 2, we will discuss our proposed method and its application. We will show the strengths and limitations of rGN and how it can be used in clinical studies. Finally, we will draw conclusions from this study.

1.2.2 Meta-analysis of different tissues in same microarray platform

Instead of studying data of different platforms from the same disease as above, we focused on meta-analysis of different diseases from the same platform. Human malignancies can occur in almost all organs, with the exception of several accessory sex organs. Most human malignancies, regardless of the origins of the tissues, contain two major characteristics: uncontrolled growth and the ability to metastasize. Abnormalities of the same signaling pathways can be found in multiple types of human cancers; a tumor may contain multiple abnormalities in signaling.

Overlapping these abnormalities among multiple types of tumors may shed light on some key alterations in carcinogenesis.

Prostate cancer is second only to skin cancer as the most commonly diagnosed malignancy in American men. Some studies suggested that up to 80% of men older than eighty were found to contain pathologically recognizable prostate cancer, while any man younger than forty rarely developed the same disease. This argues against any singular specific etiology responsible for prostate cancer besides aging. The clinical courses of most prostate cancers are long, and some are life-threatening. Hepatocellular carcinoma, on the other hand, is quite the opposite. It is not age related, and is tightly linked to cancer etiologies such as alcohol, the hepatitis B or C virus, and certain toxins. Hepatocellular carcinoma is distinctive in its well confined nodular architecture. The clinical courses of most of the hepatocellular carcinomas are short and the fatality is high. Most of the lung cancers, with the exception of small cell carcinoma, are also associated with distinctive etiologies, such as smoking or chronic exposure to certain type of carcinogens. The urothelial carcinoma of the urinary bladder, however, is primarily idiopathic or viral. Since these four types of cancer are so far apart in etiology, morphology and clinical courses, any common ground between these tumors could be interpreted as the likely common pathway of carcinogenesis.

Microarray technology has been widely applied to the analysis of many malignancies, including the four cancer types mentioned in the literature above. However, using meta-analysis to integrate multiple studies has rarely been investigated. In this project, we performed a meta-analysis on 455 arrays collected from four microarray studies in the Affymetrix U95Av2 platform: 94 samples of liver tissue (Luo, *et al*. 2006), 148 samples of prostate tissues (Yu, *et al*. 2004), 151 samples of lung tissues (Bhattacharjee, *et al*. 2001) and 62 urinary bladder tissues

(Stransky, *et al*. 2006). The liver and prostate samples contain three types of tissues: organ donor (N), adjacent to tumor (A) and tumor (T), while the lung and bladder samples only contain N and T. Detailed background introduction of this project is included in section 3.2.

In section 3.3, the materials and methods used in our project will be described. We will focus on the filtering criteria for data preprocessing. We performed an analysis on two batches of samples. In batch I, all three tissue types in the liver and prostate samples were analyzed using the analysis of variance (ANOVA) model. In batch II, the normal and tumor tissues in all four cancer types were included and a t-test was used to identify predictive biomarkers. Gene-specific scaling is used in inter-cancer-type classification. Leave-one-out cross validation is conducted by PAM. Figure 3.3 describes this leave-one-out cross validation scheme. The prediction performance is measured by a confusion matrix and the prediction performance index (PPI), which is defined as the average of sensitivity and specificity. The gene ontology (GO) database was used for pathway enrichment analysis. Finally, a set of twenty-three external prostate cancer samples was used for external validation of the batchII-PBs.

In section 3.4, the results section, we will show that the identified biomarkers have high predictability in both within-cancer-type (cross validation within a single cancer type) and inter-cancer-type (i.e., a prediction model trained in one cancer type and used to predict another cancer type) prediction via leave-one-out cross validation. Further, pathway enrichment analysis identified statistically significant function categories of the biomarkers. Validation of the 47 batch II predictive biomarkers on independent twenty-three prostate tissues yielded 96% accuracy in inter-study prediction from the original prostate, liver, and lung cancer data sets respectively, showing the robustness of the predictive biomarkers and their implications for common carcinogenesis of multiple cancer types.

In section 3.5, the forty-four batchI-PB genes will be investigated by their functions. We will discuss the two-fold clinical implication of our findings: therapeutic targeting toward some of these 44 genes will be of significant value in treating these malignancies since the prediction of liver, lung, and prostate cancer using our forty-four batchI-PBs is interchangeable. Second, the ninety-nine batchII-PB model may be able to serve as a predictor of malignancies nearby even if a biopsy misses its tumor target.

2.0 RATIO-ADJUSTED GENE-WISE NORMALIZATION TO ENHANCE CLASSIFICATION MODELS FOR INTER-STUDY PREDICTION IN MICROARRAY

Manuscript submitted.

Chunrong Cheng[1], Jian-Hua Luo[*,2]

George C. Tseng[1,3],

[1]Department of Biostatistics, University of Pittsburgh

[2]Department of Pathology, University of Pittsburgh

[3]Department of Human Genetics, University of Pittsburgh

Email: ctseng@pitt.edu, Tel: 412-624-5318

2.1 ABSTRACT

Motivation: Reproducibility of microarray experiments has been greatly improved in the past decade and its application in biomedical research is more and more prevalent. Independent studies investigating an identical disease with different array platforms and from different labs are often found in journals. Consistent high prediction accuracies are often reported in the individual studies; however, direct inter-study prediction by applying a prediction model established in one study to another usually generates poor performance.

Results: We found that gene-specific discrepancies in the expression intensity levels across studies often exist even after proper sample-wise normalization, which cause a major difficulty in direct inter-study prediction. We proposed a ratio-adjusted sample-wise normalization followed by gene-wise normalization (SN+rGN) method to solve this problem. The proposed method significantly increased the prediction accuracies when such inter-study discrepancies existed in the predictive biomarkers. The ratio of sample sizes in normal versus diseased groups could affect the performance of gene-wise normalization and an analytical method was developed to adjust for the imbalanced ratio effect. Both simulation results and applications to three lung cancer and two prostate cancer data sets, considering both binary classification and survival risk predictions, showed significant and robust improvement of our method. A calibration scheme was developed to apply our method to prospective clinical trials. The number of calibration samples needed was estimated from existing studies and suggested for future applications.

2.2 INTRODUCTION

Microarray technology has been widely used in biomedical research, for example, in prediction of cancer diagnosis (Golub, *et al*. 1999), of prognosis (van't Veer, *et al*. 2002) and of treatment outcome (Shipp, *et al*. 2002) using supervised machine learning approaches. With an increasing amount of microarray data sets available, reproducibility analysis of these independent experiments has gained more attention and has been greatly improved in the past decade. Tan (2003) and his colleagues examined three microarray platforms and showed that the correlations of detected biomarkers were moderate to poor. This finding, together with other negative studies, stimulated more researchers to investigate improvement (Kuo, *et al*. 2006). Recently, the MicroArray Quality Control (MAQC) project was initiated to address this concern and they concluded that they found satisfying intra-platform consistency across test sites as well as a high level of inter-platform concordance in terms of genes identified as differentially expressed (DE) have been reached in most popular commercial platforms (Shi, *et al*. 2006). Yauk and Berndt (2007) reviewed these studies and concluded that with improvements in the technology (the optimization and standardization of methods, including data analysis) and annotation, analysis across platforms yields highly correlated and reproducible results.

Most reproducibility studies in literature either compared and validated the DE genes or the biomarkers independently found in each study (Tan, *et al*. 2003; Shi *et al*.2006; Mitchell, *et al*. 2004) or evaluated inter-lab or inter-platform concordance by correlation (Parmigiani, *et al*. 2004). Direct application of prediction models to independent studies was, however, rarely investigated. For example, literature investigating lung cancer with different array platforms or in different labs is often reported with similar high disease prediction accuracies without mention of direct inter-study predictions (i.e., establishing a prediction model from one data set and applying

it another). This creates a barrier for the progress of array technology beyond bench work to prospective clinical use. Suppose a pilot study or clinical trial has performed in an old Affymetrix U95 platform and an effective prediction model has been constructed. The test site of another medical center may apply another commercial system (such as Agilent or Illumina platforms) or even the original medical center may migrate to a newer U133 system. The translational research of microarray would not be successful if the prediction model cannot predict inter-platform or inter-lab studies. This happens often because of the variety of microarray platforms. The major difficulties for such direct inter-study prediction may include: (1) biological differences in the sample population across studies (2) different sample preparations and experimental protocols and (3) different microarray platforms (Fishel, *et al*. 2007). Different data preprocessing and incorrect gene matching across studies have also been mentioned to as having a great impact on such an inter-study analysis (Bosotti, 2007) and some practical guidelines have been suggested.

Among the difficulties mentioned above, some can be overcome by applying consistent data preprocessing strategies or improving gene matching across studies. UniGene cluster ID has been commonly used as a gene identifier for gene matching (Parmigiani, *et al*. 2004) , but it often happens that a gene can have multiple UniGene IDs, and many genes could share the same UniGene ID, which adds to the difficulty of gene matching. Entrez Gene ID turns out to be more unique and becomes more familiar to researchers (Bosotti, *et al*. 2007). In the other extreme, if the intrinsic difference across studies is indeed large due to a differential sample population or experimental protocols as mentioned in the first two points above, the direct inter-study prediction will never be valid. In this chapter of my dissertation, we investigate a commonly encountered situation – that the gene-specific discrepancies in the expression intensity levels

across studies are found even after proper sample-wise normalization. The gene-specific discrepancies often come from differential probe sequence selections and experimental protocols that caused different gene-specific hybridization efficiencies across studies. It is easily seen from the figure by Kuo *et al* (2006) which displayed the probe matching of various microarray platforms compared to the complete probe sequence within one exon of Gas1. Based on this figure (Figure 2.1 in this dissertation), the sequences of different platforms are quite different. An extreme example exists between Affymetrix and Academic cDNA where there is no overlapping at all. In Figure 2.2A, for example, EMP2 is an ideal predictive biomarker that is down-expressed in the diseased group in the raw data (data with intra-study normalization but without inter-study normalization) of all three independent lung cancer studies (details of the data sets will be introduced later). The absolute intensities in the three studies are, however, at very different level. Direct applications of prediction models across studies using this biomarker will perform poorly in this case. For example, applying the prediction model from the Harvard study to the Michigan study will predict all subjects as cancer patients.

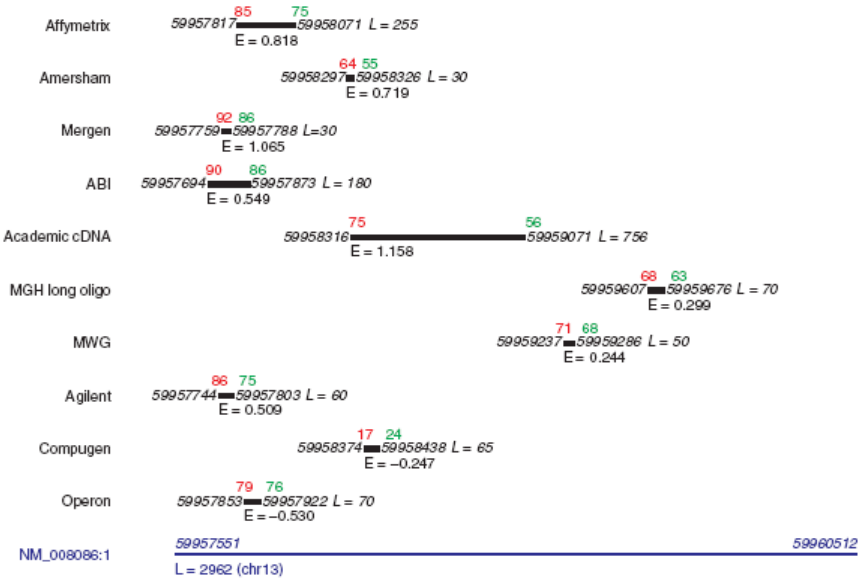

Figure 2.1 Predictive biomarker with different expression intensity levels across studies

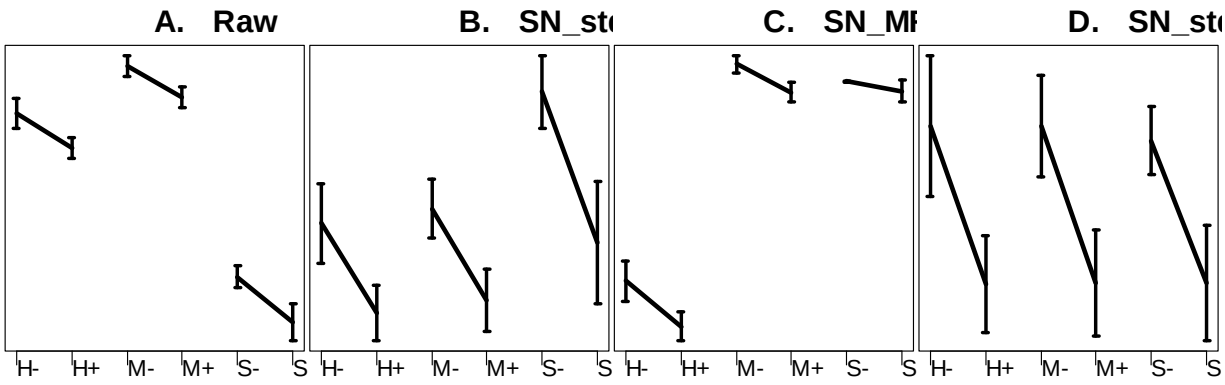

Figure 2.2 A predictive biomarker with different expression intensity levels across studies:

Figure. 2.2A: raw data before any inter-study normalization. Figure 2.2B: inter-study sample-wise normalization with standard normalization(SN_std). Figure 2.2C: inter-study sample-wise normalization with median rank scores (SN_MRS). Figure 2.2D: inter-study sample-wise normalization and then gene-wise normalization with standard normalization (SN_std+GN_std). H-: normal group for Harvard study. H+: tumor group for Harvard study. M: Michigan study; S: Stanford study. Means and error bars of expression intensities are represented on the y-axis.

2.3 METHOD

Inter-study normalization, which ensures the training data and the test data at a comparable expression level, is a critical step for direct inter-study prediction. Global sample-wise normalization methods such as standard normalization (standardizing each sample vector to zero mean and unit variance), Loess (Yang *et al*. 2002), rank invariant normalization (Tseng, *et al*. 2001), and quantile normalization (Irizarry, 2003) are popular normalization methods for intra-study normalization. They are usually effective to scale expression intensities of samples to a comparable level in a global genomic sense. Warnat *et al*. (2005) applied two intra-study sample-wise normalization methods – median rank scores (MRS) and quantile discretizations (QD) – to an inter-study normalization of three pairs of Affymetrix versus cDNA microarray studies and concluded that the two sample-wise normalization methods facilitated successful inter-study prediction in two out of the three pairs of studies. Whether or when the sample-wise

normalization methods are applicable to other inter-study predictions was, however, not discussed. In Shabalin *et al.*'s (2008) approach, data from independent studies was sample standardized and gene median centered first,, and then was combine K-means clustering on both the column and row level for parameter estimation of the block liner model. The purpose of their study was to directly merge different data sets. They did not demonstrate the magnitude of improvement that the sophisticated clock linear model could provide in addition to simple sample standardization and gene median centering, and the selection of cluster size was somewhat arbitrary. In Figures 2.2B and 2.2C standard normalization (std) and MRS are applied for inter-study sample-wise normalization. It is clearly seen that the expression levels of the gene EMP2 across studies are still not comparable and that the three pairs of direct inter-study prediction (Harvard vs Stanford, Stanford vs Michigan and Harvard vs Michigan) will fail with an average accuracy rate of 73% for std and 61% for MRS. In this paper, we proposed a sample-wise normalization followed by gene-wise normalization as a necessary step for many inter-study predictions to succeed. For example, in Figure 2.2D, the expression intensities after sample-wise and gene-wise normalization by simple standardization to zero mean and unit variance (SN_std+GN_std) will be scaled to a comparable range and an average accuracy of the three pairs of inter-study predictions can reach 94%, a magnitude similar to the accuracy level of a within-study cross validation accuracy. In this paper, we will examine properties of sample-wise normalization followed by gene-wise normalization (SN+GN). We will demonstrate through simulation and real applications that the ratio between normal versus tumor samples affects the performance of gene-wise normalization. A ratio-adjusted gene-wise normalization method is thus proposed (SN+rGN). Through the remainder of this paper, we will denote the sample-wise normalization by SN, the gene-wise normalization by GN, and the ratio-adjusted

gene-wise normalization by rGN. Since rGN requires the knowledge of sample labels (normal or tumor) to calculate the sample-size ratio, it cannot be directly applied to a prospective clinical trial. To address this issue, we proposed a calibration scheme, which allocated a small number of samples with known classification labels in the new study as a calibration set for estimating the normalization factors across studies. The concept is similar to the common practice of performing a few test samples with known classification to calibrate an experimental instrument when it is installed in a new lab environment or when it is operated by a new technician.

2.3.1 Data sets, preprocessing and gene matching

Pair-wise inter-study predictions were performed on three lung cancer data sets, including normal and adenocardinoma samples from each study. Two of these studies, named Harvard (Bhattacharjee, *et al*., 2001) and Michigan (Beer, *et al*., 2002), used Affymetrix oligonucleotide arrays Hu95a and HG6800 respectively. The third one, named Stanford (Garber, *et al*., 2001), used the cDNA platform. The raw data of these three lung cancer studies were downloaded from the public internet domain (http://www.camda.duke.edu/camda03/datasets/). Intra-study sample-normalization for Harvard and Michigan was carried out in dChip using the invariant-set normalization. Standard normalization which standardizes each sample to zero mean and unit variance was applied to the cDNA data. Genes with low average intensities or small variability were filtered out based on the criteria developed in the original studies. Detailed information is listed in Table 2.1.

Table 2.1 Description of three lung cancer data sets

	Harvard	Michigan	Stanford
Array platform	Affymetrix U95A	Affymetrix HG6800	cDNA
# of samples (normal, adenocarcinoma)	(17, 134)	(10, 86)	(5, 39)
# of original transcripts	12,625	7,129	24,192
# after filtering	4,861	5,119	4,123
# with Entrez gene ID	4,756	5,053	3,774
# after averaging probes with same EntrezID	4,107	4,467	3,399
# of overlapping genes in pair-wise studies	Harvard-Michigan: 2,493 Harvard-Stanford: 1,493 Michigan-Stanford: 1,594		

For gene matching across studies, Entrez IDs were used as the common identifiers. Several web-based gene annotation conversion tools are available, such as SOURCE (Diehn *et al*., 2003), MatchMiner (Bussey and Sunshine, 2003), and DAVID (Dennis *et al*, 2003). In this paper, R package "annotate" was used to retrieve the Entrez IDs for the two Affymetrix data sets and MatchMiner was used for the cDNA data. Averaged values were taken for multiple probes sharing an identical Entrez ID. Table 2.1 provides detailed descriptions. There were 2,493 genes that overlapped in the Harvard and Michigan data set, 1,493 genes that overlapped in the Harvard and Stanford data sets and 1,594 genes that overlapped in the Michigan and Stanford data sets that were used for the analysis of direct inter-study prediction analysis in this paper. There were 81, 86 and 22 tumor samples respectively in the Harvard, Michigan and Stanford data sets with available survival follow-up information. These samples were used in the survival risk prediction analysis.

The two prostate cancer sets, the Affymetrix U95a data: Welsh (Welsh *et al*., 2001) and cDNA data: Dhanasekaran (Dhanasekaran *et al*., 2001), were publicly available. The Welsh data set contains nine normal and twenty-five cancer samples while the Dhanasekaran has nineteen

and forteen respectively. Since these two data sets were pre-processed with only the Unigene ID provided, they were merged by Unigene ID. There were 3,078 genes left, which were the basis of the inter-study prediction analysis used in our project.

2.3.2 Classification methods and evaluation

To assess the performance of our proposed normalization method regarding binary classification (the prediction of normal versus cancer samples), we examined three popular classification methods in microarray analysis: Linear Discriminant Analysis (LDA), K-nearest neighbor (KNN), and Prediction Analysis of Microarrays (PAM). For ease of evaluation, comparisons of different normalization methods were performed with identical parameters (number of genes used in LDA, PAM and KNN) and the results were verified by varying parameters over a certain range.

Overall prediction accuracy has been widely used as the evaluation index in many publications. It is, however, often a misleading measure, especially when the data set contains unbalanced sample sizes in groups. For example, the accuracy in the Michigan data set can be as high as 89.6% (86/96), even if the classification rule predicts all samples to be adenocarcinomas. A standard alternative to this situation may be the AUC (area under ROC curve) index by varying the classification threshold in the classification rule. This measure is, however, not readily available for classical methods like KNN. Even for methods that can calculate AUC, the measure is very unstable for small sample-size situations. In this paper, we applied a simple but robust prediction performance index (PPI) that is defined as the average of sensitivity and specificity of the prediction results.

To evaluate risk prediction of survival time, Supervised Principle Components (SuperPC) (Bair & Tibshirani, 2004) method was used to cross predict the survival risks of the three lung cancer data sets. The univariate Cox model coefficient fitting the expression intensities to the survival is calculated for each gene. The most significant fifty genes are kept and singular value decomposition (SVD) is applied to select the top three principal components for fitting the Cox linear model. The risk index is defined as the linear term in the Cox model and the median risk index value of all the training samples is used as the threshold for deciding high or low risk groups. Under this criterion, about half of the patients will be classified as the high risk group and the other half will be classified as the low risk group. Leave-one-out cross validation is used in the training set. We will perform cross validation for the risk prediction based on the training model. The performance of risk prediction is determined by the separation of survival curves of high and low risk groups, which is often evaluated by the p-value and chi-square statistics of a log-rank test comparing the two Kaplan-Meier curves from predicted high and low risk groups.

2.3.3 Ratio-adjusted gene-wise normalization

Figure 2.2 shows that inter-study normalization is necessary before being able to perform inter-study prediction (2. 2A) and sample-wise normalization across studies is not sufficient to correct the bias (Figure 2.2B and 2.2C). There are several potential reasons for such gene-specific intensity discrepancies. The major cause comes from different probe design in different array manufacturings. For example, the probes from Affymetrix GeneChip are short 25-mer oligos with multiple probes (11-16 probes) representing one gene. For cDNA microarrays, the probes are cDNA fragments that are usually hundreds of bases long. As a result, probes meant to measure an identical gene always have different target sequences for hybridization in different

platforms, which, in turn, introduces differential probe efficiency and affects the final intensity levels. Even if comparing studies of the same array platform, different sample preparations, as well as labelling and hybridization protocols, can possibly introduce such gene-specific intensity discrepancies.

We proposed applying sample-wise standard normalization followed by gene-wise standard normalization (SN_std+GN_std) to alleviate the gene-specific discrepancies described above. Preprocessing by GN_std has been widely applied prior to gene clustering, to particular classification methods (SVM, KNN etc), and to dimension reduction (MDS and PCA) to obtain better scale invariant property. It is well-known that by performing GN_std, the Euclidian distance of two genes can be expressed by the correlation coefficient. Suppose $x = (x_1,..., x_{n_1})$ and $y = (y_1,...y_{n_2})$ are the expression values of two genes in a data set. The correlation coefficient and Euclidian distance are defined respectively as follows:

$$r(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad \text{and} \quad d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

After applying gene-wise normalization (GN_std), $x_i' = \frac{x_i - \bar{x}}{s_x}$ and $y_i' = \frac{y_i - \bar{y}}{s_y}$. We can express the Euclidian distance as $d(x', y') = \sqrt{2(n-1)(1 - r(x', y'))}$.

We observed that the ratios of normal and adenocarcinoma samples in the three lung cancer data sets were very close. Intuitively, GN_std is sensitive to the sample ratio between normal and diseased groups in a study. This motivated us to explore the role of the ratio played in GN_std. Simulation supports that GN_std is sensitive to the ratio imbalance of the training and test sample sizes. We proposed the following analytic approach for ratio adjustment by assuming an equal mixture in the paragraph below.

For a given gene g, we omit the subscript g and consider observed intensities $(x_1,...,x_{n_1},y_1,...y_{n_2})$ after sample-wise intra-study normalization, where the first n_1 samples are from the normal group and the next n_2 samples are from tumor patients. Suppose $(x_1,...,x_{n_1})$ are i.i.d. from distribution X and $(y_1,...y_{n_2})$ are from Y and $E(X)=u_X$, $Var(X)=\sigma_X^2$, $E(Y)=u_Y$, $Var(Y)=\sigma_Y^2$. GN_std standardizes gene vector to a mean of 0 and a standard deviation of 1 by

$$x_i^{(GN\_std)}=\frac{x_i-a_{GN\_std}}{b_{GN\_std}} \text{ and } y_i^{(GN\_std)}=\frac{y_i-a_{GN\_std}}{b_{GN\_std}},$$

where $\hat{a}_{GN\_std}=\frac{1}{n_1+n_2}(\sum x_i+\sum y_i)\approx\frac{1}{n_1+n_2}(n_1\cdot u_X+n_2\cdot u_Y)$ (1)

and

$$\hat{b}^2_{GN\_std}=\frac{1}{n_1+n_2-1}(\sum(x_i-a_{GN\_std})^2+\sum(y_i-a_{GN\_std})^2)$$

$$=\frac{1}{n_1+n_2-1}(\sum(x_i-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2+\sum(y_i-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2)$$

$$=\frac{1}{n_1+n_2-1}(\sum(x_i-\frac{\sum x_i}{n_1}+\frac{\sum x_i}{n_1}-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2+\sum(y_i-\frac{\sum y_i}{n_2}+\frac{\sum y_i}{n_2}-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2)$$

$$=\frac{1}{n_1+n_2-1}(\sum(x_i-\frac{\sum x_i}{n_1})^2+\sum(\frac{\sum x_i}{n_1}-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2$$

$$+\sum(y_i-\frac{\sum y_i}{n_2})^2+\sum(\frac{\sum y_i}{n_2}-\frac{\sum x_i+\sum y_i}{n_1+n_2})^2)$$

$$\approx\frac{1}{n_1+n_2-1}((n_1-1)\cdot\sigma_X^2+\sum(\frac{n_1\cdot\sum x_i+n_2\cdot\sum x_i-n_1\cdot\sum x_i-n_1\cdot\sum y_i}{n_1\cdot(n_1+n_2)})^2$$

$$(n_2-1)\cdot\sigma_Y^2+\sum(\frac{n_1\cdot\sum y_i+n_2\cdot\sum y_i-n_2\cdot\sum x_i-n_2\cdot\sum y_i}{n_2\cdot(n_1+n_2)})^2)$$

$$\approx \frac{1}{n_1+n_2-1}((n_1-1)\cdot\sigma_X^2+\sum(\frac{n_1\cdot n_2\cdot u_X-n_1\cdot n_2\cdot u_Y}{n_1\cdot(n_1+n_2)})^2+(n_2-1)\cdot\sigma_Y^2+\sum(\frac{n_1\cdot n_2\cdot u_Y-n_1\cdot n_2\cdot u_X}{n_2\cdot(n_1+n_2)})^2)$$

$$=\frac{1}{n_1+n_2-1}((n_1-1)\cdot\sigma_X^2+\frac{n_2^2}{(n_1+n_2)^2}\cdot\sum(u_x-u_y)^2+(n_2-1)\cdot\sigma_Y^2+\frac{n_1^2}{(n_1+n_2)^2}\cdot\sum(u_x-u_y)^2)$$

$$=\frac{(n_1-1)\cdot\sigma_X^2+(n_2-1)\cdot\sigma_Y^2+\frac{n_1\cdot n_2}{n_1+n_2}(u_x-u_y)^2}{n_1+n_2-1}$$

It is clearly seen that results of GN_std greatly depend on the sample sizes n_1 and n_2. We propose below a ratio-adjusted gene-wise normalization (rGN_std). The empirical distribution obtained from $(x_1,...,x_{n_1})$ and $(y_1,...y_{n_2})$ are denoted by X' and Y'. In other words, the cdf of X' is $F_{X'}(t)=\frac{1}{n_1}\sum_{i=1}^{n_1}I(x_i\le t)$ and similarly $F_{Y'}(t)=\frac{1}{n_2}\sum_{i=1}^{n_2}I(y_i\le t)$.

Consider Z' the mixture distribution of X' and Y', and $F_{Z'}(t)=0.5\cdot F_{X'}(t)+0.5\cdot F_{Y'}(t)$. Our goal is to find normalization factors a_{rGN_std} and b_{rGN_std} such that $E(\frac{Z'-a_{rGN\_std}}{b_{rGN\_std}})=0$ and $Var(\frac{Z'-a_{rGN\_std}}{b_{rGN\_std}})=1$.

From the derivation of a_{GN_std} and $b_{GN\_std}^2$, we can easily obtain that $\hat{a}_{rGN\_std}=(\hat{u}_X+\hat{u}_Y)/2\approx(u_X+u_Y)/2$ and

$$\hat{b}_{rGN\_std}^2=\frac{\hat{\sigma}_X^2+\hat{\sigma}_Y^2}{2}+(\frac{\hat{u}_X-\hat{u}_Y}{2})^2\approx\frac{\sigma_X^2+\sigma_Y^2}{2}+(\frac{u_X-u_Y}{2})^2$$

, where $\hat{u}_X=mean(x_1,...x_{n_1})$, $\hat{\sigma}_X^2=\mathrm{var}(x_1,...x_{n_1})$, with a similar derivation for $\hat{u}_Y$ and $\hat{\sigma}_Y^2$. The two ratio-adjusted scaling parameters are invariant to n_1 and n_2.

This analytic approach for ratio-adjustment could be easily extended when there are more than two groups in the data. Suppose K groups are available, we can generate Z' to be the mixture of all the groups with equal weight:

$$F_{Z'}(t) = \frac{1}{K}\sum_{k=1}^{K} F_{X'_k}(t)$$

The scaling parameters can be derived similarly:

$$\hat{a}_{rGN\_std} = \frac{1}{K}\sum_{k=1}^{K}\hat{u}_{X_k} \text{ and } \hat{b}^2{}_{rGN\_std} = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}^2_{X_k} + \frac{1}{K}\sum_{k=1}^{K}(\hat{u}_{X_k} - a_{rGN\_std})^2$$

2.3.4 Simulation

We performed simulations under two scenarios to examine the effects of the sample ratio on normal and diseased groups in the performance of gene-wise normalization for prediction. In scenario one, the ratios in the training and test set are equal (1:1), while the ratios are opposite in scenario two: 3:1 in the training and 1:3 in the test set. In scenario one, we simulated a ratio-balanced univariate gene scenario for the training data and for the test data. Expression intensities for 100 normal samples were simulated from *N(3.5, 1)* and 100 tumor samples were simulated from *N(6.5, 1)*. In the test data, we assumed that the hybridization efficiency was doubled and the expression intensities of 100 normal samples were simulated from *N(7, 22)* and 100 tumors were simulated from *N(13, 22)*. In scenario two, a ratio-imbalance scenario, the distributions remained the same but training data contain 150 normal and 50 tumor samples while test data contained 50 normal and 150 tumor samples. A univariate (one marker) prediction model was constructed from the training data using linear discriminant analysis

(LDA) and then was evaluated in the test data. The simulation was performed 1,000 times and the average error rate was reported. The performance of no GN (raw data), GN_std, rGN_std and the optimal Bayes error rate are evaluated. The Bayes error rates based on the Bayes optimal classifier given the underlying simulation model were calculated for both scenarios. Specifically $Error_{Bayes}(X, Y, P_x, P_y) = P_X \cdot \int_{t>\lambda} f_X(t)dt + P_Y \cdot \int_{t<\lambda} f_Y(t)dt$ if $E(X)<E(Y)$ and λ is the solution to $P_X \cdot f_X(\lambda) = P_Y \cdot f_Y(\lambda)$. In scenario 1, $P_X=P_Y=0.5$; in scenario 2, $P_X=0.25$ and $P_Y=0.75$.

2.3.5 Calibration scheme for perspective studies

SN_std+rGN_std cannot be applied directly to a test data set if the sample labels are unknown. In order to estimate the a_{GN_std} and $b^2_{GN\_std}$ for each gene in the test data in a prospective study, a small data set including both normal and disease samples are needed to serve as calibration sample. Figure 2.3 describes a calibration scheme for applying the proposed SN+rGN method to construct a prediction model from an existing training study and to perform prediction in a prospective test study. SN_std+rGN_std is performed in the training study and a classification model is obtained. In the prospective test study, a small set of calibration samples with known disease labels is available and SN_std+rGN_std is similarly applied to estimate the normalization factors. Finally, the normalization factors obtained from the calibration set and the classification model obtained from the training study are applied to all prospective test samples to generate the final prediction.

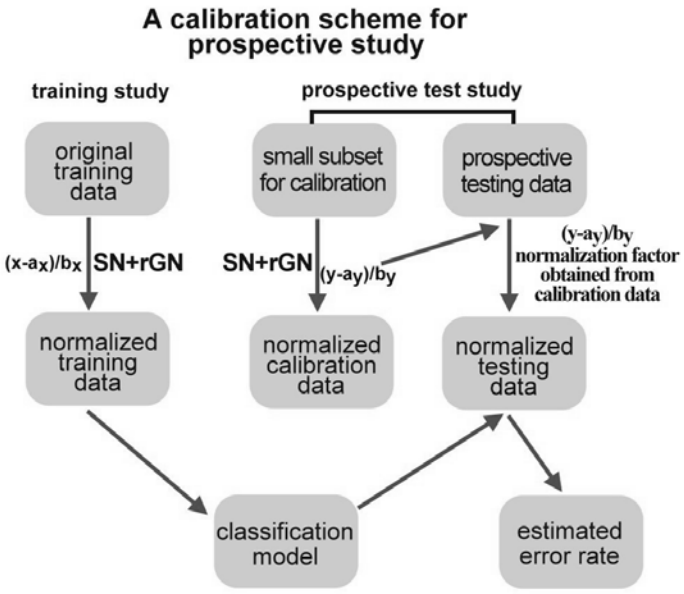

Figure 2.3 A calibration scheme for prospective study

2.4 RESULTS

2.4.1 Simulations to validate ratio-adjusted procedure

The raw and normalized data from simulation under both scenarios are displayed in Figure 2.4. By comparing Figures 2.4C and 2.4D, GN_std normalization result in near optimal prediction under scenario 1, but not under scenario 2. The ratio-imbalanced situation was corrected by rGN_std (Figure 2.4F). Table 2.2 lists the mean error rate of constructing a prediction model from the training data and predicting the test data (by LDA) using raw data, and data after applying either the GN_std or rGN_std methods based on 1,000 simulations. The optimal Bayes errors based on the simulation distribution assumptions were calculated for reference. In scenario 1, the error rates of applying GN_std and rGN_std are identical and very close to the Bayes error

rate. In scenario 2, GN_std does not perform well due to imbalanced sample ratios and applying rGN_std greatly improves GN_std.

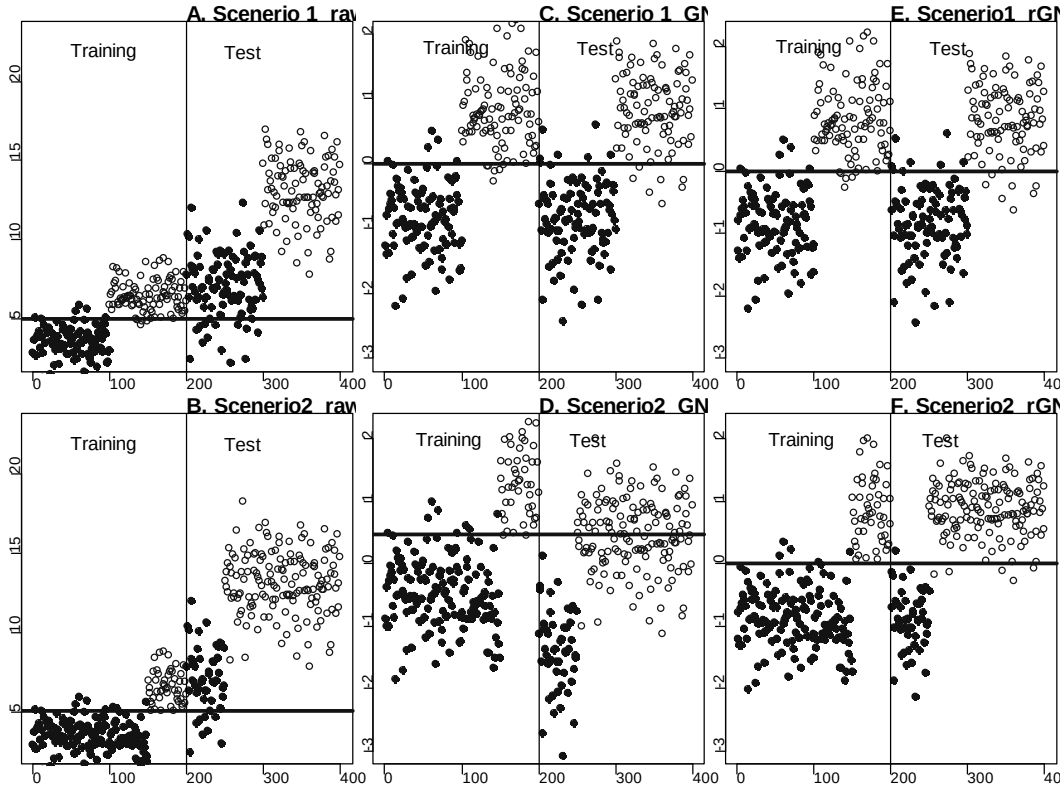

Figure 2.4 Raw Prediction in raw and normalized data from simulation

Figure 2.4A (scenerio1, raw data), Figure 2.4B (scenerio2, raw data), Figure 2.4C (scenerio1, GN_std), Figure 2.4D (scenerio2, GN_std), Figure 2.4E (scenerio1, rGN_std), Figure 2.4F (scenerio2, rGN_std),
Solid dots: normal samples, circles: tumor samples
Solid horizontal line: prediction threshold from the training set, used to predict the test set

Table 2.2 Mean error rate of 1,000 simulations

Error rate	Training => Test	
	Scenario 1	Scenario 2
Raw	42.0%	21.0%
GN_std	6.7%	37.5%
rGN_std	6.7%	6.6%
Bayes error	6.7%	5.5%

2.4.2 Inter-study prediction regarding binary classification

We compared inter-study performance using Raw, as well as data normalized by the following methods: SN_std, SN_std+GN_std, and SN_std+rGN_std to three lung cancer data sets: Harvard, Michigan and Stanford. The with-in study prediction in each individual study had nearly perfect performance in either SN_Std or SN_std+GN_std, confirming previous reports. We tested three prediction methods: PAM, LDA and KNN. The results were slightly different but similar and we only report PAM results here. The results by LDA and KNN are presented in the Appendix A. Figure 2.5 shows the PPIs of all three pair-wise inter-study predictions with 5, 10, 50, 100, 200 and 500 genes used in the training set. The lines 1, 2, 3, and 4 stand for raw data, SN_std, SN_std+GN_std, and SN_std+rGN_std respectively. Without any inter-study normalization (Raw), all the inter-study predictions performed poorly with PPIs around 50%, which is a result of almost all the samples being predicted into one group (similar to the univariate situation in Figure 2.2A). SN_std dramatically improved the inter-platform prediction between Harvard and Michigan, two Affymetrix data sets (Figure 2.5A and 2.5B). It also improved when predicting Stanford study (Figures 2.5C and 2.5E) but not using Stanford to predict the other two studies (Figures 2.5D and 2.5F). Overall, SN_std+GN_std improved SN_std and the results after applying SN_std+rGN_std performed the best for inter-study prediction. For example, in Figure 3F, SN_std+rGN_std was the only normalization method which produced good inter-study prediction.

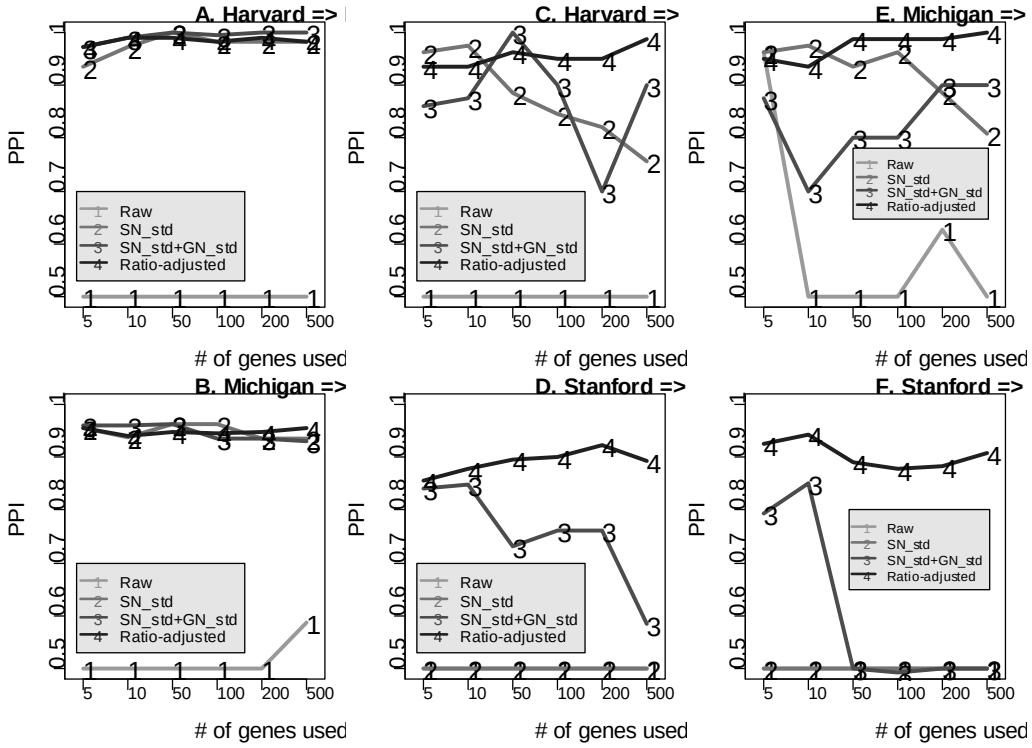

Figure 2.5 Inter-study prediction of lung cancer data by PAM

We performed the same comparison for two prostate cancer studies: Welsh and Dhanasekaran (Figure 2.6). Inter-study predictions based on the raw data had the worst performance and all the normalization methods provided some improvement. Overall, SN_std+rGN_std has the best prediction accuracy rate and it achieves a 100% accuracy rate when the Dhanasekaran study is used as training data to predict the Welsh study, shown by line 4 in Figure 2.6B.

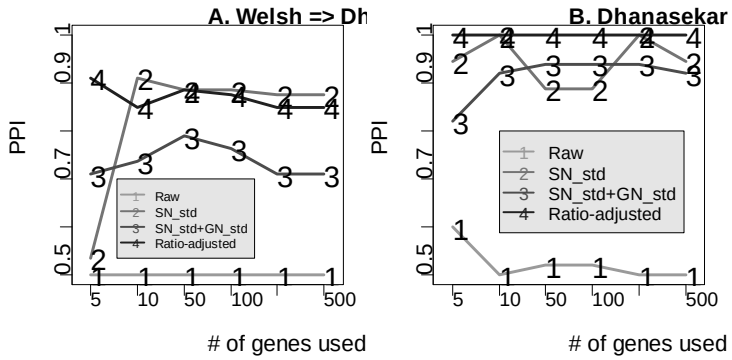

Figure 2.6 Inter-study prediction of prostate cancer data by PAM

Following the calibration scheme described in Figure 2.3, we randomly selected a few samples from the test set to as use a calibration to perform SN_std+rGN_std for inter-study prediction of the three lung cancer studies. Similarly, we only report the results by PAM here, while the results by LDA and KNN are presented in Appendix A. The numbers of the samples in the calibration are 1:1 (1 normal and 1 adenocarcinoma), 2:2, and 3:3 shown by line 1, 2, and 3 in Figure 2.7 respectively. The prediction results were based on an average of ten random draws of calibration samples. In general, larger numbers of calibration samples provide better estimate of normalization factors. The improvement from 1:1 to 2:2 was significant while not much improvement was gained from 2:2 to 3:3. Most of the three pair-wise cross predictions have PPIs above 70%, and line 2 and 3 have PPIs above 90% most of the time. Line 1 with only 2 samples has the worst results when compared to lines 2 and 3. The result provides a practical guideline that a calibration set of three normal samples and three tumor samples should be used when designing a prospective clinical trial in an independent medical center or using a different array platform.

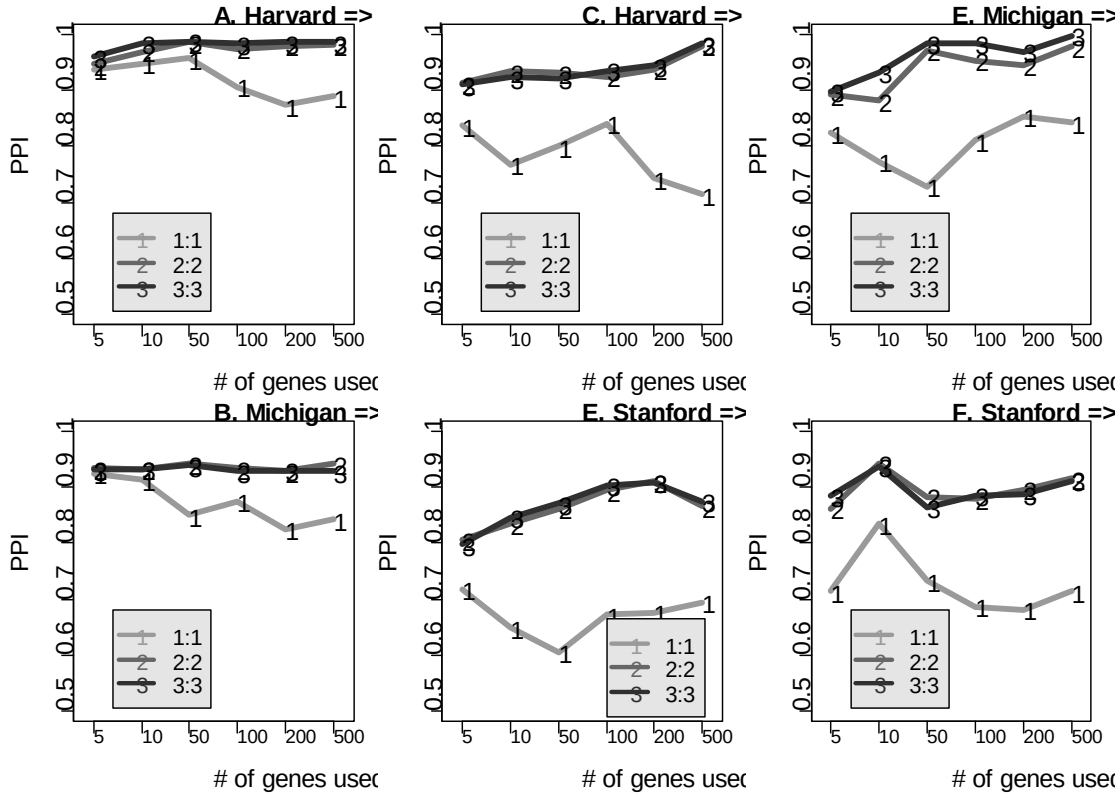

Figure 2.7 Inter-study predictions by SN_std+rGN_std with different number of calibration samples

2.4.3 Inter-study prediction regarding survival risk

Figure 2.8 shows the within-study prediction results (with leave-one-out cross-validation) based on a processed data set without inter-study gene-matching. We omitted the prediction based on the raw data because it performed poorly. The with-study predictions after SN_std are displayed on the left and the with-study predictions after SN_std+GN_std are displayed on the right side of Figure 2.8. In general, prediction results from SN-std+GN_std performed better than SN_std only in that the log-rank test p-values were more significant and the numbers of predicted high-risk and low-risk samples were more balanced, which is expected since the median threshold

from the training data was used. For example, the log-rank test p-value of the Michigan study improved from 0.38 to 0 in Figure 2.8C and 2.8D. The predicted risk group sizes were sixty-two and twenty-four after applying SN_std and became more balanced (forty-four and forty-two) after SN_std+GN-std.

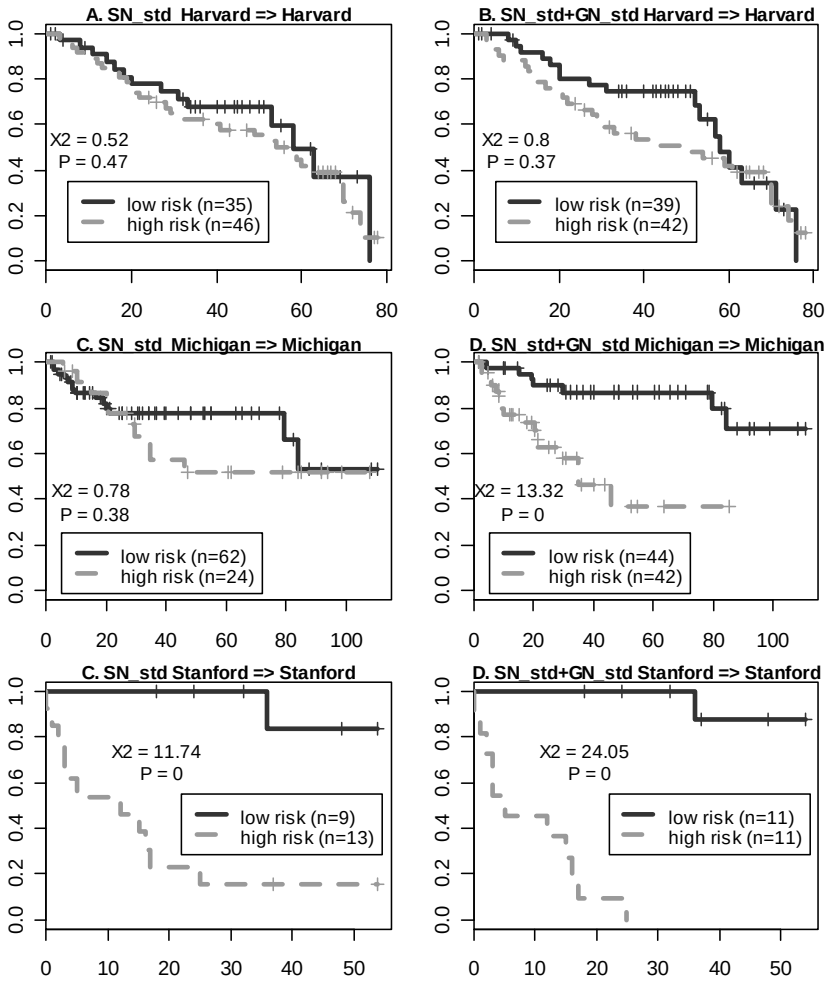

Figure 2.8 Within-study survival prediction

Figure 2.8A: Within-study prediction in the Harvard study after SN_std ,based on 50 genes. Figure 2.8B: Within-study prediction in the Harvard study after SN_std+GN_std, based on 50 genes.

Figure 2.9 shows pair-wise inter-study prediction for all three lung cancer studies comparing SN_std and SN_std+Gn_std. As Figure 2.8 shows as well, inter-study predictions after SN_std are displayed on the left and inter-study predictions after SN_std+GN_std are displayed on the

right. The only improvement happened in Michigan versus Stanford. Similar to the results in Figure 2.8, SN_std produced confusing results (Figures 2.9I and 2.9K), where samples predicted as low-risk had worse survival rates than the high-risk group. Results from SN_std+GN_std (Figures 2.9J and 2.9L), corrected the error and provided more reasonable predictions. Although there was no obvious improvement from SN_std to SN_std+GN_std in Harvard versus Michigan and Harvard versus Stanford, we still noticed some differences. First, the numbers of samples predicted in the low and high risk groups are more even for SN_std+GN_std than SN_std. For example, the numbers are seventy-two and nine (Figure 2.9C) and they became forty-one and forty (Figure 2.9D). Also, in Figure 2.9G, the prediction model built on the Stanford study could not distinguish the between high risk and low risk in the Harvard samples, and all of the eight-one Harvard samples were predicted to be in the low risk group after SN_std; in Figure 2.9H, after applying additional GN_std, twenty-one samples were predicted as low risk and sixty samples were predicted as high risk.

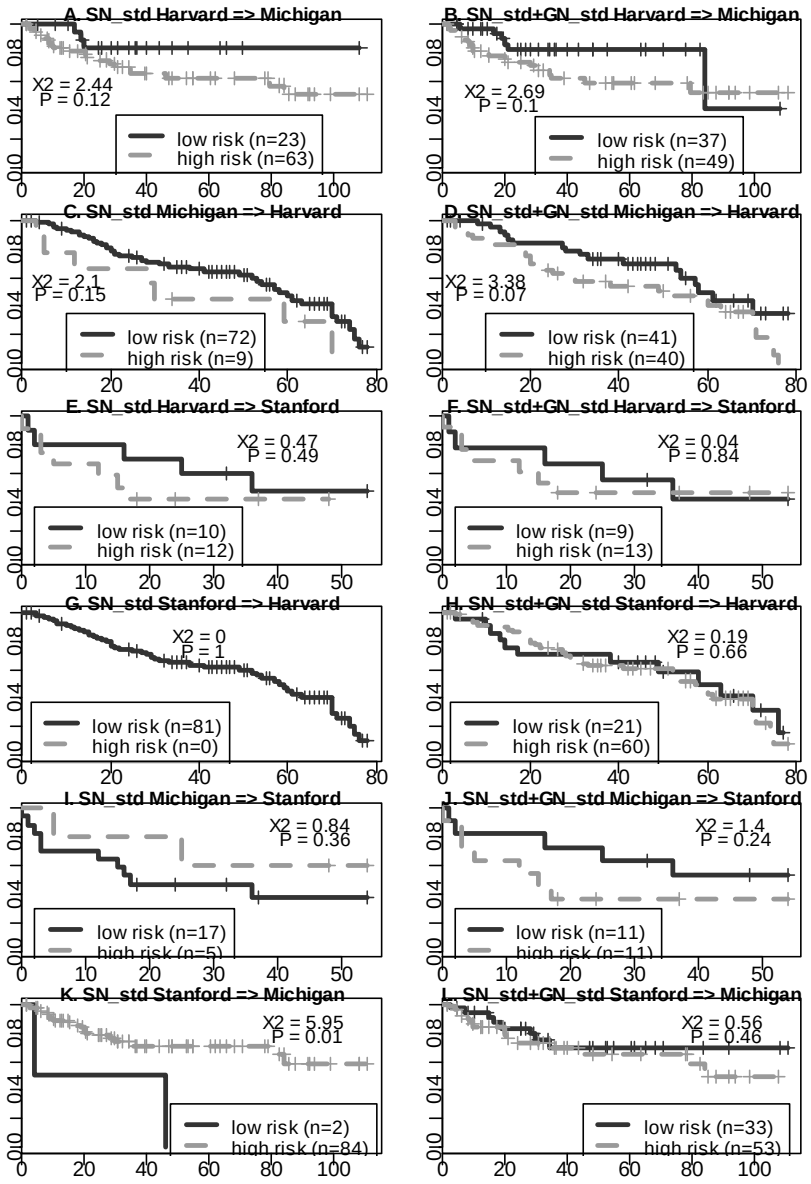

Figure 2.9 Inter-study survival prediction

Figure 2.9A: Prediction of the Michigan study after SN_std, using the Harvard model.
Figure 2.9B: Prediction of the Michigan study after SN_std+GN_std, using the Harvard model.
Figure 2.9C: Prediction of the Harvard study after SN_std, using the Michigan model.
Figure 2.9D: Prediction of the Harvard study after SN_std+GN_std, using the Michigan model.

2.5 DISCUSSION

2.5.1 Gene-wise Normalization (GN)

Sample-wise normalization (SN) is a routine step in supervised machine learning for microarray data and it works well in many within-study cases, thus GN_std is seldom discussed in the literature. However, GN could be a critical factor when dramatic differences exist between different platforms and none of the SN methods is expected to bring the expression values to the same level for all genes. As we have shown in this study, implementation of GN is fast and a significant improvement is often obtained compared to SN only. We suggest that GN should be applied in the analyses to directly carry prediction models from one study to another. However, GN should be applied cautiously because it is sensitive to the ratio of disease groups in the training and test set.

2.5.2 Applicability of calibration

Calibration of experimental instruments is a common and necessary practice when installed in a new lab or operated by a new technician. It is especially necessary when preparing for a large survey or screen test. Our proposed SN_std+rGN_std method requires knowledge of sample labels in the test set for proper gene-wise normalization. As a result, a calibration scheme was developed for application to a real prospective clinical trial. In our analysis of the three lung cancer data sets, a small calibration set of three normal and three diseased samples was sufficient. Although experimental quality and genetic variation may be different in other diseases, it provides a rough guideline for real life applications.

2.5.3 Application in survival analysis by SuperPC

To further investigate the cause of improvement in GN, we conducted a Principle Component Analysis (PCA) on the three individual lung cancer data sets based on the top 50 genes, comparing normalization by SN_std and SN_std+GN_std. It is clearly shown in Figure 2.10 that the PCA is heavily dominated by the first component after applying SN_std for all three studies. By additional GN_std, proportions of information after the second principal component are increased and can be better utilized to construct the prediction model. This explains why additional GN improves the prediction performance.

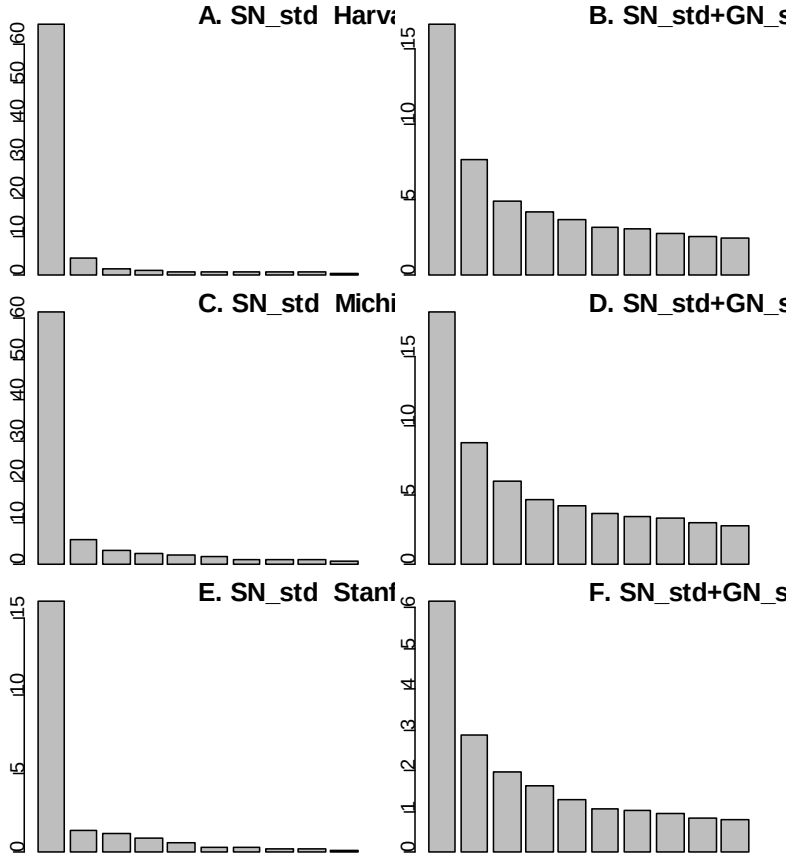

Figure 2.10 PCA based on unmatched data

In the SN_std+GN_std normalization for the survival analysis of these three lung cancer data sets, we did not adjust for the ratio because the prediction of survival risk, in some sense, is

not a supervised analysis, unlike the prediction of disease classification. Furthermore, because all of the samples in the training and test set are tumor samples, the difference between the low-risk and high-risk groups is much smaller than the one between the normal samples and the tumor samples. The sample ratio issue is not as critical here. In general, predicting survival is more difficult than predicting disease classification, and we only observed obvious improvement of SN_std+GN_std over SN_std in the inter-study prediction between Michigan and Stanford. However, we did observe the importance and necessity of performing GN_std based on the within-study leave-one-out cross validation in Figure 2.8.

2.5.4 Clinical implication

As we mentioned in the introduction, studies performing direct inter-study prediction of microarrays are rare, especially when the training and the test studies are from different microarray platforms due to all kinds of technical difficulties. Suppose we need to do cross-prediction of a prospective microarray study from a certain platform. It could not be successfully accomplished if no prior study from the same platform exists, which happens often because of the variety of microarray platforms. Even for a platform like Affymetrix, the probe designs have been changed gradually and extensively. In reality, in order to predict a prospective study, we usually need to have a training study from the same platform as the test study, which is not always available. By applying our proposed SN+GN normalization method, together with an appropriate data preprocessing strategy and gene matching, we can cross-predict the prospective study using training set in any platform most of the time. On the other hand, for a certain disease, if there is a microarray study of a very good data quality, such as good completeness, a large enough sample size, and a high self-prediction accuracy, we can trust this study and use it to

predict possibly any study. Also, with more and more microarray studies available, there is an urgent need to combine the information from all the studies. Our proposed normalization method makes that possible.

2.5.5 Limitations

This study performed inter-study analyses for the same disease and showed that with appropriate normalization method, high prediction accuracy could be obtained. However, if the analyses are conducted between different organs or different diseases, such as the liver and prostate cancer data, it is highly possible that some genes could never be brought to the same level between training and test data sets due to the intrinsic differences even with SN+GN. Also, when few similar significant genes exist across the studies, cross prediction could be very difficult and SN+GN cannot solve it. As we see from the survival prediction between Harvard and Michigan, and Harvard and Stanford, the additional GN after SN is not better, although neither is it worse. Last, when the calibration data is not available or not reliable, the rGN cannot be accomplished.

2.6 CONCLUSIONS

In this study, we investigated the normalization issue for enhancing inter-study disease prediction, a critical issue for microarray translational research. Instead of developing more sophisticated sample-wise normalization methods (SN), we observed that gene-wise discrepancies of expression levels across studies are often significant in impeding successful inter-study prediction and that gene-wise normalization (GN) is necessary. We further found that

differential sample size ratios of diseased and normal groups greatly deteriorate the gene-wise standardization procedure. An analytical method with equal mixture assumption was proposed for ratio-adjusted gene-wise normalization (rGN). Finally, since the sample labels are needed to perform rGN, we developed a practical calibration scheme for the design of a prospective clinical trial.

3.0 COMMON PREDICTIVE BIOMARKERS AND CROSS-PREDICTABILITY IN THE EXPRESSION PROFILES OF MULTIPLE CANCER TYPES

Manuscript submitted.

George C. Tseng[#,1,2], Chunrong Cheng[#,1], Yan Ping Yu[3],

Joel Nelson[3], George Michalopoulos[3] and Jian-Hua Luo[*,3]

[1]Department of Biostatistics, University of Pittsburgh

[2]Department of Human Genetics, University of Pittsburgh

[3]Department of Pathology, University of Pittsburgh

[#] the first two authors should be regarded as joint first authors.

Email: ctseng@pitt.edu, Tel: 412-624-5318

Email: luoj@msx.upmc.edu, Tel: 412-648-8791

3.1 ABSTRACT

Microarray technology has been widely applied to the analysis of many malignancies, however, integrative analyses across multiple studies are rarely investigated. In this study we performed a meta-analysis on the expression profiles of four published studies analyzing organ donor, benign tissues adjacent to tumor and tumor tissues from liver, prostate, lung and bladder samples. We identified 99 distinct predictive biomarkers in the comparison of all three tissues in liver and prostate and 44 in the comparison of normal versus tumor in liver, prostate and lung. The bladder samples appeared to have a different list of predictive biomarkers from the other three cancer types. The identified predictive biomarkers achieved high accuracy similar to using whole genome in the within-cancer-type prediction. They also performed superior than the one using whole genome in inter-cancer-type prediction. To test the validity of the predictive biomarkers, 23 independent prostate cancer samples were evaluated and 96% accuracy was achieved in inter-study prediction from the original prostate, liver and lung cancer data sets respectively. The result suggests that the compact lists of predictive biomarkers are important in cancer development and represent the common signatures of malignancies of multiple cancer types. Pathway analysis revealed important tumorogenesis functional categories.

3.2 INTRODUCTION

Human malignancies can occur in almost all organs, with the exception of several accessory sex organs. Most human malignancies, regardless of the origins the of tissues, contain two major characteristics: uncontrolled growth and the ability to metastasize. Abnormalities of the same

signaling pathways can be found in multiple types of human cancers; a tumor may contain multiple abnormalities in signaling. Overlapping these abnormalities among multiple types of tumors may shed light on some key alterations in carcinogenesis.

Prostate cancer is second only to skin cancer as the most commonly diagnosed malignancy in American men: at current rates of diagnosis, one man in six will be diagnosed with the disease during his lifetime (Jemal, *et al*. 2005). Even though nutritional and environmental etiology has been implicated for prostate cancer development, such link has yet to be firmly established in general population. Some studies suggested that up to 80% of men older than 80 were found to contain pathologically recognizable prostate cancer, while men below 40 rarely developed the disease. This argues against any singular specific etiology responsible for prostate cancer besides aging. Histologically, prostate cancer cells closely interact with their neighbor stromal cells to form some distinctive architectural patterns that make up the basis of Gleason's grading (Gleason 1966). The clinical courses of most prostate cancers are long, and some are life-threatening. Hepatocellular carcinoma, on the other hand, is quite the opposite. It is not age related, and is tightly linked to cancer etiologies such as alcohol, the hepatitis B or C virus and certain toxins. Hepatocellular carcinoma is distinctive in its well confined nodular architecture. The clinical course of most of the hepatocellular carcinomas are short and the fatality is high. Most of the lung cancers, with the exception of small cell carcinoma, are also associated with distinctive etiologies, such as smoking or chronic exposure to certain type of carcinogens. The urothelial carcinoma of the urinary bladder, however, is primarily idiopathic or virus related. Since these four types of cancers are so far apart in etiology, morphology, and clinical courses, any common ground between these tumors could be interpreted as the likely common pathway of carcinogenesis.

Microarray technology has been widely applied to the analysis of many malignancies, including the four cancer types mentioned in the literature above. However, using meta-analysis to integrate multiple studies has rarely been investigated. Segel *et al* (Segal, *et al*. 2004) proposed a systematic approach to incorporate 1,975 arrays in 22 tumor types and constructed a large gene module map. The resulting module map was, however, too complex to follow up and the modules were based on 2,849 known biologically meaningful gene sets instead of learning new sets of predictive biomarkers. The gene matching of heterogeneous array types also potentially deteriorate the analysis accuracy. In this report, we performed a meta-analysis on 455 arrays collected from four microarray studies in Affymetrix U95Av2 platform: 94 samples of liver tissue (Luo, *et al*. 2006) (43 liver cancer, 30 hepatic tissues adjacent to liver cancer, 21 normal liver from organ donors), 148 samples of prostate tissues (Yu, *et al*. 2004) (66 prostate cancer, 59 prostate tissues adjacent to prostate cancer and 23 organ donors), 151 samples of lung tissues (Bhattacharjee, *et al*. 2001) (134 tumors and 17 normal lung tissues) and 62 urinary bladder tissues (Stransky, *et al*. 2006) (5 normal and 57 tumors). The use of common array platform has avoided the problem of incorrect gene matching and gene annotation, a common cause to deteriorate the performance of meta-analysis in microarray (Kuo, *et al*. 2006). We performed an analysis on two batches of samples. In batch I, all three tissue types in the liver and prostate samples were analyzed using the analysis of variance (ANOVA) model. In batch II, the normal and tumor tissues in all four cancer types were included and a t-test was used to identify predictive biomarkers (see Table 3.1 for data description). The identified biomarkers were found to have high predictability in both within-cancer-type (cross-validation within a single cancer type) and inter-cancer-type (i.e. a prediction model trained in one cancer type and used to predict another cancer type) prediction via leave-one-out cross validation. Further pathway enrichment

analysis identified statistically significant function categories of the biomarkers. Validation of the 47 batch II predictive biomarkers on 23 prostate tissues yielded 96% accuracy in inter-study prediction from the original prostate, liver and lung cancer data sets respectively, showing the robustness of the predictive biomarkers and their implications to common carcinogenesis of multiple cancer types.

Table 3.1 Overview of data sets used in batch I and batch II analyses.

	Batch I Analysis			
	Organ Donor (N)	Adjacent to Tumor (A)	Tumor (T)	Total
Liver	21	30	43	94
Prostate	23	59	66	148
	Batch II Analysis			
	Organ Donor (N)		Tumor (T)	Total
Liver	21		43	64
Prostate	23		66	89
Lung	17		134	151
Bladder	5		57	62

3.3 MATERIALS AND METHODS

3.3.1 Data and preprocessing

We collected four published microarray data sets (Bhattacharjee, *et al*. 2001; Luo, *et al*. 2006; Stransky, *et al*. 2006; Yu, *et al*. 2004) to perform meta-analysis on prostate, liver, lung and bladder samples. A total of 455 U95Av2 arrays were analyzed (94 liver, 148 prostate, 151 lung and 62 bladder tissues) with each covering 12,625 genes and EST sequences. The common array platform eliminated technical difficulties including gene matching and inter-platform discrepancies. In liver and prostate data sets, three types of samples were collected: organ donor

(N), normal tissues adjacent to tumor (A) and tumor tissues (T). In lung and bladder tissues, only organ donor and tumor tissues were available. We analyzed the data through two batches of analyses. In the first batch, both liver and prostate data sets with all three tissues were included. The expression patterns across the three types of samples were the major targets for investigation. In the second batch, data of all four organ types were included and only normal and tumor samples were compared. For details see Table 3.1.

The raw data (CEL files) were preprocessed in each cancer type separately using dChip software for array quality assessment, normalization, expression intensity extraction and log-transformation (base 2). Genes of low information content in each data set were filtered respectively and the union gene set of the four data sets was retrieved for further analysis. Specifically, in each data set, the top 50% genes with the largest average intensities were first selected. Among them the top 50% genes with the largest standard deviations were further identified, resulting in 25% genes (3,156 genes) selected in each data set. The union list of these most informative 25% genes in four data sets was used for subsequent downstream analysis (a total of 5,917 genes).

3.3.2 Biomarker selection by ANOVA and t-test

In batch I analysis, ANOVA model was fitted for the organ donor (N), adjacent to tumor (A) and tumor (T) samples with a β parameter for field effect and a γ parameter for tumor effect. Stepwise algorithm was used to select the best regression model. The ANOVA model is described in the following:

$$Y_{in} = \alpha_i + \beta_i \cdot F_{in} + \gamma_i \cdot T_{in} + \varepsilon_{in}$$

where i =1,… 5917 for all the genes, n =1,… 94 for liver samples and n=1,… 148 for prostate samples. The field effect binary covariate F_{in} = 1 for A or T group; F_{in} = 0 for N group. The tumor effect covariate T_{in} = 1 for T group; T_{in} = 0 for N or T group. Field effect is defined as the expression difference between normal tissues (N) compared to tissues adjacent to tumor (A) and tumor tissue (T). Tumor effect is defined as a further difference between A and T. Genes satisfying the following criteria were selected: (a) statistical significance: adjusted q-value for the final stepwise-selected ANOVA model after Benjamini-Hochberg correction is less than 0.05 (i.e. to control false discovery rate smaller than 0.05); (b) biological significance: field effect or tumor effect is larger than 0.4 (correspond to ~32% fold change). The field effect and tumor effect parameter β and γ both have three possibilities-- positive, negative and no change --, resulting in eight patterns as described in Figure 3.1A. Figure 3.1B and 3.1C shows the number of genes selected in liver and prostate samples respectively and their distribution in the eight pattern categories. The intersection of selected ANOVA genes in liver and prostate with concordant pattern categories were used to construct prediction model for within-cancer-type (Liv→Liv and Pro→Pro) and inter-cancer-type (Liv→Pro and Pro→Liv) analysis. To summarize a list of gene markers in batch I for further analysis, genes selected in more than 70% of the times in leave-one-out cross validation (see section below for more detail) in the above procedure were identified as the “batch I predictive biomarkers” (batchI-PBs).

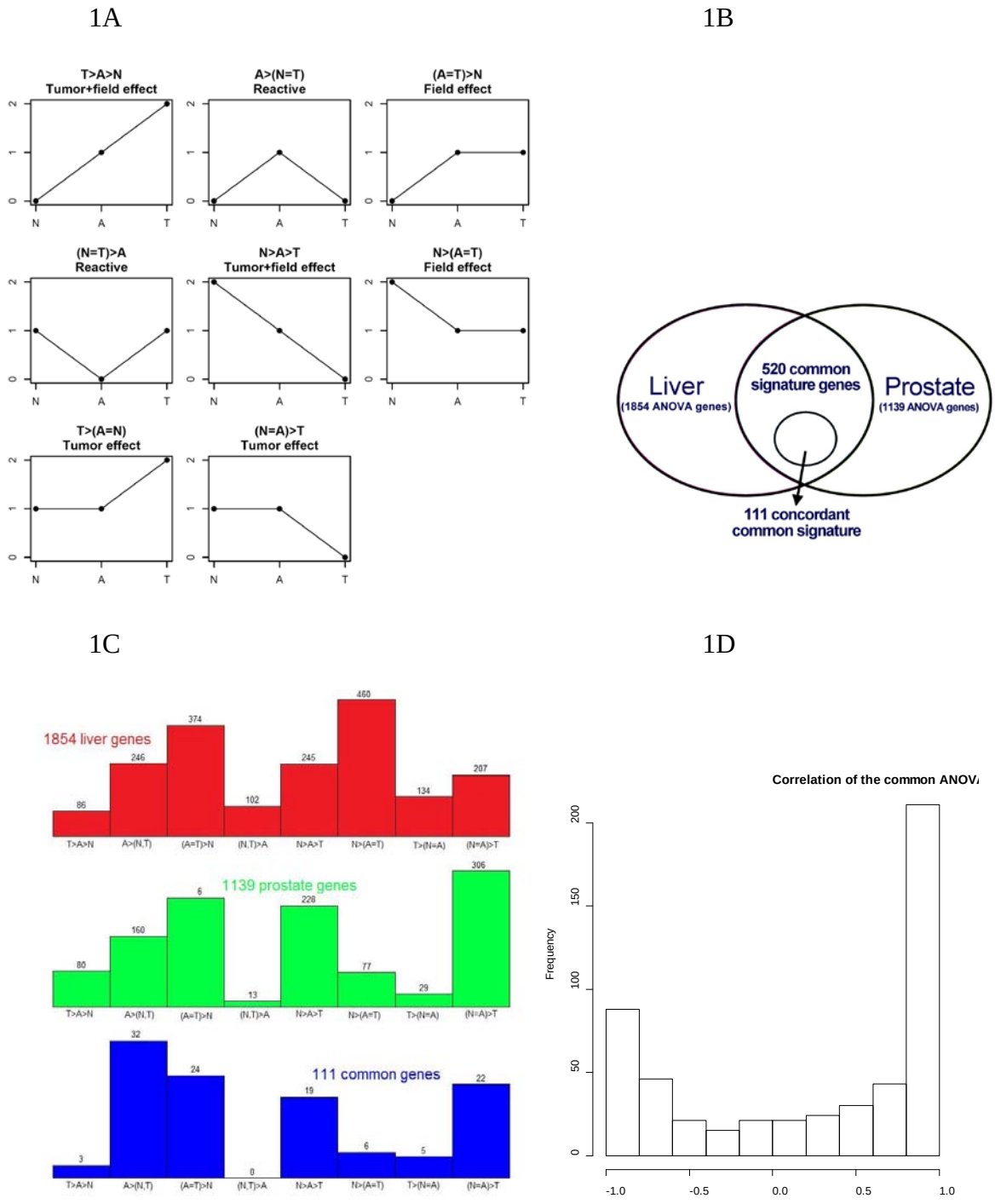

Figure 3.1 ANOVA model for batch I analysis

(A) Eight categories of ANOVA patterns used to select predictive biomarkers. N denotes normal, A tissue adjacent to cancer, and T tumor sample. (B) Venn diagram representation of the number of ANOVA genes found to be significantly altered in liver and prostate tissues when comparing N, A and T groups. (C) Bar graph of genes that were altered in liver (1854), prostate (1139) or both tissue samples with same pattern (111). (D) Histogram of correlations of N-A-T patterns across prostate and liver of the 520 common ANOVA genes.

In the batch II analysis, similar gene selection procedure was performed. Instead of ANOVA, simple t-test was performed to distinguish normal and tumor. Given the comparison of a pair of cancer types (e.g. liver vs lung), genes satisfying the two criteria used in batch I were first selected and the intersection of the gene lists obtained from the two compared cancer types were identified. Among them, genes with concordant differential expression direction (up- or down-regulation) were used to construct prediction model for within-cancer-type (Liv→Liv and Lun→Lun) and inter-cancer-type (Liv→Lun and Lun→Liv) analysis. Leave-one-out cross validation was similarly performed. For each pair of cancer type comparison, gene lists of more than 70% appearance in the leave-one-out cross validation signatures were identified and were denoted as "liv-pro-PBs" (i.e. predictive biomarkers in liver-prostate comparison), "liv-lun-PBs" etc. The intersection genes of "liv-pro-PBs", "liv-lun-PBs" and "pro-lun-PBs" are denoted as "batchII-PBs" (See Figure 3.5; bladder cancer data appear to be very different from liver, prostate and lung as will be describe later).

3.3.3 Gene-specific scaling in inter-cancer-type classification

Figure 3.2 demonstrates expression patterns of one selected gene for each of the eight pattern categories (the category (N=T)>A had no gene and is omitted). We observed that gene-specific scaling was needed for many of the biomarkers so the prediction information could be carried across organs. For example in "APBA2BP", the expression of group A is consistently greater than N and group T is further greater than A in both liver and prostate samples. However, the levels of expression intensities in liver and prostate are in different scale even though all the liver and prostate samples are preprocessed and properly normalized across data sets. This phenomenon may be due to differential sample preparation, tissue physiology and/or

hybridization conditions in different studies. As a result, we conducted gene-specific scaling in all inter-cancer-type classification. Conceptually the scaling parameters are estimated so that the gene vectors in each study are standardized to mean 0 and standard deviation 1. However, since each study has a different ratio of normal versus tumor samples, we performed a bootstrap sampling before scaling so that the gene vectors were standardized under a synthetic condition that groups (N, A and T) are of equal sample size in each study (see Appendix for more details).

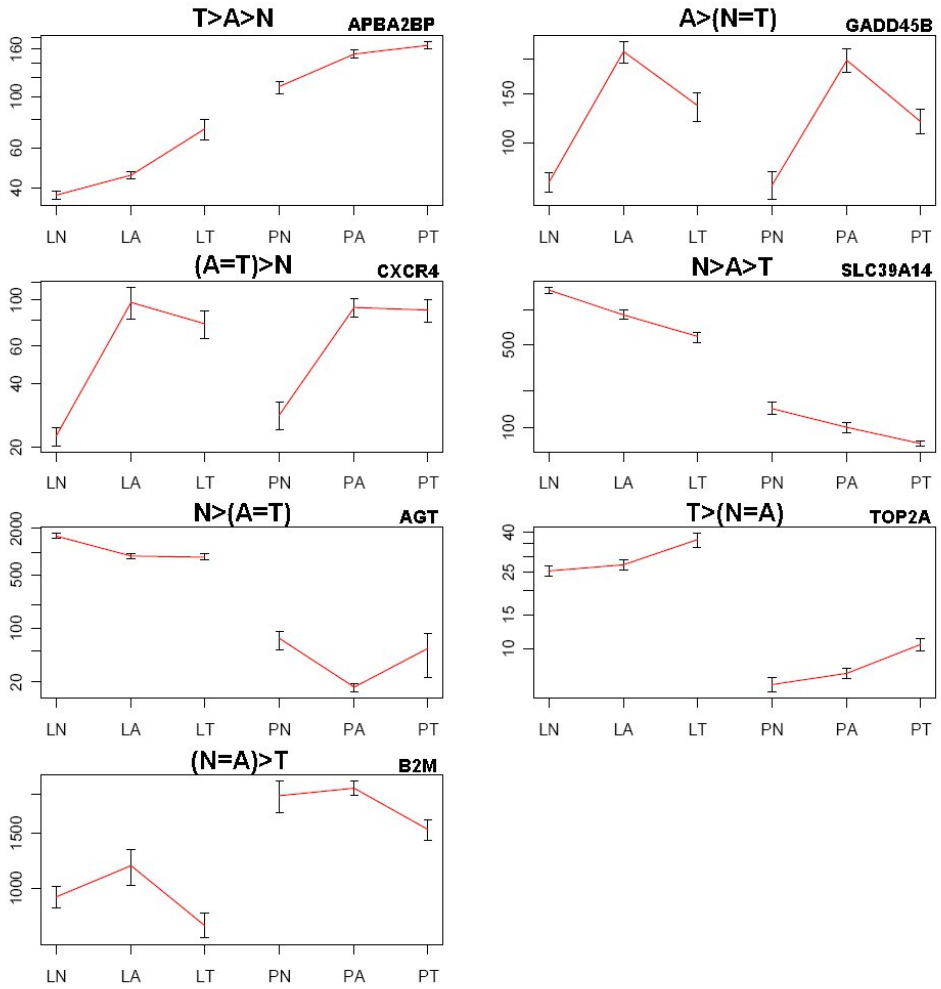

Figure 3.2 Expression patterns of selected representative genes in liver and prostate samples.

Selected genes of seven pattern categories from the 111 common concordant ANOVA genes in liver and prostate samples. Global sample normalization has been performed across prostate and liver data sets. It is clearly seen that although all these biomarkers demonstrate concordant patterns across prostate and liver, many of them (APBA2BP, SLC39A14, AGT, TOP2A and B2M) are at different expression level and direct application of a prediction model developed in one data set will likely perform poor in the other data set.

3.3.4 Classification method and leave-one-out cross validation

PAM (Prediction Analysis of Microarrays) was used to construct the prediction models in this paper (Tibshirani, *et al*. 2002). The method has been found effective in many microarray prediction analyses and has the merit that gene selection is embedded in the method. When "all genes" are used, the predictive genes are automatically chosen from the total of 5,917 genes to construct the prediction model. When "common signatures" are used, the common biomarkers are selected according to the description in the section "Biomarker selection by ANOVA and t-test" and no gene selection is further performed in PAM. Results of both gene selection procedures are reported and compared. To avoid over-fitting in the evaluation of cross-predictability of the predictive biomarkers, we conducted rigorous leave-one-out cross validation (see the prediction scheme in Figure 3.3A and 3.3B), i.e. the left out sample does not participate in the selection of marker genes.

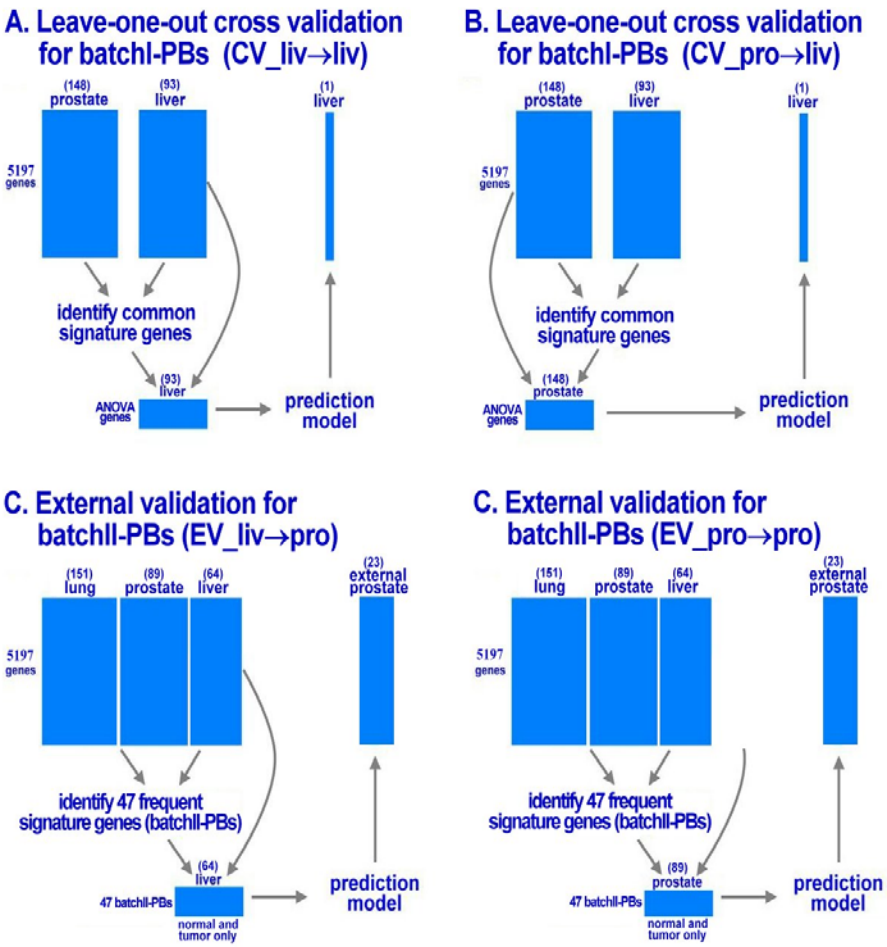

Figure 3.3 Schemes of leave-one-cross-validation or external validation for batchI-PBs and batchII-PBs.

Upper: scheme for leave-one-out cross validation to evaluate the procedure of selecting batchI-PBs and batchII-PBs. The test sample is first left aside. The remaining samples are used for selecting predictive biomarkers and construct the prediction model to be used to evaluate the set-aside test sample. This scheme is used to evaluate procedures of selecting both batchI-PBs and batchII-PBs to generate Table 3.2 and Table 3.3. (A) an example to evaluate liv→liv in Table 3.2 (B) an example to evaluate pro→liv in Table 3.2. Lower: scheme for external validation of batchII-PBs by 23 independent prostate cancer samples. (C) external evaluation of the prediction model based on liver data and batchII-PBs (EV_liv→pro). (D) external evaluation of the prediction model based on the old prostate data and batchII-PBs (EV_pro→pro).

3.3.5 Confusion matrix and prediction index

In the literature, the overall accuracies from different methods are usually reported to compare performance. It is, however, often a misleading index in practice. Supplementary Table B.1

demonstrates an example. Among 42 tumor patients, one false negative was made and among six normal patients, five false positives were made. The overall accuracy is pretty high (87.5 %) but it is a result of predicting almost all tissues as tumor with high sensitivity (97.6%) but extremely low specificity (16.7%). To solve this problem, reporting both sensitivity and specificity or plotting the receiver operating characteristic (ROC) curve is commonly used. In this paper, we report the confusion matrixes that convey the entire prediction results in the appendix. A 2×2 table is used to summarize the number of patients in true and predicted status of normal or tumor groups. The two off-diagonal numbers represent the false positives and false negatives in the prediction and their sum represent to total errors made (see Supplement Table B.1). We then further summarize the prediction results by a prediction performance index (PPI) that is defined as the average of sensitivity and specificity, to be used throughout this paper for performance evaluation.

3.3.6 Pathway analysis

For each gene list of predictive biomarkers, the gene ontology (GO) database was used for pathway enrichment analysis. For each GO term, a Fisher's exact test was performed to determine the enrichment of the gene list and a p-value was generated (Draghici, *et al*. 2003). We performed this analysis in batchI-PBs, batchII-PBs and all pairwise comparison predictive biomarkers in batch II (liv-pro-PBs, liv-lun-pro-PBs etc). The p-value results were summarized in a heatmap (Figure 3.4).

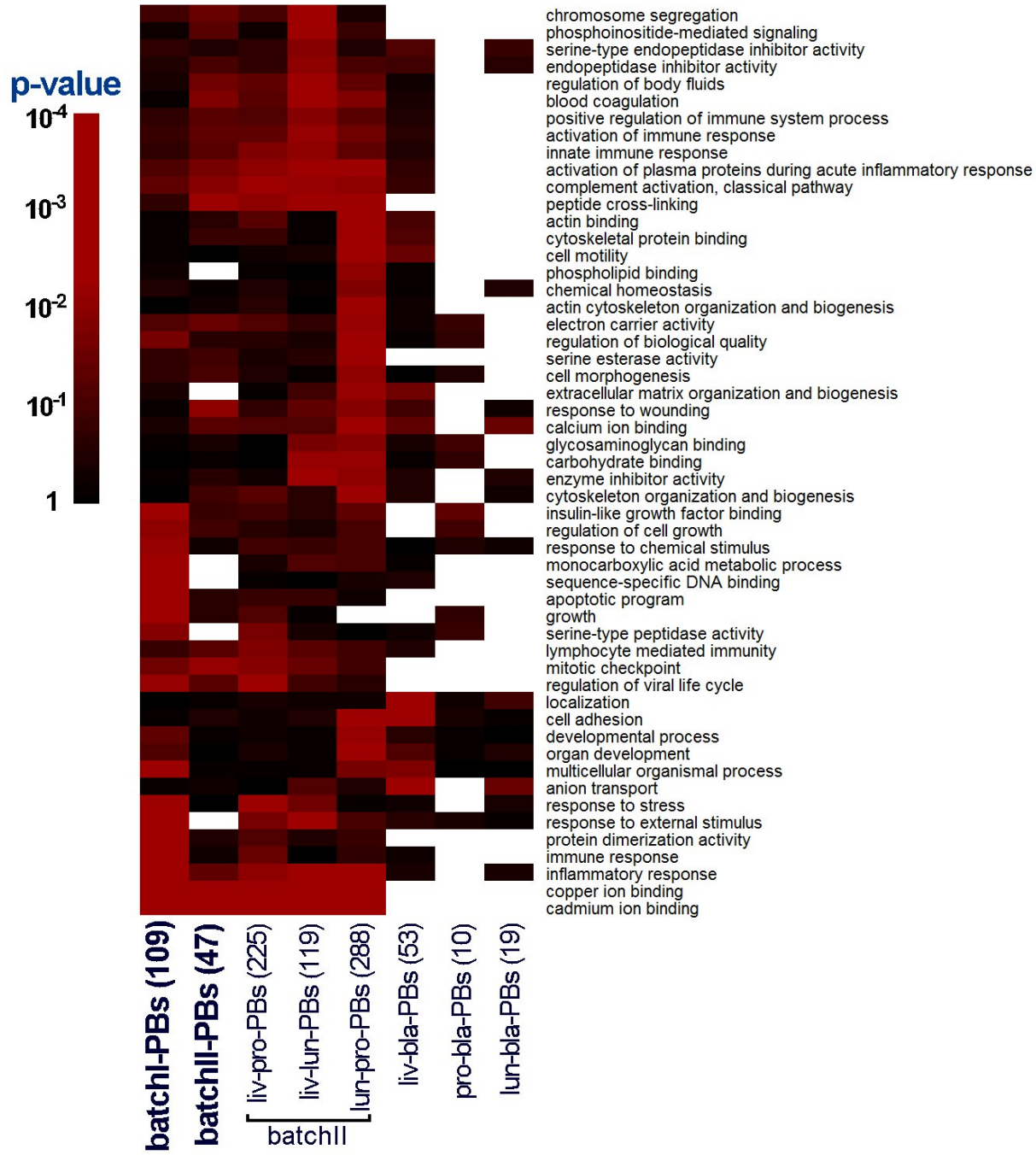

Figure 3.4 Pathway analysis heatmap.

The enriched Gene Ontology terms are demonstrated on rows and lists of predictive biomarkers are shown on columns. The significance (p-values) is represented by gradient red color.

3.3.7 External evaluation of batchII-PBs by independent prostate data

A data set of 23 prostate cancer samples independently performed by LaTulippe *et al*. (LaTulippe, *et al*. 2002) was used for external validation of the batchII-PBs. A toltal of 47 batchII-PBs were identified from the normal and tumor samples in liver, prostate and lung data sets. To evaluate the robustness and inter-cancer-type cross-predictability, a prediction model based on the 47 batchII-PBs in the normal and tumor samples of liver data set was constructed and was used to evaluate the 23 external prostate cancer samples (see "EV_liv→pro in Figure 3.3C). The evaluation of prediction model generated by the old prostate data is denoted by "EV_pro→pro" in Figure 3.3D. Similarly we also perform "EV_lun→pro" evaluation. The data preprocessing of the 23 new samples was conducted similarly to the four analyzed data sets and simple constant normalization was adopted against the original prostate data set. Additional gene-wise normalization against the original prostate is also applied so the liver and lung data sets can be used to predict the 23 new prostate samples.

3.4 RESULT

To identify common signature genes among four types of malignancies, we started with the prostate and liver data sets in batch I analysis because of more balanced numbers of tumor and normal samples and availability of benign tissues adjacent to tumor. In this analysis, 1,854 genes from liver data set and 1,139 genes from the prostate data set were found to fit the ANOVA model and meet the gene selection criteria. Among these genes, 520 genes were common in both organs (Venn diagram in Figure 3.1B). The histogram of correlations of N vs A vs T patterns

(average intensities of each group) across two organs in each gene is shown in Figure 3.1D. Majority of the genes were highly correlated across prostate and liver but surprisingly 113 genes presented strong negative correlation (<-0.7), which may reflect the differences in tissue types. The 520 selected genes were categorized into eight patterns as demonstrated in Figure 3.1A. These patterns represent either tumor specific alteration, field effect, or reactive changes. Among these 520 genes, 111 genes were in the same pattern categories in liver and prostate (Figure 3.1C) based on our definition in Figure 3.1A. Further analysis of expression of the 111 genes in both organs indicated that even though the expression patterns for these genes across N, A and T were identical in both organs, the levels of expressions may vary greatly (for example, APBA2BP and SLC39A14 in Figure 3.2). This suggests that direct application of classification model constructed in one cancer type may not predict the histology of tissues in the other cancer type. To resolve this problem, an adequate gene-specific scaling across organs was carried out for the inter-cancer-type prediction. The gene-specific scaling procedure described in the Method section and Appendix is applied for all analyses hereafter.

We performed leave-one-out cross validation throughout the prediction analyses. There are 242 samples in liver and prostate data sets. Among the 242 leave-one-out cross validation analysis, a total of 109 common biomarkers were identified in more than 70% leave-one-out cross validation and all of them belong to the 111 gene list using all liver and prostate samples described above. These 109 frequently identified biomarkers are named “batchI-PBs”. 99 (out of 109) were identified as distinct predictive biomarkers (Supplement Table B.4). Subsequently we assessed the cross-predictability of the identified biomarkers. When using all genes, we observed high PPI between normal and tumor comparison (N vs T) with 96.5% in liver dataset and 93.9% prostate dataset while lower accuracy was observed between adjacent and tumor (79.9% in liver

and 71.4% in prostate) (Table 3.2). When only common signature biomarkers were used, the prediction accuracy remained comparable to using all genes (N vs T: 96.5% in liver and 98.8% in prostate; A vs T: 75.6% in liver and 66.7% in prostate). The result suggests that the common signature biomarkers carry as good predictive information as the entire 5,917 genes. We then further conducted inter-cancer-type classification analysis. We used either "all genes" (the entire 5,917 genes) or the common signatures to construct a prediction model in one cancer type and predict in another cancer type. The prediction evaluation was performed in a manner of leave-one-out cross validation. We denoted "prostate→liver" as constructing prediction models using prostate samples and predicting liver samples. We found that prediction with "all genes" did not perform well with only 47.4% in liver→prostate and 66.3% in prostate→liver among N vs T comparison and 55.7% in liver→prostate and 51.9% in prostate→liver among A vs T comparison. On the other hand, the model using common signature genes achieved much superior performance, nearly as good as the within-cancer-type classification (96.3% in liver→prostate and 93% in prostate→liver among N vs T comparison and 65.1% in liver→prostate and 74.7% in prostate→liver among A vs T comparison). The results clearly demonstrate the cross-predictability of the common signatures.

Table 3.2 Prediction performance indexes (PPI) in batch I analysis

Liver vs Prostate (Normal vs Tumor)				
	liv→liv	pro→liv	pro→pro	liv→pro
All genes	96.5%	66.3%	93.9%	47.4%
Common signature	96.5%	93.0%	98.8%	96.3%
Liver vs Prostate (Normal vs Adjacent)				
	liv→liv	pro→liv	pro→pro	liv→pro
All genes	92.6%	77.9%	96.6%	54.6%
Common signature	98.2%	96.0%	98.3%	96.6%
Liver vs Prostate (Adjacent vs Tumor)				
	liv→liv	pro→liv	pro→pro	liv→pro
All genes	79.9%	51.9%	71.4%	55.7%
Common signature	75.6%	74.7%	66.7%	65.1%

Pairwise two-group comparisons (N vs T, N vs A and A vs T) are performed.

Table 3.3 Prediction performance indexes (PPI) in batch II analysis.

Liver vs Prostate				
	liv→liv	pro→liv	pro→pro	liv→pro
All genes	96.51%	66.28%	93.94%	47.36%
Common signature	97.67%	97.67%	95.55%	94.14%
Liver vs Lung				
	liv→liv	lun→liv	lun→lun	liv→lun
All genes	96.51%	56.98%	90.72%	45.32%
Common signature	95.23%	93.02%	95.94%	94.72%
Lung vs Prostate				
	lun→lun	pro→lun	pro→pro	lun→pro
All genes	90.72%	69.03%	93.94%	62.88%
Common signature	94.82%	94.45%	79.61%	72.76%
Liver vs Bladder				
	liv→liv	bla→liv	bla→bla	liv→bla
All genes	96.51%	62.79%	88.60%	49.65%
Common signature	91.74%	91.86%	98.25%	98.25%
Prostate vs Bladder				
	pro→pro	bla→pro	bla→bla	pro→bla
All genes	93.94%	36.30%	88.60%	42.63%
Common signature	92.92%	86.86%	97.81%	88.25%
Lung vs Bladder				
	lun→lun	bla→lun	bla→bla	lun→bla
All genes	90.72%	51.87%	88.60%	50.88%
Common signature	89.38%	85.91%	97.37%	85.61%

Table 3.4 PPI summary of within-cancer-type and inter-cancer-type predictions in batch II analysis.

		test data			
		Liver	Prostate	Lung	Bladder
training data	Liver	96.5% (69)*	94.1% (225) +	94.7% (119) +	98.3% (53) +
	Prostate	97.7% (225)+	93.9% (55)*	94.5% (288) +	88.3% (10) +
	Lung	93.0% (119) +	72.8% (288) +	90.7% (57)*	85.6% (19) +
	Bladder	91.9% (53) +	86.9% (10) +	85.9% (19) +	88.6% (135)*

*: All genes are used in the within-cancer-type prediction to allow PAM for automatic predictive gene selection. Numbers of genes used in PAM are shown in parentheses.
+: In all inter-cancer-type predictions, only common signature genes are used in PAM and PAM does not perform further gene selection. The numbers of genes appeared more than 70% of leave-one-out cross validations are shown in the parentheses (i.e. liv-pro-PBs, liv-lun-PBs and pro-lun-PBs).

Table 3.5 The 44 batchII-PBs overlapped by pair-wise comparisons of liver, prostate and lung data sets.

Probe Set ID	Gene Title	Gene Symbol
39597_at*	actin binding LIM protein family, member 3	ABLIM3
37599_at*	aldehyde oxidase 1	AOX1
34736_at*	cyclin B1	CCNB1
37302_at*	centromere protein F, 350/400ka (mitosin)	CENPF
37203_at*	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	CES1
32168_s_at*	Down syndrome critical region gene 1	DSCR1
34311_at*	glutaredoxin (thioltransferase)	GLRX
1737_s_at*	insulin-like growth factor binding protein 4	IGFBP4
609_f_at*	metallothionein 1B	MT1B
36130_f_at*	metallothionein 1E	MT1E
31622_f_at*	metallothionein 1F	MT1F
39594_f_at*	metallothionein 1H	MT1H
35699_at	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	BUB1B
38796_at	complement component 1, q subcomponent, B chain	C1QB
35276_at	claudin 4	CLDN4
36668_at	cytochrome b5 reductase 3	CYB5R3
33295_at	Duffy blood group, chemokine receptor	DARC
41225_at	dual specificity phosphatase 3 (vaccinia virus phosphatase VH1-related)	DUSP3
38052_at	coagulation factor XIII, A1 polypeptide	F13A1
37743_at	fasciculation and elongation protein zeta 1 (zygin I)	FEZ1
38326_at	G0/G1switch 2	G0S2
1597_at	growth arrest-specific 6	GAS6
411_i_at	interferon induced transmembrane protein 2 (1-8D)	IFITM2
37484_at	integrin, alpha 1	ITGA1
38116_at	KIAA0101	KIAA0101
37883_i_at	Hypothetical gene supported by AK096951	LOC400879
242_at	microtubule-associated protein 4	MAP4
31623_f_at	metallothionein 1A	MT1A
39081_at	metallothionein 2A	MT2A
37736_at	protein-L-isoaspartate (D-aspartate) O-methyltransferase	PCMT1
35752_s_at	protein S (alpha)	PROS1
34163_g_at	RNA binding protein with multiple splicing	RBPMS
34887_at	radixin	RDX
39150_at	ring finger protein 11	RNF11
41096_at	S100 calcium binding protein A8	S100A8
33443_at	serine incorporator 1	SERINC1
39775_at	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1, (angioedema, hereditary)	SERPING1
1798_at	solute carrier family 39 (zinc transporter), member 6	SLC39A6
33131_at	SRY (sex determining region Y)-box 4	SOX4
40419_at	stomatin	STOM
1897_at	transforming growth factor, beta receptor III	TGFBR3
38404_at	transglutaminase 2 (C polypeptide, protein-glutamine-gamma-glutamyltransferase)	TGM2
40145_at	topoisomerase (DNA) II alpha 170kDa	TOP2A
35720_at	WD repeat domain 47	WDR47

The first 12 genes with asterisk overlapped batchI-PBs.

Subsequently, we expanded our analysis to prostate, lung, liver and bladder data sets (batch II analysis) with only normal and tumor tissues to test whether common signature genes can be found across these four types of cancers. Similar analyses were performed except that ANOVA was replaced by t-test for two class normal and tumor comparison. Each pair of the cancer types was analyzed. Similar to batch I analysis, only common signature genes with consistent regulation direction (up-regulation or down-regulation) in both cancer types were selected. Table 3.3 (see also Supplement Table B.3 for the entire confusion matrix results) summarizes the prediction results of batch II analysis. Similar to the result of batch I analysis, we observed high prediction accuracy for within-cancer-type prediction when using all genes in PAM (96.5% for liver, 93.9% for prostate, 90.7% for lung and 88.6% for bladder). The prediction models using common signature biomarkers generated similar high accuracy compared to using all genes (91.7%-97.7% in liver, 79.6%-95.6% in prostate, 89.4%-96.0% in lung and 97.4%-98.3% in bladder). The result confirms that the common signature biomarkers carry as good predictive information as the entire 5,917 genes. For the inter-cancer-type classification analysis, we repeatedly found that prediction with all genes did not perform well. In contrast, using common signature genes achieved much superior performance (Table 3.4). Liver particularly seemed to be the most robust either used as training or test data. Bladder, however, showed slightly lower cross-predictability with the other three cancer types. The numbers of common signature genes of bladder with other cancer types are also much smaller. Following the same criterion of selecting 70% frequency of being selected as common signatures in the cross-validations, we identified predictive biomarkers of the comparison in each pair of cancer types in Table 3.4 (255 liv-pro-PBs, 119 liv-lun-PBs, 288 lun-pro-PBs, 53 liv-bla-PBs, 10 pro-bla-PBs and 19 lun-bla-PBs). When all possible pairs of comparisons among liver, prostate and lung are overlapped (liv-

pro-lun-PBs), a number of 47 genes was identified. After deleting replicates, 44 (out of 47) distinct predictive biomarkers in liver, prostate and lung cancers were identified as batchII-PBs (Table 3.5). However, these common signature genes do not overlap with those from bladder data set, indicating a lack of common signature between these cancers and bladder cancer. There are 12 overlapping genes (Figure 3.5; p<1E-10 with significantly high overlapping) between batchI-PBs and batchII-PBs (marked with asterisk in Table 3.5 and Supplement Table B.4). Pathway analysis was performed on these predictive biomarkers indicating that fewer numbers of predictive biomarkers and GO terms were identified when bladder samples were analyzed in the inter-cancer-type prediction.

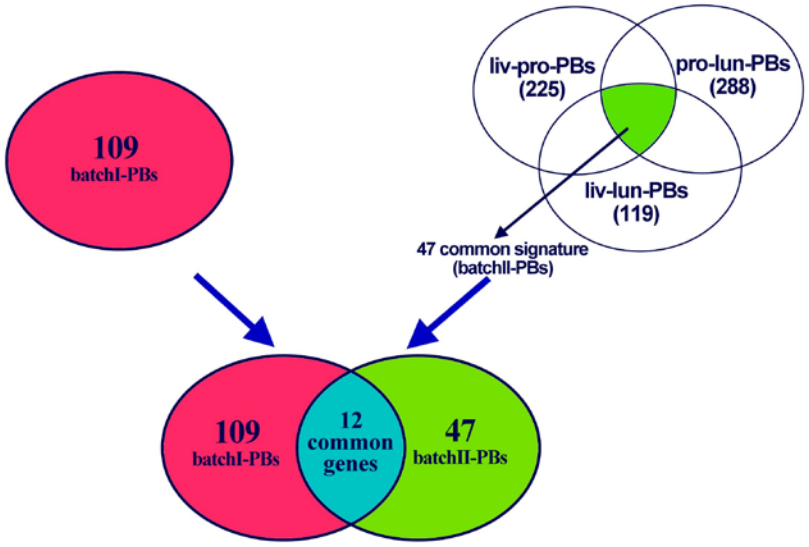

Figure 3.5 Diagram of batchI-PBs and batchII-PBs and their intersection genes.
The 47 batchII-PBs are listed in Table 3.5 and 109 batchII-PBs are listed in Supplement Table B.4.

To validate the robustness and cross-predictability of batchII-PBs, a data set of 23 independent prostate cancer samples obtained from another institute (LaTulippe, *et al*. 2002) was evaluated. The prediction model based on the 47 batchII-PBs in the 64 normal and tumor liver

samples achieved 96% (22/23) accuracy in predicting the 23 independent prostate samples (the "EV_liv→pro" scheme in Figure 3.3C). Evaluation of "EV_pro→pro" and "EV_lun→pro" also gave the same results (96% accuracy). Since we only have tumor samples in the external prostate data, there is a potential pitfall that the high accuracy may be an accidental result of study discrepancies between the new 23 prostate samples and the normal and tumor samples in analyzed data sets. We performed multi-dimension scaling (MDS) plots to visualize the new and old samples and excluded this possibility (Figure 3.6). The new prostate tumor samples are scattered and mixed with the old tumors but separated from old normal samples. As a result, the high accuracy of the prediction on this new data set is not caused by pure "accident".

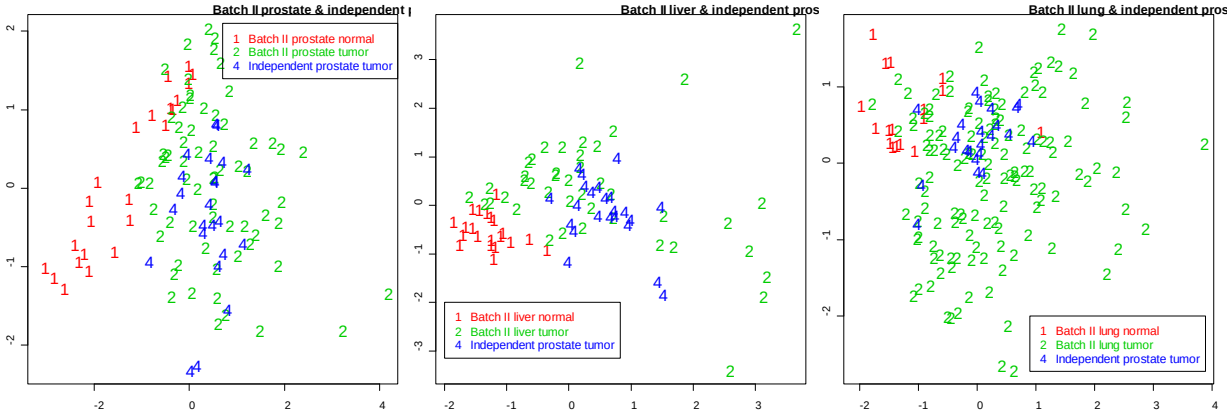

Figure 3.6 MDS plot of existing training data set and independent prostate cancer data.

Three MDS plots of the existing liver, prostate and lung training data sets respectively with the 23 independent prostate tumor samples. The mixing of the 23 tumor samples and old tumor samples exclude the possibility of accidental high accuracy due to study differences.

3.5 DISCUSSION

Meta-analyses had been performed for several types of human malignancies (Best, *et al*. 2003; Mehra, *et al*. 2005; Ramaswamy, *et al*. 2003; Rhodes, *et al*. 2002; Rubin, *et al*. 2004; Warnat, *et al*. 2005). However, to our knowledge, this is the first report showing that a microarray gene

expression model demonstrates inter-cancer predictability between different types of cancers using the identified predictive biomarkers. These results not only were evaluated in cross-validation analysis of existing working data sets but also were validated by independent prostate tissues collected and preprocessed separately. This argues strongly in favor of the reproducibility of the predictive biomarkers and the models. The 44 batchII-PBs appear to represent the common gene expression alteration among hepatocellular carcinoma, lung and prostate cancer. They follow similar patterns of differential expression in normal and tumor tissues for prostate, lung and liver cancer. Surprisingly, these gene signatures predict prostate, lung and hepatocellular carcinoma with similarly high accuracy as using the entire genome information of 5917 genes in each within-cancer-type prediction in prostate, lung or liver cancer. This suggests that the 44 genes are the major determinant of gene expression alteration in these three types of cancers.

The high level of inter-organ cancer predictability using just 44 genes implies that the core of cancer gene alterations may actually be quite small. The alterations of the expression of these genes could represent the common features of the three types of malignancies. None of these genes was, however, identified as the most significantly altered in bladder cancer suggest the dis-resemblance of bladder cancer to these three types of cancers. Among these genes includes a interferon inducible protein, *1-8D* (IFITM2, 411_i_at). This gene was a known important mediator of interferon induced in cell growth inhibition and induction of cell death (Deblandre, *et al*. 1995; Lewin, *et al*. 1991). 1-8D was down-regulated in hepatocellular carcinoma, lung cancer and prostate cancer, while pro-growth genes such as *cyclin B1* (CCNB1, 34736_at) was significantly up-regulated in three types of tumor samples. Other genes involving in growth controls including growth arrest specific 6 (GAS6, 1597_at), G0/G1swtich 2 (GOS2, 38326_at)

are also abnormally expressed in these tumors. The 44 gene list also includes six metallothioneins including 1A, 1B, 1E, 1F, 1H and 2A (MT1A, 31623_f_at; MT1B, 609_f_at; MT1E, 36130_f_at; MT1F, 31622_f_at; MT1H, 39594_f_at; MT2A, 39081_at). Metallothioneins are some low molecular weight zinc binding proteins that play important role in regulating transcriptional activity for variety of genes, and play crucial role in zinc signaling (Foster, *et al*. 1988; Tsuji, *et al*. 1992). Abnormal up-regulation of these genes may result in global pattern of gene expression alteration. Up-regulation of metallothioneins were thought to contain prognostic value in invasive ductal breast cancer (Schmid, *et al*. 1993). CCNB1 and most of the metallothioneins were also identified in batchI-PBs where adjacent tissues were included in the analysis. In the pathway analysis, we also observe many cancer related functional categories, including “mitotic checkpoint”, “apoptotic program”, “copper ion binding” and “cadmium binding”. Investigation into the abnormalities of these pathways may yield important insight into the common carcinogenesis mechanism of the tumors.

The clinical implication of our finding is two-fold: If the prediction of hepatocellular carcinoma, lung cancer and prostate cancer using our 44 batchI-PBs is interchangeable, we like to hypothesize that the abnormalities in the expression of the 44 genes represent a common features of these malignancies. Therapeutic targeting toward some of these genes will be of significant value in treating these malignancies. Second, the 99 batchII-PBs predicts tissues adjacent to malignancies versus completely normal organ tissues with high accuracy. This model may be able to serve as a predictor of malignancies nearby even if a biopsy misses its tumor target. This may serve as an indicator for a quick follow-up re-biopsy until the tumor(s) is identified. Alternatively, the detection of a strong cancer field effect change may argue for some prophylactic treatments before morphological cancer appears.

4.0 CONCLUSION, DISCUSSION AND FUTURE WORK

The microarray technology has been proven useful in the biomedical field. The analysis of multiple microarray studies faces many challenges. My dissertation treats the statistical analyses of multiple microarray studies.

Literature investigating the same disease with different array platforms or in different labs is often reported with similar high disease prediction accuracies without mention of direct inter-study predictions. In section 2.2, we described the many technical difficulties that exist for direct application of prediction models to independent studies. Inter-study normalization is a critical step for inter-study prediction, especially when data sets are from different platforms. The first part of this dissertation aims to develop an improved and robust normalization method for the inter-study prediction. Our proposed method yielded better inter-study prediction results and the improvement is more obvious for inter-platform prediction.

In the literature, multiple microarray studies for the same disease are very common. However, meta-analysis to integrate multiple studies has rarely been investigated. Inter-study prediction for better disease diagnosis and information integration for better biomarker detection are important issues to enhance the clinical utility of microarray. Therefore, our other goal is to detect the common predictive biomarkers in the microarray studies of four different cancer types. Based on the gene list we identified, high cross prediction accuracy was achieved. Although we

have talked about the conclusions and discussions in each chapter, we will summarize it and demonstrate our future works based on what we have found in this dissertation.

4.1 RATIO-ADJUSTED GENE-WISE NORMALIZATION

4.1.1 Conclusion

We draw the following conclusions from the ratio-adjustment gene-wise normalization project: most of the time, sample-wise normalization is helpful, but is not enough in inter-study prediction, especially for inter-platform prediction. Thus we turn to gene-wise normalization. The sample size ratio of different disease groups in one study has influence on the gene-wise normalization. Our proposed method, SN_std+rGN_std with calibration had accurate and robust results in the inter-study and inter-platform prediction. GN can also be used in analyses involving principle components analysis.

4.1.2 Discussion

In section 2.5, we focused our discussions on our proposed normalization method, particularly on its strength, application, and limitations. Here, we will discuss two issues in a bigger scope, the inter-study and inter-platform prediction.

First, one important issue regarding inter-study analysis is gene-matching. From Table 2.1, we see the number of overlapping genes between the two Affymetrix sets: Harvard and Michigan have 2,494 out of over 4,000 overlapping genes in each unmatched study. There are only around

1,500 genes remaining after merging between Affymetrix and cDNA sets. Therefore, we lose more than half of the information for the inter-study analysis. Even for these overlapping genes, some of those expression intensities are quite different due to the probe designs as well as other protocol problems as we introduced in section 2.2. Even worse, some genes could have opposite expression patterns across disease groups in different studies, which makes it more difficult to build a powerful and compact prediction model. As we see from Figures 2.5, 2.6, and 2.7, the PPI increases as more genes are used in the training set. Furthermore, this big model cannot be used to predict a third study directly due to the gene-matching problem. A metagene (i.e. a linear weighting of expression intensities of all genes) approach which captures the major, invariant biological features and reduces noise has often been used by researchers to combine and to compare different microarray studies (Shen, et al 2004; Pittman *et al*, 2004; Tamayo *et al* 2007*)*. However, this approach still cannot resolve the gene-matching problem. In other words, if only one element of the metagene - a single gene - is missing in the test set, this approach will not work.

Second, our proposed method, SN_std+GN_std outperforms SN_st in the inter-study prediction of the first two examples with binary outcome. For the survival prediction of the three lung cancer data sets, we only observed obvious improvement in one of the three pairs shown by Figure 2.9. However, the significant improvement exists in all three within-study leave-one-out predictions shown in Figure 2.8. We also observed that the sample sizes of the two risk groups after SN_std+GN_std are more even as they are supposed to be. Although we explained it through the PCA data in section 2.5.3, there are still some questions that remain unanswered. For example, how can the inter-study survival prediction be improved? How important is the ratio of

sample size in true low and high risk groups? How can our proposed method be extended to studies involving Principle Component Analysis (PCA)?

4.1.3 Future work

In the first part of the above discussions, we listed a few problems for gene-based inter-study prediction and while the metagene approach is popular and useful, it has its weaknesses. An ongoing project of Zhibao Mi, one of our group members, is performing methodological exploration on module-based prediction. This project is being conducted under the guidance of Dr. George Tseng, and I have participated in this project as well. This approach does not require complete gene matching and is robust to the existence of non-matched genes across studies. The cross prediction is based on the modules generated by the clustering method on the entire genome before merging. For example, if a module in the training set consists of ten genes, it will work as long as at least one of these ten genes exists in the test set to form the corresponding module. Our proposed method, SN_std+rGN_std is used for inter-study and inter-platform normalization. In the future, I will work on two aspects of this project: 1) cluster size estimation on various level of gene overlapping and 2) comparison of gene-based approaches with module-based approach.

For the second issue in section 4.1.2 concerning survival prediction, we plan to do further research on the three lung cancer data sets. Regarding the improvement of within-study survival prediction, we will first pull out the top 50 genes selected for SuperPC, and then investigate each of them for their contribution and influence in the Principle Component Analysis (PCA). We will repeat and compare this for raw data, for example,data after applying SN_std and SN_std+GN_std. We will find out which genes dominate the PCA without additional GN_std

and see how these genes behave in the survival analysis. We will work on improving the inter-study prediction if it is possible. At last, we will employ more examples for the prediction of prognosis, such as these three famous publicly available breast cancer data sets: Sorlie(2001), VanDeVijver(2002) and Huang(2003).

4.2 COMMON PREDICTIVE BIOMARKERS IN MULTIPLE CANCER TYPES

4.2.1 Conclusion

The identified common predictive biomarkers achieved high accuracy similar to using whole genomes in the within-cancer-type and inter-cancer-type prediction. They also performed superior to the method using the whole genome in inter-cancer-type prediction. Compact lists of predictive biomarkers are important in cancer development and represent the common signatures of malignancies of multiple cancer types.

4.2.2 Discussion

This project is more focused on the interpretation of the detected common biomarkers from the microarray data of four cancer types. Although we did not apply sophisticated statistical methods, the result serves as a starting point for information integration of such multi-class data sets. We performed an ANOVA analysis on the liver and prostate cancer sets in batch I individually, and a t-test was used for the analysis of batch II. Next, we selected the significant genes by certain criteria and merged them to get the common predictive biomarkers. Because all

the studies are from the same Affymetrix chip, there are not as many technical difficulties as we encountered in the first project. Strictly speaking, it is not a traditional meta-analysis. It would be better if we analyze the data again by more rigorous statistical meta-analysis methods. For example, Fisher's (1932) equal weight approach uses the test statistic involving the log-transformation of p-values to Chi-square scores. Tippett's (1931) approach uses the minimum p-value instead.

4.2.3 Future work

One of our group members, Jia Li, under the supervision of Dr. George Tseng, compared most of the current meta-analysis methods regarding detection of differently expressed (DE) genes. They illustrated two statistical hypothesis settings and proposed an optimally weighted statistic and a maximum p-value statistic for the two questions, respectively. Permutation analysis was then applied to control the false discovery rate (FDR). They further showed the advantage of their proposed test procedures over existing methods by power comparison, simulation study and real data analyses of a multiple-tissue energy metabolism mouse model data sets and prostate cancer data sets. In the future, we will adopt their approach to reanalyze the four cancer data sets. We will compare the results of their method with the results we got from our simple method regarding the gene list and prediction accuracy.

APPENDIX A: SUPPLEMENT MATERIAL OF CHAPTER 2

A. 1 CROSS PREDICTION FIGURES FOR LUNG CANCER DATA

LDA was used for all pair-wise cross prediction for the three lung cancer data sets. Figure A.1 display the prediction results comparing raw data and data after three normalization approaches: SN_std, SN_std+GN_std and SN_std+rGN_std. As we know, the number of genes in the training model of LDA cannot exceed the number of samples; thus we varied the gene numbers from 5 to 30. The prediction results are not stable based on these genes. Contradicting genes which have different expression patterns across studies could be selected in the training model and those genes may destroy the cross prediction severely. This phenomenon happens more often in the inter-platform scenario than the intra-platform scenario. This is why we observed a dramatic decrease of prediction accuracy in subfigures D, E and F in Figure A.1. Overall, line 4 (SN_std+rGN_std) has the best prediction performance.

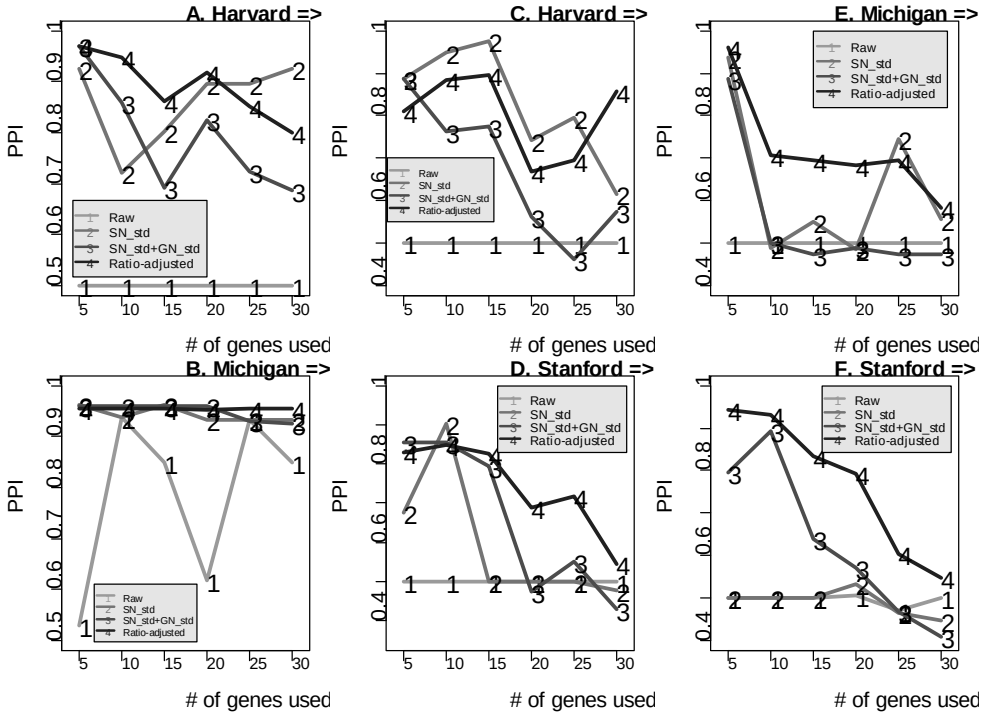

Figure A.1 Inter-study prediction by LDA

Figure A.2 shows the inter-study prediction results by KNN with K fixed at 5 and the number of genes in the training model varying from 5 to 500 like the prediction by PAM. KNN has very similar performance to PAM as shown in Figure 2.5. For the cross prediction between the Harvard and Michigan studies, except for in the raw data, all three normalization approaches achieve good cross-predictions. In the inter-platform prediction, line 4 (SN_std+rGN_std) outperforms the other two normalization methods and yields very accurate and stable prediction performance. As we observed from Figure A.2F, line 4 has PPIs over 80% all the time, while line 1, line 2, and line 3 completely failed with the prediction.

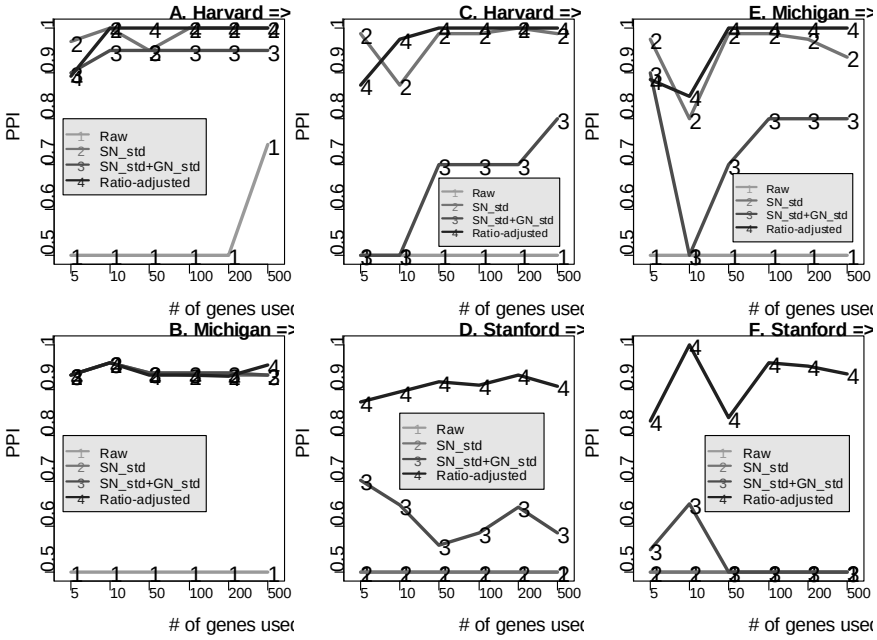

Figure A.2 Inter-study prediction by KNN

Figures A.3 and A.4 show the inter-study predictions of three lung cancer data sets by applying the normalization method: SN_std+rGN_std with calibration using prediction tools LDA and KNN respectively. Both methods require at least two normal and two adencarcinoma samples for calibration and both yield satisfactory normalization.

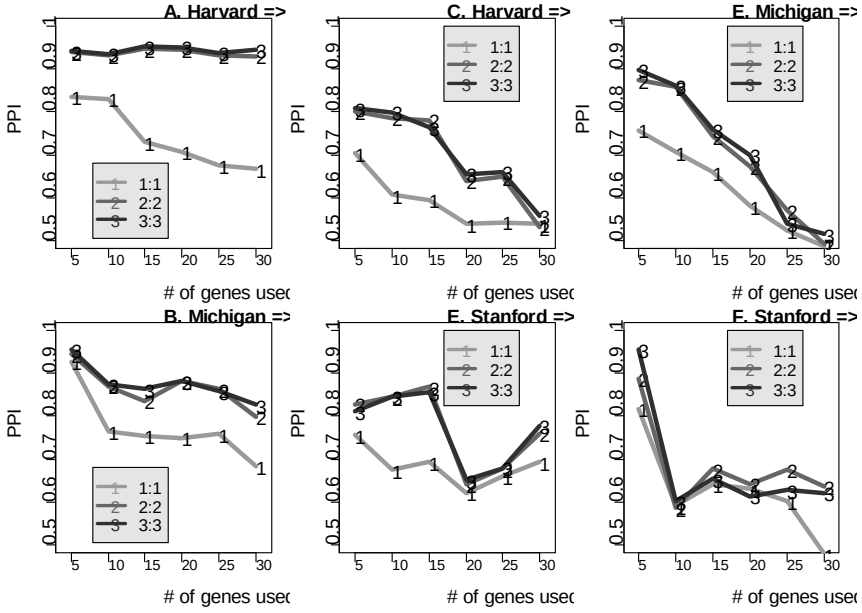

Figure A.3 Inter-study prediction applying SN_std+rGN_std with calibration by LDA

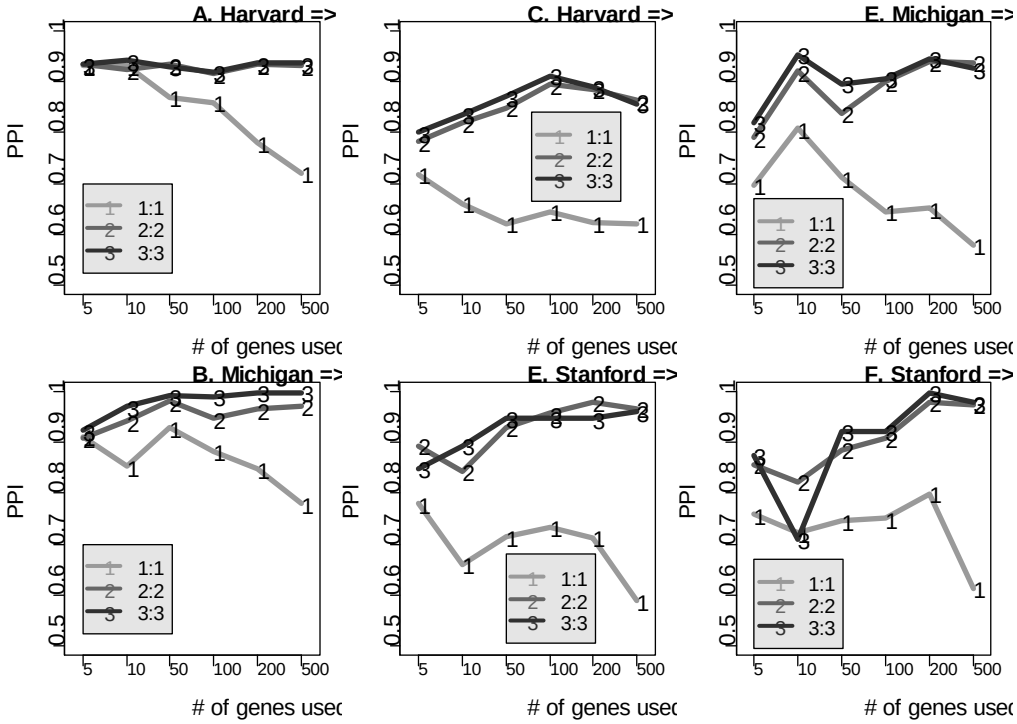

Figure A.4 Inter-study prediction applying SN_std+rGN_std with calibration by KNN

A. 2 CROSS PREDICTION FIGURES FOR PROSTATE CANCER DATA

Figure A.5 and A.6 show the inter-study prediction of two prostate cancer data sets comparing raw data and data after three normalization approaches: SN_std, SN_std+GN_std and SN_std+rGN_std using LDA and KNN respectively. Similar to the lung cancer studies, the LDA method is not as stable as PAM and KNN. Also, line 4 (SN_std+rGN_std) has the best performance for both methods.

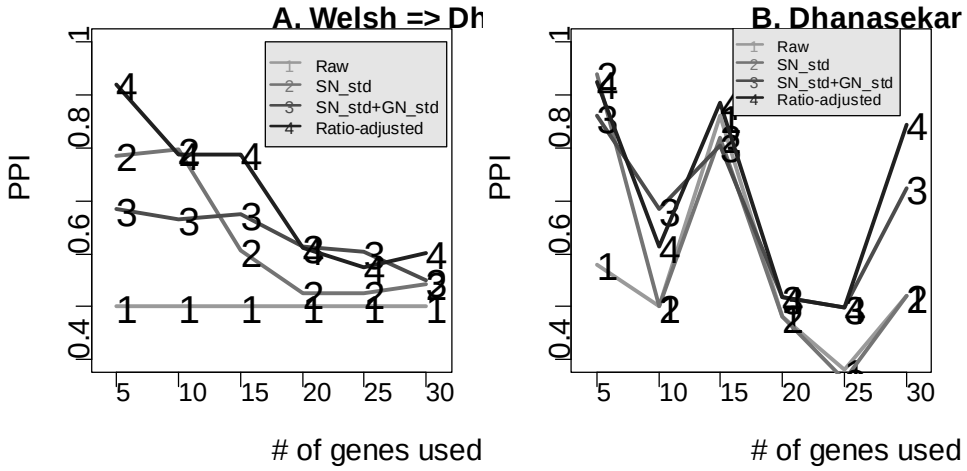

Figure A.5 Inter-study prediction of two prostate studies by LDA

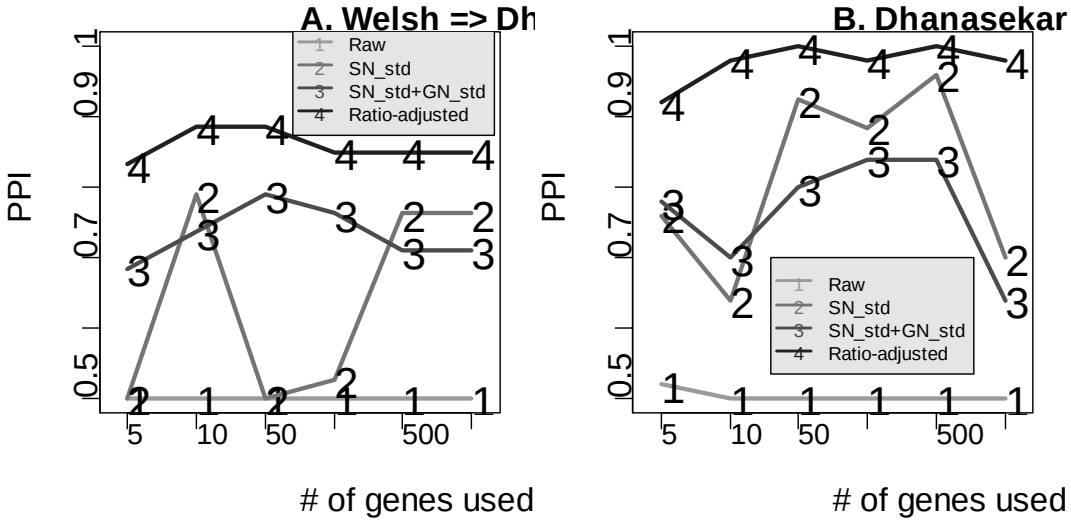

Figure A.6 Inter-study prediction of two prostate studies by KNN

APPENDIX B: SUPPLEMENT MATERIAL OF CHAPTER 3

B. 1 BOOTSTRAP PROCEDURE FOR GENE-WISE NORMALIZATION

Conceptually we standardize each gene vector to mean 0 and standard deviation 1 to accommodate different expression range of a predictive biomarker across different studies (e.g. APBA2BP, SLC39A14, AGT, TOP2A and B2M in Figure 3.2). Since the ratios of normal and tumor groups can vary in different studies, simple standardization can cause bias and deteriorate the prediction performance. Instead we perform bootstrap to sample a gene vector of B=1,000 samples in each group and standardize the vector of 2,000 (3,000 if N, A and T groups are all compared) bootstrapped samples to mean 0 and standard deviation 1 to estimate the standardization factors. Essentially we perform standardization under the simulated condition that normal and tumor groups have the same sample sizes.

Table B.1 An example of confusion matrix.

	true normal tissues	true tumor tissues
predicted as normal tissues	1	1
predicted as tumor tissues	5	41

Two false negatives and five false positive are made in the prediction, which sum up to seven total errors (with 42/48=87.5% overall accuracy). The sensitivity is 41/42=97.6%, specificity 1/6=16.7% and prediction performance index (PPI) (97.6%+16.7%)/2 =57.2%.

Table B.2 Batch I leave-one-out cross validation analysis result (confusion matrix)

Liver vs Prostate (Normal vs Tumor)												
			liver→liver		prostate→liver				prostate→prostate		liver→prostate	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	69	Predicted N	21	3	21	29	55	Predicted N	23	8	19	58
		Predicted T	0	40	0	14		Predicted T	0	58	4	8
Common signature	111.3	Predicted N	21	3	20	4	111.9	Predicted N	23	1.6	22	2
		Predicted T	0	40	1	39		Predicted T	0	64.4	1	64
Liver vs Prostate (Normal vs Adjacent)												
			liver→liver		prostate→liver				prostate→prostate		liver→prostate	
			True N	True A	True N	True A			True N	True A	True N	True A
All genes	66	Predicted N	20	3	18	9	63	Predicted N	23	4	15	33
		Predicted A	1	27	3	21		Predicted A	0	55	8	26
Common signature	111.2	Predicted N	21	1.1	20	1	110.4	Predicted N	23	2	23	4
		Predicted A	0	28.9	1	29		Predicted A	0	57	0	55
Liver vs Prostate (Adjacent vs Tumor)												
			liver→liver		prostate→liver				prostate→prostate		liver→prostate	
			True A	True T	True A	True T			True A	True T	True A	True T
All genes	64	Predicted A	27	13	13	17	266	Predicted A	44	21	46	44
		Predicted T	3	30	17	26		Predicted T	15	45	13	22
Common signature	111.5	Predicted A	27	16.7	26	16	112.0	Predicted A	42	25	41	26
		Predicted T	3	26.3	4	27		Predicted T	17	41	18	40

*The numbers marked in dark gray are the number of genes used to construct the prediction model. When "all genes" are used, the PAM method performs automatic gene selection to construct the model. When "common signature genes" are used, no gene selection is performed in PAM and the results (number of genes and confusion matrix) shown are averages of leave-one-out cross-validation results.

Table B.3 A Batch II leave-one-out cross validation analysis result (confusion matrix)

Liver vs Prostate												
			liver→liver		prostate→liver				prostate→prostate		liver→prostate	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	69	Predicted N	21	3	21	29	55	Predicted N	23	8	19	58
		Predicted T	0	40	0	14		Predicted T	0	58	4	8
Common signature	225.9	Predicted N	21	2	21	2	222.0	Predicted N	21.3	1	21	2
		Predicted T	0	41	0	41		Predicted T	1.7	65	2	64
Liver vs Lung												
			liver→liver		lung→liver				lung→lung		liver→lung	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	69	Predicted N	21	3	21	37	57	Predicted N	16	17	13	115
		Predicted T	0	40	0	6		Predicted T	1	117	4	19
Common signature	120.1	Predicted N	21	4.1	21	6	120.3	Predicted N	16	3	16	6
		Predicted T	0	38.9	0	37		Predicted T	1	131	1	128
Lung vs Prostate												
			lung→lung		prostate→lung				prostate→prostate		lung→prostate	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	57	Predicted N	16	17	17	83	55	Predicted N	23	8	23	49
		Predicted T	1	117	0	51		Predicted T	0	58	0	17
Common signature	289.5	Predicted N	16	6	16	7	288.8	Predicted N	17	9.7	15	13
		Predicted T	1	128	1	127		Predicted T	6	56.3	8	53
Liver vs Bladder												
			liver→liver		bladder→liver				bladder→bladder		liver→bladder	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	69	Predicted N	21	3	21	32	135	Predicted N	5	13	4	46
		Predicted T	0	40	0	11		Predicted T	0	44	1	11
Common signature	51.7	Predicted N	21	7.1	21	7	54.5	Predicted N	5	2	5	2
		Predicted T	0	35.9	0	36		Predicted T	0	55	0	55
Prostate vs Bladder												
			prostate→prostate		bladder→prostate				bladder→bladder		prostate→bladder	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	55	Predicted N	23	8	16	64	135	Predicted N	5	13	4	54
		Predicted T	0	58	7	2		Predicted T	0	44	1	3
Common signature	9.5	Predicted N	20.3	1.6	18	3	10.1	Predicted N	5	2.5	4	2
		Predicted T	2.7	64.4	5	63		Predicted T	0	54.5	1	55
Lung vs Bladder												
			lung→lung		bladder→lung				bladder→bladder		lung→bladder	
			True N	True T	True N	True T			True N	True T	True N	True T
All genes	57	Predicted N	16	17	17	129	135	Predicted N	5	13	5	56
		Predicted T	1	117	0	5		Predicted T	0	44	0	1
Common signature	19.1	Predicted N	14.9	11.9	15	22	19.2	Predicted N	5	3	4	5
		Predicted T	2.1	122.1	2	112		Predicted T	0	54	1	52

The confusion matrixes in the gray shaded regions are used to generate the PPI in shaded regions in Table 3.3 and corresponding Table 4.

Table B.4 A List of 109 biomarkers identified in batch I

Probe Set ID	Gene Title	Gene Symbol
39597_at*	actin binding LIM protein family, member 3	ABLIM3
37599_at*	aldehyde oxidase 1	AOX1
34736_at*	cyclin B1	CCNB1
37302_at*	centromere protein F, 350/400ka (mitosin)	CENPF
37203_at*	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	CES1
32168_s_at*	Down syndrome critical region gene 1	DSCR1
34311_at*	glutaredoxin (thioltransferase)	GLRX
1737_s_at*	insulin-like growth factor binding protein 4	IGFBP4
609_f_at*	metallothionein 1B	MT1B
36130_f_at*	metallothionein 1E	MT1E
31622_f_at*	metallothionein 1F	MT1F
39594_f_at*	metallothionein 1H	MT1H
41530_at	acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)	ACAA2
34050_at	acyl-CoA synthetase medium-chain family member 1	ACSM1
684_at	angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	AGT
32747_at	aldehyde dehydrogenase 2 family (mitochondrial)	ALDH2
33756_at	amine oxidase, copper containing 3 (vascular adhesion protein 1)	AOC3
41306_at	amyloid beta (A4) precursor protein-binding, family A, member 2 binding protein	APBA2BP
287_at	activating transcription factor 3	ATF3
201_s_at	beta-2-microglobulin	B2M
2011_s_at	BCL2-interacting killer (apoptosis-inducing)	BIK
39409_at	complement component 1, r subcomponent	C1R
40496_at	complement component 1, s subcomponent	C1S
1943_at	cyclin A2	CCNA2
33950_g_at	corticotropin releasing hormone receptor 2	CRHR2
408_at	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	CXCL1
649_s_at	chemokine (C-X-C motif) receptor 4	CXCR4
38772_at	cysteine-rich, angiogenic inducer, 61	CYR61
36643_at	discoidin domain receptor family, member 1	DDR1
33393_at	DEAD (Asp-Glu-Ala-As) box polypeptide 19B	DDX19B
32600_at	docking protein 4	DOK4
37827_r_at	dopey family member 2	DOPEY2
34823_at	dipeptidyl-peptidase 4 (CD26, adenosine deaminase complexing protein 2)	DPP4
36088_at	Down syndrome critical region gene 2	DSCR2
167_at	eukaryotic translation initiation factor 5	EIF5
1519_at	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	ETS2
36543_at	coagulation factor III (thromboplastin, tissue factor)	F3
1915_s_at	v-fos FBJ murine osteosarcoma viral oncogene homolog	FOS
36669_at	FBJ murine osteosarcoma viral oncogene homolog B	FOSB
39822_s_at	growth arrest and DNA-damage-inducible, beta	GADD45B

Table B.4 Continued		
290_s_at	G protein-coupled receptor 3	GPR3
35127_at	histone cluster 1, H2ae	HIST1H2AE
31521_f_at	histone cluster 1, H4k	HIST1H4J
152_f_at	histone cluster 2, H4a	HIST2H4A
38833_at	major histocompatibility complex, class II, DP alpha 1	HLA-DPA1
38096_f_at	major histocompatibility complex, class II, DP beta 1	HLA-DPB1
36878_f_at	major histocompatibility complex, class II, DQ beta 1	HLA-DQB1
37039_at	major histocompatibility complex, class II, DR alpha	HLA-DRA
36617_at	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein	ID1
676_g_at	interferon induced transmembrane protein 1 (9-27)	IFITM1
41745_at	interferon induced transmembrane protein 3 (1-8U)	IFITM3
37319_at	insulin-like growth factor binding protein 3	IGFBP3
36227_at	interleukin 7 receptor	IL7R
35372_r_at	interleukin 8	IL8
38545_at	inhibin, beta B (activin AB beta polypeptide)	INHBB
36355_at	involucrin	IVL
1895_at	jun oncogene	JUN
41483_s_at	jun D proto-oncogene	JUND
217_at	kallikrein-related peptidase 2	KLK2
35118_at	lecithin-cholesterol acyltransferase	LCAT
41710_at	hypothetical protein LOC54103	LOC54103
35926_s_at	lysozyme (renal amyloidosis)	LYZ
36711_at	v-maf musculoaponeurotic fibrosarcoma oncogene homolog F (avian)	MAFF
33146_at	myeloid cell leukemia sequence 1 (BCL2-related)	MCL1
33241_at	microfibrillar-associated protein 3-like	MFAP3L
668_s_at	matrix metallopeptidase 7 (matrilysin, uterine)	MMP7
870_f_at	metallothionein 3	MT3
36933_at	N-myc downstream regulated gene 1	NDRG1
37544_at	nuclear factor, interleukin 3 regulated	NFIL3
190_at	nuclear receptor subfamily 4, group A, member 3	NR4A3
31886_at	5'-nucleotidase, ecto (CD73)	NT5E
31733_at	purinergic receptor P2X, ligand-gated ion channel, 3	P2RX3
32210_at	phosphoglucomutase 1	PGM1
36980_at	proline-rich nuclear receptor coactivator 1	PNRC1
39366_at	protein phosphatase 1, regulatory (inhibitor) subunit 3C	PPP1R3C
36159_s_at	prion protein (p27-30) (Creutzfeldt-Jakob disease, Gerstmann-Strausler-Scheinker syndrome, fatal familial insomnia)	PRNP
216_at	prostaglandin D2 synthase 21kDa (brain)	PTGDS
1069_at	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	PTGS2
37701_at	regulator of G-protein signalling 2, 24kDa	RGS2
41471_at	S100 calcium binding protein A9	S100A9
33305_at	serpin peptidase inhibitor, clade B (ovalbumin), member 1	SERPINB1

Table B.4 Continued		
36979_at	solute carrier family 2 (facilitated glucose transporter), member 3	SLC2A3
38797_at	solute carrier family 39 (zinc transporter), member 14	SLC39A14
38994_at	suppressor of cytokine signaling 2	SOCS2
34666_at	superoxide dismutase 2, mitochondrial	SOD2
38763_at	sorbitol dehydrogenase	SORD
38805_at	TGFB-induced factor homeobox 1	TGIF1
39411_at	TCDD-inducible poly(ADP-ribose) polymerase	TIPARP
1715_at	tumor necrosis factor (ligand) superfamily, member 10	TNFSF10
904_s_at	topoisomerase (DNA) II alpha 170kDa	TOP2A
32793_at	T cell receptor beta variable 19	TRBC1
38469_at	tetraspanin 8	TSPAN8
40198_at	voltage-dependent anion channel 1	VDAC1
36909_at	WEE1 homolog (S. pombe)	WEE1
40448_at	zinc finger protein 36, C3H type, homolog (mouse)	ZFP36
32588_s_at	zinc finger protein 36, C3H type-like 2	ZFP36L2
1514_g_at		
1662_r_at		
40487_at	Transcribed locus	

A total of 109 biomarkers are identified in more than 70% of leave-one-out cross validation in batch I (batchI-PBs). After deleting duplicates, 99 distinct predictive biomarkers are listed below.

BIBLIOGRAPHY

[1] Bair E & Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**: E108.

[2] Beer DG, Kardia SL, Huang CC, *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* **8**: 816-824.

[3] Best CJ, Leiva IM, Chuaqui RF, *et al.* (2003) Molecular differentiation of high- and moderate-grade human prostate cancer by cDNA microarray analysis. *Diagn Mol Pathol* **12**: 63-70.

[4] Bhattacharjee A, Richards WG, Staunton J, *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**: 13790-13795.

[5] Bolstad BM, Irizarry RA, Astrand M & Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**: 185-193.

[6] Bosotti R, Locatelli G, Healy S, *et al.* (2007) Cross platform microarray analysis for robust identification of differentially expressed genes. *BMC Bioinformatics* **8 Suppl 1**: S5.

[7] Brown PO & Botstein D (1999) Exploring the new world of the genome with DNA microarrays. *Nat Genet* **21**: 33-37.

[8] Bussey KJ, Kane D, Sunshine M, *et al.* (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol* **4**: R27.

[9] Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO & Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. *Science* **282**: 699-705.

[10] Crick F (1970) Central dogma of molecular biology. *Nature* **227**: 561-563.

[11] Deblandre GA, Marinx OP, Evans SS, *et al.* (1995) Expression cloning of an interferon-inducible 17-kDa membrane protein implicated in the control of cell growth. *J Biol Chem* **270**: 23860-23866.

[12] Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3.

[13] DeRisi JL, Iyer VR & Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680-686.

[14] Dhanasekaran SM, Barrette TR, Ghosh D, *et al.* (2001) Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**: 822-826.

[15] Diehn M, Sherlock G, Binkley G, *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **31**: 219-223.

[16] Draghici S, Khatri P, Martins RP, Ostermeier GC & Krawetz SA (2003) Global functional profiling of gene expression. *Genomics* **81**: 98-104.

[17] Eisen MB, Spellman PT, Brown PO & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**: 14863-14868.

[18] Fishel I, Kaufman A & Ruppin E (2007) Meta-analysis of gene expression data: a predictor-based approach. *Bioinformatics* **23**: 1599-1606.

[19] Fisher RA (1932) *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

[20] Foster R, Jahroudi N, Varshney U & Gedamu L (1988) Structure and expression of the human metallothionein-IG gene. Differential promoter activity of two linked metallothionein-I genes in response to heavy metals. *J Biol Chem* **263**: 11528-11535.

[21] Garber ME, Troyanskaya OG, Schluens K, *et al.* (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* **98**: 13784-13789.

[22] Gleason DF (1966) Classification of prostatic carcinomas. *Cancer Chemother Rep* **50**: 125-128.

[23] Golub TR, Slonim DK, Tamayo P, *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531-537.

[24] Huang E, Cheng SH, Dressman H, *et al.* (2003) Gene expression predictors of breast cancer outcomes. *Lancet* **361**: 1590-1596.

[25] Huang E, Ishida S, Pittman J, *et al.* (2003) Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat Genet* **34**: 226-230.

[26] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B & Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**: e15.

[27] Irizarry RA, Warren D, Spencer F, *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**: 345-350.

[28] Jemal A, Ward E, Wu X, Martin HJ, McLaughlin CC & Thun MJ (2005) Geographic patterns of prostate cancer mortality and variations in access to medical care in the United States. *Cancer Epidemiol Biomarkers Prev* **14**: 590-595.

[29] Kuo WP, Liu F, Trimarchi J, *et al.* (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol* **24**: 832-840.

[30] Lamb J, Ramaswamy S, Ford HL, *et al.* (2003) A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**: 323-334.

[31] Larkin JE, Frank BC, Gavras H, Sultana R & Quackenbush J (2005) Independence and reproducibility across microarray platforms. *Nat Methods* **2**: 337-344.

[32] Li C & Hung Wong W (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2**: RESEARCH0032.

[33] Li J, Pankratz M & Johnson JA (2002) Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays. *Toxicol Sci* **69**: 383-390.

[34] Lipshutz RJ, Fodor SP, Gingeras TR & Lockhart DJ (1999) High density synthetic oligonucleotide arrays. *Nat Genet* **21**: 20-24.

[35] Lockhart DJ, Dong H, Byrne MC, *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**: 1675-1680.

[36] Luo JH, Ren B, Keryanov S, *et al.* (2006) Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology* **44**: 1012-1024.

[37] Maglott D, Ostell J, Pruitt KD & Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **33**: D54-58.

[38] Mehra R, Varambally S, Ding L, *et al.* (2005) Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Res* **65**: 11259-11264.

[39] Mitchell SA, Brown KM, Henry MM, Mintz M, Catchpoole D, LaFleur B & Stephan DA (2004) Inter-platform comparability of microarrays in acute lymphoblastic leukemia. *BMC Genomics* **5**: 71.

[40] Park PJ, Cao YA, Lee SY, Kim JW, Chang MS, Hart R & Choi S (2004) Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *J Biotechnol* **112**: 225-245.

[41] Parmigiani G, Garrett-Mayer ES, Anbazhagan R & Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res* **10**: 2922-2927.

[42] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP & Fodor SP (1994) Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci U S A* **91**: 5022-5026.

[43] Pittman J, Huang E, Dressman H, *et al.* (2004) Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proc Natl Acad Sci U S A* **101**: 8431-8436.

[44] Potti A, Dressman HK, Bild A, *et al.* (2006) Genomic signatures to guide the use of chemotherapeutics. *Nat Med* **12**: 1294-1300.

[45] Ramaswamy S, Ross KN, Lander ES & Golub TR (2003) A molecular signature of metastasis in primary solid tumors. *Nat Genet* **33**: 49-54.

[46] Rhodes DR, Barrette TR, Rubin MA, Ghosh D & Chinnaiyan AM (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* **62**: 4427-4433.

[47] Rubin MA, Varambally S, Beroukhim R, *et al.* (2004) Overexpression, amplification, and androgen regulation of TPD52 in prostate cancer. *Cancer Res* **64**: 3814-3822.

[48] Schena M, Shalon D, Davis RW & Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467-470.

[49] Schena M (1999) *DNA Microarrays: A Practical Approach* Oxford University Press, Oxford.

[50] Schmid KW, Ellis IO, Gee JM, *et al.* (1993) Presence and possible significance of immunocytochemically demonstrable metallothionein over-expression in primary invasive ductal carcinoma of the breast. *Virchows Arch A Pathol Anat Histopathol* **422**: 153-159.

[51] Segal E, Friedman N, Koller D & Regev A (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**: 1090-1098.

[52] Segal E, Friedman N, Kaminski N, Regev A & Koller D (2005) From signatures to models: understanding cancer using microarrays. *Nat Genet* **37 Suppl**: S38-45.

[53] Shabalin AA, Tjelmeland H, Fan C, Perou CM & Nobel AB (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* **24**: 1154-1160.

[54] Shen R, Ghosh D & Chinnaiyan AM (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* **5**: 94.

[55] Shi L & Reid LH & Jones WD, *et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**: 1151-1161.

[56] Shipp MA, Ross KN, Tamayo P, *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* **8**: 68-74.

[57] Sorlie T, Perou CM, Tibshirani R, *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**: 10869-10874.

[58] Stransky N, Vallot C, Reyal F, *et al.* (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat Genet* **38**: 1386-1396.

[59] Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW & Mesirov JP (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc Natl Acad Sci U S A* **104**: 5959-5964.

[60] Tan PK, Downey TJ, Spitznagel EL, Jr., *et al.* (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**: 5676-5684.

[61] Tibshirani R, Hastie T, Narasimhan B & Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**: 6567-6572.

[62] Tippett LHC (1931) *The Methods in Statistics*. Williams and Norgate,. Ltd., London.

[63] Tongbai R, Idelman G, Nordgard SH, *et al.* (2008) Transcriptional networks inferred from molecular signatures of breast cancer. *Am J Pathol* **172**: 495-509.

[64] Tseng GC, Oh MK, Rohlin L, Liao JC & Wong WH (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* **29**: 2549-2557.

[65] Tsuji S, Kobayashi H, Uchida Y, Ihara Y & Miyatake T (1992) Molecular cloning of human growth inhibitory factor cDNA and its down-regulation in Alzheimer's disease. *EMBO J* **11**: 4843-4850.

[66] van 't Veer LJ, Dai H, van de Vijver MJ, *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**: 530-536.

[67] Warnat P, Eils R & Brors B (2005) Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* **6**: 265.

[68] Warrington JA, Dee, S.,Trulson, M. (2000) Large-scale genomic analysis using Affymetrix GeneChip(R) probe arrays. *Microarray Biochip Technology*,(Schena M, ed.^eds.), p.^pp. 119-148. Eaton Publishing, Natick, MA.

[69] Welsh JB, Sapinoso LM, Su AI, *et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* **61**: 5974-5978.

[70] Woo Y, Affourtit J, Daigle S, Viale A, Johnson K, Naggert J & Churchill G (2004) A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. *J Biomol Tech* **15**: 276-284.

[71] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J & Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**: e15.

[72] Yauk CL & Berndt ML (2007) Review of the literature examining the correlation among DNA microarray technologies. *Environ Mol Mutagen* **48**: 380-394.

[73] Yu YP, 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.

[74] Yuen T, Wurmbach E, Pfeffer RL, Ebersole BJ & Sealfon SC (2002) Accuracy and calibration of commercial oligonucleotide and custom cDNA microarrays. *Nucleic Acids Res* **30**: e48.