ROBUST CROSS-PLATFORM DISEASE PREDICTION USING GENE EXPRESSION MICROARRAYS

by

Zhibao Mi

Ph.D. in Epidemiology, Institute of Microbiology and Epidemiology, Beijing China, 1992

M. Med. in Epidemiology, Tianjin Medical College, Tianjin China, 1987

B. Med. in Preventive Medicine, Shanxi Medical College, Taiyuan China, 1984

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

Zhibao Mi

It was defended on

October 15th, 2008

and approved by

Dissertation Advisor:
George C. 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:
Gong Tang, Ph.D. Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Naftalie Kaminski, M.D. Professor
Department of Medicine
School of Medicine
University of Pittsburgh

ROBUST CROSS-PLATFORM DISEASE PREDICTION USING GENE EXPRESSION MICROARRAYS

Zhibao Mi, PhD

University of Pittsburgh, 2008

Microarray technology has been used to predict patient prognosis and response to treatment, which is starting to have an impact on disease intervention and control, and is a significant measure for public health. However, the process has been hindered by a lack of adequate clinical validation. Since both microarray analyses and clinical trials are time and effort intensive, it is crucial to use accumulated inter-study data to validate information from individual studies. For over a decade, microarray data have been accumulated from different technologies. However, using data from one platform to build a model that robustly predicts the clinical characteristics of a new data from another platform remains a challenge. Current cross-platform gene prediction methods use only genes common to both training and test datasets. There are two main drawbacks to that approach: model reconstruction and loss of information. As a result, the prediction accuracy of those methods is unstable.

In this dissertation, a module-based prediction strategy was developed to overcome the aforementioned drawbacks. By the current method, groups of genes sharing similar expression patterns rather than individual genes were used as the basic elements of the model predictor. Such an approach borrows information from genes' similarity when genes are absent in test data. By overcoming the problems of missing genes and noise across platforms, this method yielded robust predictions independent of information from the test data. The performance of this method was evaluated using publicly available microarray data. *K*-means clustering was used to group

genes sharing similar expression profiles into gene modules and small modules were merged into their nearest neighbors. A univariate or multivariate feature selection procedures was applied and a representative gene from each selected module was identified. A prediction model was then constructed by the representative genes from selected gene modules. As a result, the prediction model is portable to any test study as long as partial genes in each module exist in the test study. The newly developed method showed advantages over the traditional methods in terms of prediction robustness to gene noise and gene mismatch issues in inter-study prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my gratitude to my advisor, Dr. George Tseng, who mentored and directed me through the whole process of my study and research. I am greatly thankful to the members of my dissertation committee for their guidance and encouragement in my research and the dissertation preparation. I also want to express my sincere appreciation of the efforts of all the professors who taught me and the departmental staff who helped me. My special thanks to Mr. Mike Gabrin and my colleagues at Precision Therapeutics, Inc., without their support, I would not have completed this work. Finally I want to thank my family and friends for their encouragement, love and support. I dedicate this work to my wife, Xiaoli and my son, Michael.

1.0 INTRODUCTION

1.1 MICROARRAY TECHNOLOGY AND ITS APPLICATION

Microarray technology is originated from a nucleotide hybridization technique, Southern blotting, where fragmented DNA are spotted on a supporting material, usually celluloid membrane and then probed with a known DNA fragment to identify unknown genes (Southern 1975). However, Southern blotting is limited for single or fewer gene processes. Early microarrays were spotted multiple cDNAs onto a miniaturized filter paper or glass slide to monitor panel of gene expression profile (Kulesh, Clive et al. 1987; Schena, Shalon et al. 1995). Now this miniaturized technology has extended to other molecular studies known as 'omics', e.g. genomics, transcriptomics, and evolved to many platforms using advanced technologies, such as Affymetrix GeneChip, GE (Amersham) Codelink, Illumina BeadChip, and Agilent SurePrint.

With the advance of genome sequencing, microarray technology has been developed rapidly in many aspects: from hundreds of gene probes to tens of thousands of gene probes, from spotted cDNA microarrays to photolithography oligonucleotide gene chips, from manual results reading system to automated data processing (Schulze and Downward 2001). Although various techniques are involved in microarray technology, the principle of microarray experiments is simply nucleotide hybridizations on micro-scaffolds. Use of this technology to systemically measure gene expression on a global level has evolved from large scale gene mapping and

sequencing (Poustka, Pohl et al. 1986; Cantor, Mirzabekov et al. 1992) to transcript level analysis and gene signaling pathway identification (Schena, Shalon et al. 1995; Schulze, Nicke et al. 2004), and it has even spread to develop gene signatures for disease classification and prognosis prediction (Luo, Duggan et al. 2001; Beer, Kardia et al. 2002; van 't Veer, Dai et al. 2002; Potti, Mukherjee et al. 2006). No other technology has drawn as much dedicated attention in the biomedical field, and microarrays led the way from related gene expression level to human global activities.

1.2 CLINICAL RISK PREDICTION USING GENOMIC TECHNOLOGIES

Microarray technology is becoming a promising methodology for predicting prognosis and response to treatment for cancer patients, and an emerging component for individualized medicine. Though many gene signatures developed from microarray technology reported have prediction values for various cancer patients, so far only few of them are in clinical practices (van 't Veer, Dai et al. 2002; Paik, Shak et al. 2004; Ross, Hatzis et al. 2008). The many others are suffering from either lack of a standardized molecular class prediction methods or independent clinical validations (Pusztai, Mazouni et al. 2006; Ioannidis 2007; Pusztai and Leyland-Jones 2008), which severely hinders the utility of individual genomic information. Since microarray analyses and clinical trials are expensive as well as time and effort intensive, therefore, to validate information and to predict patient outcomes from individual studies, it is crucial to utilize accumulated inter-study data. A stable prediction model requires features selected from a large training data set (Dobbin and Simon 2005; Dobbin, Zhao et al. 2008). For over a decade, microarray data have been accumulated from different array technologies or

different versions within technologies performed on similar clinical samples. However, to use a data set or integrated data sets from one platform to build a model that robustly and accurately predicts clinical characteristics of a new data set or a new sample from another platform remains a challenge (Tan, Downey et al. 2003; Park, Cao et al. 2004); however, the public accessible data provide an alternative to validate the genomic information when new clinical trials are inapproachable.

1.3 METHODOLOGICAL ISSUES IN INTER-STUDY MICROARRAY ANALYSES

Even though performing microarray experiment is straightforward with current automated microarray assay systems, accurate use of the genomic information from microarray analysis to classify patients or to predict patient prognosis is not trivial. An ideal approach to use genomic data for prediction includes microarray data pre-processing, gene selection, and model construction based on training study, and finally, the constructed model is validated on an independent test data. Commonly in literature, prediction models were cross validated only based on the same data used for the model construction (Pusztai and Leyland-Jones 2008). Cross validated models are usually under represented or over fitted due to a lack of heterogeneity of sampling and do not reveal cross platform problems when training data and test data are from different microarray platforms and protocols. A common cross platform problem is missing genes when genes in prediction model based on training data cannot be matched to test data. Traditional prediction methods using microarray data select individual genes as model components. This individual gene-based prediction (GBP) approach is sensitive to cross-platform missing genes.

Current cross-platform gene prediction methods use only those genes common to both training and test data sets (Shi, Tong et al. 2004; Irizarry, Warren et al. 2005; Shi, Tong et al. 2005). One drawback of this approach is that the prediction model has to be reconstructed, depending on the test data. Thus, the model cannot be created independent of knowledge about the test data. The model elements need to be adjusted every time test data is predicted.

In addition, because many genes in the training set and not in the test set are ignored, important information from the training set may be lost. Finally, the prediction accuracy of the GBP methods is unstable. In addition to missing genes, another reason for this instability is that these methods are sensitive to gene noise. In this study, a module-based prediction (MBP) method is developed to overcome these aforementioned drawbacks. In the MBP, groups of genes sharing similar expression patterns rather than individual genes are used as model predictors. Such an approach borrows information from genes' similarity when genes are absent in test sets. By overcoming gene noise and avoiding the problem of missing genes across platforms, the MBP method was hypothesized to yield robust predictions completely independent of information from the test data. The mechanism of the GBP versus the MBP is illustrated in Figure 1.1.

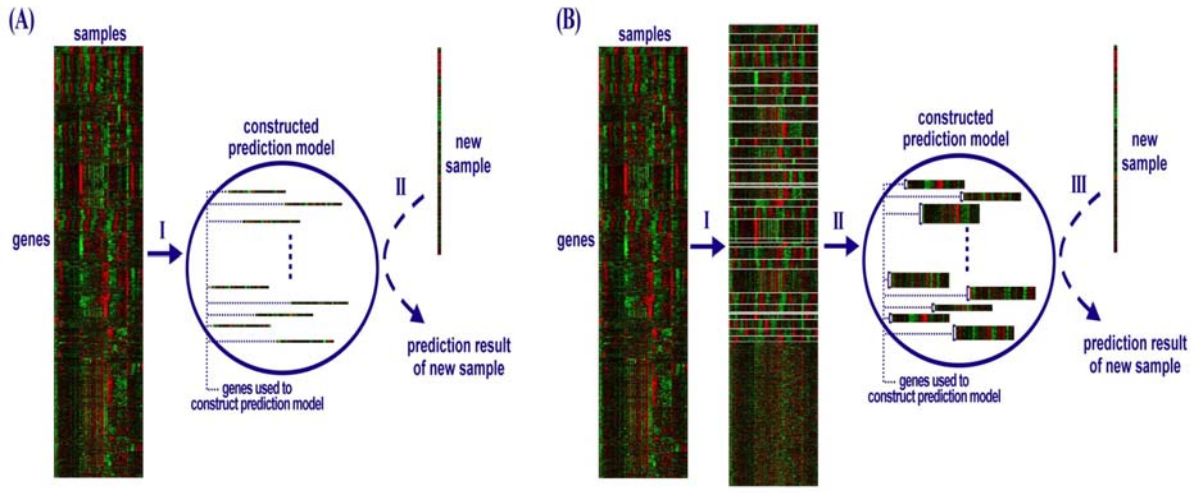

Figure 1.1 The GBP method versus the MBP method: The GBP method selects individual genes from training samples to construct a prediction model and uses the model to predict new sample (A); whereas, the MBP method selects groups of genes to form gene modules and uses these modules to construct a prediction model (B).

1.4 RELEVANT APPROACHES

Recently, there have been methods developed to use gene cluster information instead of individual gene information as predictors although they are not designed to solve the problems mentioned above. These methods can be summarized as three categories, metagenes, supergenes, and gene pathway modules, sharing a nature of using information of gene clusters. The metagene, defined as aggregate patterns of gene expression, was originally proposed by a group of researchers from Duke University (West, Blanchette et al. 2001; Spang, Zuzan et al. 2002; Huang, Cheng et al. 2003; Pittman, Huang et al. 2004; Potti, Mukherjee et al. 2006; West, Ginsburg et al. 2006). The metagene approach first dealt with an array dimension reduction either using supervised feature selection based on t test or correlation coefficient (Potti, Dressman et al. 2006) or unsupervised *K*-means clustering (Huang, Cheng et al. 2003), then took

the linear combination of the group of genes within cluster (metagene) based on the principal component of singular value decomposition (SVD) (Huang, Cheng et al. 2003; Potti, Dressman et al. 2006; Potti, Mukherjee et al. 2006) as predictors to fit prediction model (Figure 1.2.A). Meanwhile, they try to identify the biological pathways of the metagenes (Bild, Potti et al. 2006; Bild, Yao et al. 2006; Potti, Dressman et al. 2006). This method was then modified by group of researchers from MIT and Harvard University, whose metagene was extracted from a standard preprocessed array data applying nonnegative matrix factorization (NMF) to factor the resulting expression matrix and yield a metagene model by deriving the Moore-Penrose pseudoinverse matrix, and then the model data set was projected into metagene space and refined by trimming outliers using support vector machine (SVM) classifier (Figure 1.2.A). The refined model data set was then refactored using NMF and a refined projection map was established by pseudoinverse matrix used to analyze new test data set(Tamayo, Scanfeld et al. 2007). The goal of metagene methods was to obtain more accurate and stable prediction. A group of researchers from Stanford University took a slightly different approach from metagene. They focused on gene cluster method, named supergene, in one aspect, to control cluster reproducibility that a cluster defined in training dataset (model dataset) can be found in the test dataset (Kapp and Tibshirani 2007) and further to take average of gene expressions within cluster at each hierarchical level to fit a lasso regression model (Figure 1.2B), yielding a more accurate prediction results (Park, Hastie et al. 2007). Another approach of using information of group of genes for prediction is pathway module prediction, of which genes are grouped according to their functional pathways (Segal, Shapira et al. 2003; Segal, Friedman et al. 2004; Segal, Friedman et al. 2005; van Vliet, Klijn et al. 2007; Wong, Nuyten et al. 2008). This method tried to make gene clusters more enriched with biological meanings (Figure 1.2C). However, unlike the MBP

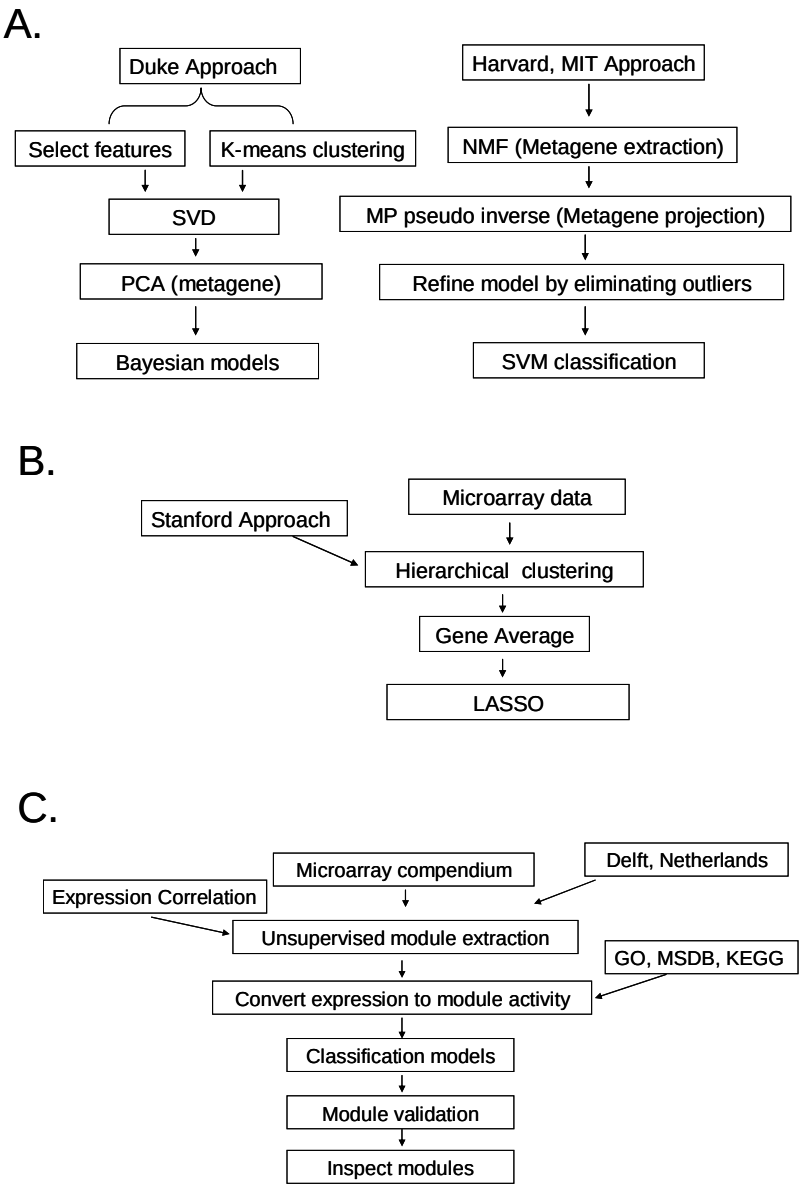

Figure 1.2 Approaches related to the MBP: Metagene approaches (A), Suppergene approach (B), Model based on biological pathway approach(C)

method being proposed, none of these methods was targeted to fully utilize gene information of training data or to deal with missing genes in the test dataset. A comparison of these methods is shown in Table 1.1. Their models still involve all individual genes. Missing one gene in the model will invalid the entire prediction.

Table 1.1 Comparison of prediction methods using gene cluster information

Methods	Authors	Institution	Data reduction	Feature selection	Model	Missing controlled	Test dependent
Metagene	Huang et al	Duke	K-means clustering	SVD	Bayesian decision tree	No	Yes
	Potti et al	Duke	Pearson correlation	SVD	Bayesian decision tree	No	Yes
	Tamayo et al	MIT, Harvard	NMF	pseudoinverse	SVM	No	Yes
Supergene	Park et al	Stanford	Hierarchical Clustering	average genes	LASSO	N/A	N/A
Pathway Module	van Vliet et al	Delft, Netherland	Pathway Compendia	rank of p-values	Bayes classifier	No	N/A
MBP	Mi et al	Pitt	K-means clustering	representative gene	Multiple classifiers	Yes	No

1.5 THE MBP VERSUS METAGENE APPROACH

Though both the MBP and metagene approaches use the information from group of genes they are different by nature. In metagene approach, a subset of genes is selected by either supervised (Potti, Dressman et al. 2006) or unsupervised method (Huang, Cheng et al. 2003), then projection method, SVD or NMF is used to identifying linear combinations of the subset of genes as metagenes (Potti, Dressman et al. 2006; Tamayo, Scanfeld et al. 2007). In this approach, if a gene selected in the subset of genes to form a metagene is missing in a test data the prediction would fail. Whereas in the MBP approach, all genes are grouped into *K* subsets by an unsupervised clustering method, then top *k* subsets of genes are selected by a supervised method, moderate t statistics (Tibshirani, Hastie et al. 2002), to form prediction modules. For each module, a representative gene of group of genes is selected to build prediction model. In this approach, if a gene selected in the prediction module is missing in a test data, a representative gene can be chosen from the remaining of genes in the module and the prediction would not fail. Further, for construction of a prediction model, only information from one gene subset is used by metagene approach, whereas, the information from multiple subsets is used by the MBP

approach. The differences indicate that the MBP is robust to gene missing and noise in inter-study microarray analyses, while, metagene approach is sensitive to the gene missing and noise.

1.6 HIGHLIGHT OF THE MBP

The MBP is developed to gain model robustness by solving gene missing and gene noise problems, not designed to increase prediction accuracy. As long as no significant loss of predictive power or accuracy, the MBP has its practical advantage of clinical utility over the GBP. The MBP development is motivated by two major issues existing in current GBP inter-study microarray analyses, i.e. information loss and model reconstruction when only common genes of both training and test data sets are used. To avoid the issues the prediction models need to be built using all genes in training dataset and are independent of test dataset. The GBP models are invalid if there are missing genes in the test data. The MBP can overcome the problem by borrowing information from the other genes within the same prediction module. Further, the MBP is robust to gene noise by selecting a representative gene from a prediction module. The prediction robustness to gene missing and noise is inherited from the MBP design.

1.7 THE GOALS OF THE PROJECT

The current study was designed to determine whether the MBP method has the similar predictive power as the GBP method both within-study and inter-studies and whether the method gained the prediction robustness in terms of gene missing and noise. The overall goal to build the MBP

model is to yield robust prediction models by solving gene noise and gene missing problems existing in the traditional GBP, and to obtain simple models by using all genes of a dataset to build prediction model without knowing information of test datasets. The following three goals are to be achieved in the dissertation project.

1.7.1 MBP algorithm development

Microarray data are preprocessed and standardized. The genes in the standardized array data are clustered by *K*-means and PW-*K*-means approaches. For those cluster size smaller than a threshold δ are merged into their nearest neighbors. The clusters are selected based on the moderated *t* statistics and the representative gene for each selected cluster was determined by minimum sum distance among the genes within the cluster. The final prediction model is constructed based on the selected representative genes.

1.7.2 Prediction failure control

An intrinsic disadvantage of using the GBP to build cross-platform prediction model is that gene missing problem can not be solved. The key of solving this problem using the MBP is that when a gene or multiple genes are missing in a test sample, the MBP method can borrow information from other genes in the same cluster by presenting a representative gene. When gene missing increases, the probability of the MBP method fails also increases. This especially happens when the cluster sizes are small and the cluster merging strategy is developed to avoid this problem. If a cluster size is smaller than the threshold δ, the cluster is merged into its nearest neighbor cluster to avoid the prediction failure due to clusters missing. The cluster sizes generated by *K*-means clustering could be viewed as a random vector depending on the original data matrix. A data mining is performed to explore the distribution of cluster sizes and in turn, generate

probability calculation under different number of genes, number of clusters and probability of missing genes.

1.7.3 The MBP performance evaluation

The newly developed MBP algorithm is tested on its prediction accuracy using publicly available microarray datasets. The prediction accuracy is assessed in both within-study and inter-studies scenarios and is compared with that of the GBP method using both univairate and multivariate feature selection methods. The MBP method is designed for model simplicity and robustness. The simplicity indicates that the method provides simple and easy use prediction model and the robustness indicates that the method can perform robust prediction with presence of gene missing and noise. Simulation studies are performed by randomly generating gene missing and noise in real gene expression profiles to evaluate the robustness of the MBP method.

2.0 EXPERIMENTAL DESIGN AND METHODOLOGY

2.1 DATA SETS AND GENE MATCH

Eight publicly available datasets were used to test the validity and adequacy of the MBP method (Table 2.1). Five prostate cancer datasets, Luo (Luo, Duggan et al. 2001), Yu (Yu, Landsittel et al. 2004), Welsh (Welsh, Sapinoso et al. 2001), Dhan (Dhanasekaran, Barrette et al. 2001),and Lap (Lapointe, Li et al. 2004) were downloaded from a public available web site. The malignant prostate cancer and its matched adjacent prostate tissue samples from Yu, Welsh, and Lap datasets, and the malignant prostate cancer and its matched donor samples from Luo and Dhan datasets were used for two sets of pair-wise cross-platform analyses. Three lung cancer data sets, Beer (Beer, Kardia et al. 2002), Bhat (Bhattacharjee, Richards et al. 2001), and Garber (Garber, Troyanskaya et al. 2001), were downloaded from publicly accessible information supporting the published manuscripts. Only the normal and the adenocarcinoma samples were used for analysis. All three datasets were from different platforms or different versions, and pair-wise inter-study analyses were performed.

All pair-wise inter-study analyses relied on matching genes between training data and test data. Entrez ID was used to match Affymetrix datasets using the R package “annotate” (Kuhn, Luthi-Carter et al. 2008), and a web-based match tool, MatchMiner, was used for cDNA datasets

(Bussey, Kane et al. 2003). The genes sharing the same Entrez ID was averaged for their expression.

Table 2.1 Datasets used in the study

Tumor Type	Dataset	Authors	Platform	Sample histology
Lung Cancer	Beer	Beer et al	Affymetrix U95A	86 AC, 10 normal
	Bhat	Bhattacharjee et al	Affymetrix HG 6800	134 AC, 17 normal
	Garber	Garber et al	cDNA	39 AC, 5 normal
Prostate Cancer	Luo	Luo et al	cDNA	16 PC, 9 donors
	Dhan	Dhanasekaran et al	cDNA	14 PC, 19 donors
	Yu	Yu et al	Affymetrix U95A	66 PC, 59 Adjacent
	Welsh	Welsh et al	Affymetrix U95A	25 PC, 9 Adjacent
	Lap	Lapointe et al	cDNA	62 PC, 41 Adjacent

2.2 DATA DESCRIPTION AND PREPROCESSING:

Beer dataset The data was originally published in *Nature Medicine* 2002 by Beer *et al* (Beer, Kardia et al. 2002). The 86 lung adenocarcinoma samples were collected from the University of Michigan Hospital between May 1994 and July 2000 from 67 stage I and 19 stage III patients, and 10 non-neoplastic lung tissues were also obtained during that time. The total 96 samples were analyzed using Affymetrix HG6800 microarray chips. After the data preprocessing, 4467 genes remained in the dataset.

Bhat dataset The data was originally published in *PNAS* 2001 by Bhattacharjee *et al* (Bhattacharjee, Richards et al. 2001). The data used in the project was the subset of the total 203 snap-frozen samples, including 134 lung adenocarcinoma samples and 17 normal lung specimens. The samples were collected from the Thoracic Oncology Tumor Bank at the Brigham

and Women's Hospital/Dana–Farber Cancer Institute, Harvard University. Total 151 samples were analyzed using Affymetrix U95A microarray chips. After the data preprocessing, 4107 genes remained in the dataset.

Garber dataset The data was originally published in *PNAS* 2001 by Garber *et al* (Garber, Troyanskaya et al. 2001). The data included 39 lung adenocarcinoma samples and 5 normal lung specimens. Total 44 samples were analyzed using cDNA microarrays. After the data preprocessing, 3399 genes remained in the dataset.

Luo dataset The data was originally published in *Cancer Research* 2001 by Luo *et al* (Luo, Duggan et al. 2001). The data included 16 prostate adenocarcinoma samples from Johns Hopkins Hospitals during October 1998 and March 2000, and 9 benign prostatic hyperplasia specimens from Johns Hopkins Hospital during February 1999 and November 2000. Total 25 samples were analyzed using cDNA microarrays. After the data preprocessing, 3673 genes remained in the dataset.

Dhan dataset The data was originally published in *Nature* 2001 by Dhanasekaran *et al* (Dhanasekaran, Barrette et al. 2001). The data included 14 prostate adenocarcinoma samples from University of Michigan Prostate SPORE tumor bank and 19 noncancerous. Total 33 samples were analyzed using cDNA microarrays. After the data preprocessing, 7784 genes remained in the dataset.

Lap dataset: The data was originally published in *PNAS* 2004 by Lapointe *et al* (Lapointe, Li et al. 2004). The data included 62 primary prostate cancer samples and 41 matched normal prostate tissues from Stanford University, Karolinska Institute, and Johns Hopkins University. Total 33 samples were analyzed using cDNA microarrays. After the data preprocessing, 1735 genes remained in the dataset.

Welsh dataset The data was originally published in *Cancer Research* 2001 by Welsh *et al* (Welsh, Sapinoso et al. 2001). The data included 25 primary prostate cancer samples and 9 matched normal prostate tissues from University of Virginia. Total 33 samples were analyzed using Affymetrix U95A microarray chips. After the data preprocessing, 9494 genes remained in the dataset.

Yu dataset The data was originally published in *JCO* 2004 by Yu *et al*(Yu, Landsittel et al. 2004). The data included 66 primary prostate cancer samples and 59 matched normal prostate tissues. Total 125 samples were analyzed using Affymetrix U95A microarray chips. After the data preprocessing, 9109 genes remained in the dataset.

All the data were preprocessed using standard data filtering by eliminating genes with low expressions and genes not varying sufficiently across the training samples. The data were standardized by first column-wise and then row-wise normalization by subtracting column or row means and dividing by the corresponding column or row standard deviations.

2.3 NOTATIONS AND GENERAL CONCEPT

Suppose a prediction model is to be constructed from a training study and will be applied to a test study. Let G^{tr} be the gene set covered in the training study and G^{te} in the test study, where normally $G^{tr} \neq G^{te}$ if the training and test studies are of different microarray platforms. Denote by $D^{tr} = \left\{ x_{gs}^{tr}, g \in G^{tr}, s \in S^{tr} \right\}$ the expression intensity matrix of the training study and $D^{te} = \left\{ x_{gs}^{te}, g \in G^{te}, s \in S^{te} \right\}$, where g represents gene indexes and s are samples. For traditional GBP methods, gene sets covered in the training and test studies have to be identical so that the

prediction model can be applied across studies. The prediction model is usually constructed in the submatrix $\widetilde{D}^{tr} = \left\{ x_{gs}^{tr}, g \in G^{tr} \cap G^{te}, s \in S^{tr} \right\}$ and applied to $\widetilde{D}^{te} = \left\{ x_{gs}^{te}, g \in G^{tr} \cap G^{te}, s \in S^{te} \right\}$. As a result, the prediction model is not totally independent of the test study information. In the MBP approach, the gene set G^{tr} in the training study is clustered into *K* clusters by *K*-means such that $G^{tr} = \bigcup_{k=1}^{K} G_k^{tr}$ and $G_i^{tr} \cap G_j^{tr} = \phi$ for $1 \le i < j \le K$. The prediction model will be constructed based on the *K* cluster modules, of the form $F(r_1^{tr}, r_2^{tr}, ..., r_K^{tr})$, where r_k^{tr} is the representative gene vector in cluster module G_k^{tr}. In this study, the "median gene" that has the smallest sum of distances to other genes in the cluster was used as the representative gene vector. Namely when $\widetilde{g}_k^{tr} = \arg\min_{m \in G_k^{tr}} \sum_{g \in G_k^{tr}} \left\| x_{m\bullet}^{tr} - x_{g\bullet}^{tr} \right\|$ and $\left\| x_{m\bullet}^{tr} - x_{g\bullet}^{tr} \right\|^2 = \sum_{s \in S^{tr}} (x_{ms}^{tr} - x_{gs}^{tr})^2$, $r_k^{tr} = (x_{\widetilde{g}_k^{tr} s}^{tr}, s \in S^{tr})$. For application to the test study, denote by $G_k^{te} = G_k^{tr} \cap G^{te}$ the genes in the *k*-th cluster module that appear in the test study. If $G_k^{te} \neq \phi \quad \forall k$, the representative gene vectors can be similarly calculated as $r_k^{te} = (x_{\widetilde{g}_k^{te} s}^{te}, s \in S^{te})$ where $\widetilde{g}_k^{te} = \arg\min_{m \in G_k^{te}} \sum_{g \in G_k^{te}} \left\| x_{m\bullet}^{te} - x_{g\bullet}^{te} \right\|$ and $\left\| x_{m\bullet}^{te} - x_{g\bullet}^{te} \right\|^2 = \sum_{s \in S^{te}} (x_{ms}^{te} - x_{gs}^{te})^2$. The proposed MBP model can then be applied to the test study by $F(r_1^{te}, r_2^{te}, ..., r_K^{te})$.

2.4 ALGORITHM DEVELOPMENT

The MBP algorithm was developed under the rationale that genes sharing similar expression profiles could be grouped together and that a representative gene can be selected from the group

of genes. The algorithm involved four major steps: gene clustering, cluster merging, cluster selection, and model construction. The MBP schema is shown in Figure 2.1.

Gene clustering The processed data were clustered into K clusters by the classical K-means method (Hartigan and Wong 1979) or PW-K-means method (Tseng 2007). The clusters were defined as gene modules. Normally $K = 100$ or 150 was chosen.

Module merging When the number of genes within a module was less than a given threshold δ, the small module was merged into its nearest-neighboring module based on the minimum distance between module centroids. The selection of δ was determined by a probabilistic model described below to avoid missing genes of the entire module in the test study with high probability.

Module selection Although in the MBP approach, the number of features has been reduced to hundreds, the dimensionality is still high and proper feature (module) selection is needed to achieve better performance. Thus, both univariate and multivariate feature selection methods were explored. For univariate feature selection, the top k modules are selected according to their ranks of average absolute value of moderated t statistics (Tibshirani, Hastie et al. 2002), and the representative genes within each selected module were used to construct the prediction model. For multivariate feature selection methods, the module selection was embedded with the prediction model construction.

Model construction The selected representative genes of the prediction module were used to fit the prediction models. For univariate feature selection methods, three classical classification methods included linear discrimination analysis (LDA) (Mardia, Kent et al. 1979), , k-nearest neighbor (KNN) (Dasarathy 1991), and supporting vector machine (SVM) (Cristianini and

Shawe-Taylor 2000) were examined. For multivariate feature selection, prediction analysis of microarray (PAM) (Tibshirani, Hastie et al. 2002) was explored.

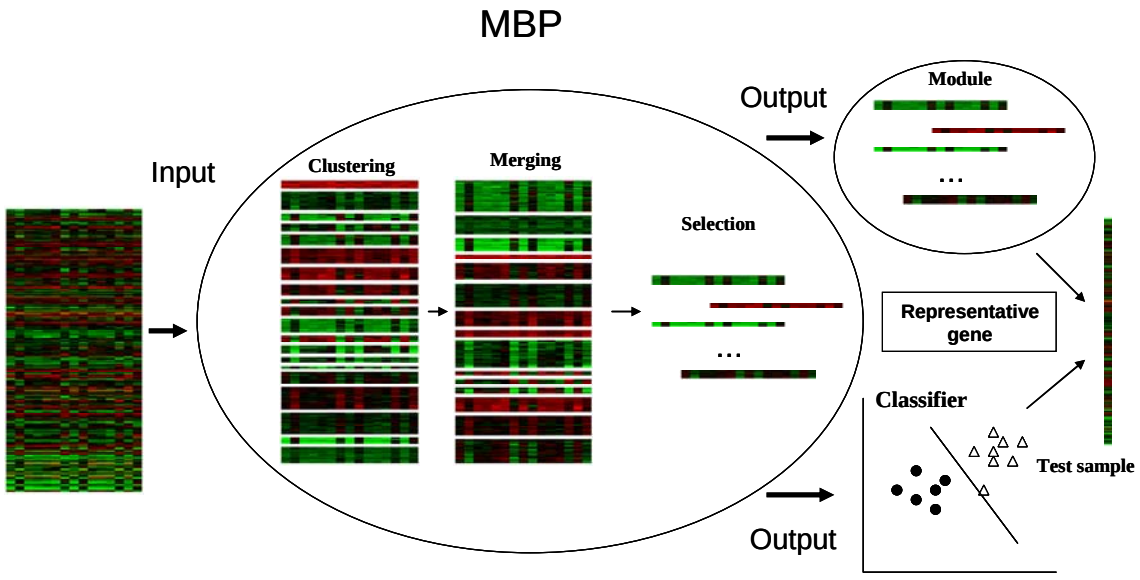

Figure 2.1 Schema of the module based prediction (MBP) method: a microarray data with sizable samples is selected as training input for the MBP; then the genes are clustered into groups using unsupervised method based on the gene similarity, such as *K*-means correlation method; for those gene cluster sizes smaller than a threshold δ, the clusters are merged into their nearest neighbor; the clusters after merging are selected by moderate t statistics to form prediction modules which comprise of the prediction model as the output; for each module, a representative gene is selected based on the minimum sum of distance among the genes within the module; then the representative genes are used to predict outcome for a new test sample.

The algorithm was implemented using R 2.7.1 (free software downloaded from www.r-project.org). The MBP core algorithm was written using R program, however, some functions, such as *K*-means clustering, classifiers (LDA, KNN, SVM, PAM) were downloaded from R packages. The PW-*K*-means clustering program C code was provided by Dr. Tseng.

2.5 MINIMUM CLUSTER SIZE IN MODULE MERGING AND DISTRIBUTION DIAGNOSIS

One of the motivations to develop the MBP method is to build a prediction model solely on the training data, independent of test data and portable across studies with different microarray platforms. A necessary condition for the MBP procedure to succeed is that the test study should contain one or more genes in each gene cluster module in order to calculate the representative gene vector in the prediction model. Below a simplified probabilistic model to estimate the smallest δ needed to achieve the goal is provided. Assume π is the probability for a gene in the training study to be missing in the test study and that gene missing is independent of each other. The probability that the MBP method obtains no less than N genes in each of the *K* modules in the test study is

$$\begin{aligned}\widetilde{p}(\pi,\widetilde{G}_K) &= \Pr\left(\text{no module has less than N genes in test study}\middle|\widetilde{G}_K\right)\\ &= \prod_{k=1}^{K}\left(1-\sum_{n=0}^{N-1}\binom{n(G_k^{tr})}{n}\cdot\pi^{n(G_k^{tr})-n}\cdot(1-\pi)^n\right) \qquad (1)\end{aligned}$$

where $\widetilde{G}_K=\left(n(G_1^{tr}),\cdots,n(G_K^{tr})\right)$ and $n(G_k^{tr})$ is the number of genes in cluster G_k^{tr}. In this paper, we require $\widetilde{p}>99\%$ and $N = 3$. The probability calculation depends only on the gene missing probability π and the module sizes, $n(G_k^{tr})$. In the data analyses of the eight data sets used in this paper, we found that the cluster sizes generated by *K*-means clustering follow multinomial distributions very well, i.e. $\widetilde{G}_K=\left(n(G_1^{tr}),\cdots,n(G_K^{tr})\right)\sim Multinomial\left(n(G^{tr}),1/K,\cdots,1/K\right)$. (see QQ-plots in Figure 3.1). Thus the probability of each module in test study to have no less than *N* genes becomes

$$p(n(G^{tr}), K, \pi) = P(\text{no module has less than N genes in test study})$$
$$= \int P(\text{no module has less than N genes in test study}) \cdot p(\widetilde{G}_k) d\widetilde{G}_k \qquad (2)$$
$$= \sum_{\widetilde{G}_k} \widetilde{p}(\pi, \widetilde{G}_K) \cdot p(\widetilde{G}_k)$$

To estimate the minimal δ required in the module merge procedure, the following simulation to calculate the probability was performed such that each module in the test study has no less than *N* genes after *δ-merge*:

1. Suppose $n(G^{tr})$, *K*, π, δ are given. Simulate $\widetilde{G}_K$ from $Multinomial\left(n(G^{tr}), 1/K, \cdots, 1/K\right)$.

2. Given δ, merge clusters with less than δ genes into a random cluster. Suppose the resulting cluster sizes become $\widetilde{G}'_{K'}$ ($K' \leq K$).

3. Compute the conditional probability, $\widetilde{p}(\pi, \widetilde{G}'_{K'})$ from equation (1).

4. Repeat 1-3 for B times (B=1,000 in this study). The probability of successful application to the test study can be estimated by

$$p(n(G^{tr}), K, \pi, \delta) = \text{P(no module has less than N genes in test study} \mid \delta \text{ - merge)}$$
$$= \frac{1}{\mathrm{B}} \sum_{b=1}^{B} \widetilde{p}^{(b)}(\pi, \widetilde{G}'_{K'}) \qquad (3)$$

5. Find the smallest δ such that $p(n(G^{tr}), K, \pi, \delta) \geq 0.99$.

The advantage of the probabilistic model is that the estimation of minimal δ only depends on the total number of genes in the training data ($n(G^{tr})$), the number of clusters *K* used in *K*-means and the probability of gene missing in the test study. It does not depend on the observed data and a table can be computed for a rapid decision in future applications. For example, when 4,000 genes are analyzed in the training study, Kmeans clustering generated *K* = 100 modules and one expects π = 50% gene missing probability in the test study, $\delta \geq 3$ was required in *δ-merge* to

guarantee successful application of the MBP in inter-study prediction with 99% probability (see Table 3.1 in the result section).

2.6 CLASSIFICATION METHODS

In this study, binary outcome prediction or classification was the main focus; however, the method can be extended to other outcome predictions, such as those for time to event data. To evaluate the MBP performance, four commonly used classifiers, LDA, KNN, PAM and SVM were chosen for univairate or multivariate selections. For univariate selection, the features were chosen by selecting top k clusters for the MBP or top q individual genes for the GBP based on moderate t statistics; whereas for multivariate selection, the features were chosen by considering correlations among genes and minimizing error rate using the two state-of-the-art methods, PAM and Recursive-SVM (Zhang, Lu et al. 2006).

Linear discriminate Analysis (LDA) Originally conceptualized by RA Fisher and developed by others (Fisher 1936; Mardia, Kent et al. 1979), LDA is a very popular statistical classification method, which identifies linear combinations of features that accurately separate two or more classes of events. The method also is used for data dimension reduction. In this study, LDA was used for classification purposes.

K-nearest neighbor (KNN) KNN is a simple machine learning method to classify the events based on the majority vote of its neighbors (Dasarathy 1991). It is very useful for features in multidimensional space. KNN was used for classification purposes here.

Support vector machine (SVM) SVM is a supervised machine learning method that identify the classes by separating hyperplanes, which maximizes the margin between two classes

(Cristianini and Shawe-Taylor 2000). Generally, SVM is used to classify two groups of data, but it can be extended to separate more than two groups of data (Hsu and Lin 2002). In this study, binary SVM classification method was used to compare results among different classifiers.

Nearest centroid This method was used as a special case when no shrinkage operation involved for feature selection in PAM (Tibshirani, Hastie et al. 2002) and marked as PAM-U or PAM for PAM univariate feature selection.

2.7 FEATURE SELECTION METHODS

Two feature selection approaches were used to select prediction modules for the MBP, i.e. univairate and multivariate feature selections. For univariate approach, once the genes in training set are grouped in K clusters, the top k clusters are selected based on the absolute value of moderate t statistics to form the prediction modules; whereas for multivariate approach, the prediction modules are selected by the following two feature methods from K clusters.

Prediction analysis for microarrays (PAM) Published in 2002 (Tibshirani, Hastie et al. 2002), PAM has become a very popular method for gene classification and prediction. The unique part of the PAM is its centroid shrunken algorithm (Tibshirani, Hastie et al. 2002). Unlike LDA, KNN, and SVM, PAM was used for both feature selection and classification in this study.

Recursive supporting vector machine (R-SVM) R-SVM was developed for genes classification of noisy data. Similar method was reported in *machine learning* 2002 by Guyon et al (Guyon, Weston et al. 2002) using SVM recursive feature elimination or SVM-RFE to select gene features robust to outliers. However, according to Zhang et al, R-SVM was better than

SVM-RFE in regarding to robustness to noise (Zhang, Lu et al. 2006). R-SVM was used as a multivariate selection method in the study.

2.8 EVALUATION AND SIMULATION

Prediction accuracy versus prediction performance index In common practice, prediction performance is evaluated by overall accuracy, calculated as the correctly predicted number of subjects divided by total number of subjects tested. Sometimes, the accuracy may misinterpret the prediction performance when samples are imbalanced between the two groups of subjects, e.g. if a hundred of subjects are tested, ninety five of them are cancer patients and five of them are normal. For a useless predictor, all subjects are classified into the cancer group and the prediction accuracy is still 95%. To avoid this problem, besides prediction accuracy, a prediction performance index (PPI), computed as the average of sensitivity and specificity, was used to evaluate the prediction performance. As for previous example, PPI is 50%.

The MBP versus the GBP within study prediction Prediction performance was assessed for every dataset using a leave-one-out cross validation (LOOCV) approach (Kohavi 1995). Since there were random factors during *K*-means in the MBP method, the LOOCV was run thirty or hundred times and used LDA, KNN, and SVM as the classifiers. The means and standard deviations of accuracies were calculated and compared with the accuracies obtained from the traditional GBP method based on moderate *t* statistics feature selection and PAM method.

The MBP versus the GBP inter-study prediction Cross-platform prediction was performed by the standard MBP algorithm stated above. The test data used in the pair wise inter-study analyses were three lung cancer datasets, three prostate cancer datasets matched with adjacent

tissues as controls, and two prostate cancer datasets using donors' samples as controls. The prediction accuracies and PPI were calculated based on three classifiers, LDA, KNN, and SVM. The results were compared with those of the GBP method, which only used genes common to both training and test datasets. PAM classification method with univariate and multivariate feature selections were also evaluated.

Simulation with varying gene variability To determine whether using gene cluster information would create a model that was robust even when gene noise was present, white noise was randomly added to the Luo dataset. The noise followed a Gaussian distribution with mean $\mu = 0$ and standard deviation σ. The magnitude of noise was determined by size of σ, and the range of noise added was based various different proportions of noise (p). The prediction accuracies were evaluated by the LOOCV approach across different σs and different ps and were compared with their counterparts from the GBP methods.

Simulation with gene mismatches in cross-platform scenario The robustness to missing genes was evaluated by randomly splitting the Luo dataset into a training dataset and a test dataset by a 1:1 ratio, and serial proportions of genes were randomly deleted from the training and test dataset to create missing genes. The prediction accuracies and prediction successful rate (PSR), defined as number of successful predictions divided by the total number of prediction tests attempted when genes used for prediction are missing in test set, were compared between the MBP method and the GBP method using LDA, KNN, and SVM classifiers.

3.0 RESULTS

3.1 DISTRIBUTION OF CLUSTER SIZE

In order to estimate minimum cluster size, δ, one needs to know the distribution of the cluster size generated by *K*-means' method. Since *K*-means' clusters were randomly generated, the cluster size, G_k^{tr} was considered as a random variable, which might affect the prediction or classification accuracy and stability for approaches involving clustering as the data reduction method. G_k^{tr} usually depends on, G^{tr} K, and individual dataset. To have a good estimate of δ, it is essential for one to estimate the distribution of G_k^{tr}. Both conditional Poisson and multinomial distributions were tested to fit the distribution of G_k^{tr}. The distribution was estimated using the eight datasets under either Poisson or multinomial conditions by fitting QQ-plots. The results showed that the Poisson fits were not good (data not shown), but the eight datasets fitted well under multinomial assumptions (Figure 3.1). Determination of G_k^{tr} distribution provided useful information for estimation of δ.

3.2 ESTIMATION OF MINIMUM CLUSTER SIZE

As a key parameter of the MBP model, δ was used to control or minimize prediction failure caused by missing genes when genes in a model built on training set do not exist in a test set. When a cluster size G_k^{tr} was smaller than δ, the cluster was merged into its nearest cluster to minimize the probability of prediction failure. A smaller G_k^{tr} leads to a higher probability of prediction failure. It is crucial to estimate δ given the proportion of genes missing in test set, π, and the probability of successful prediction, α (considered acceptable when α = 99%).

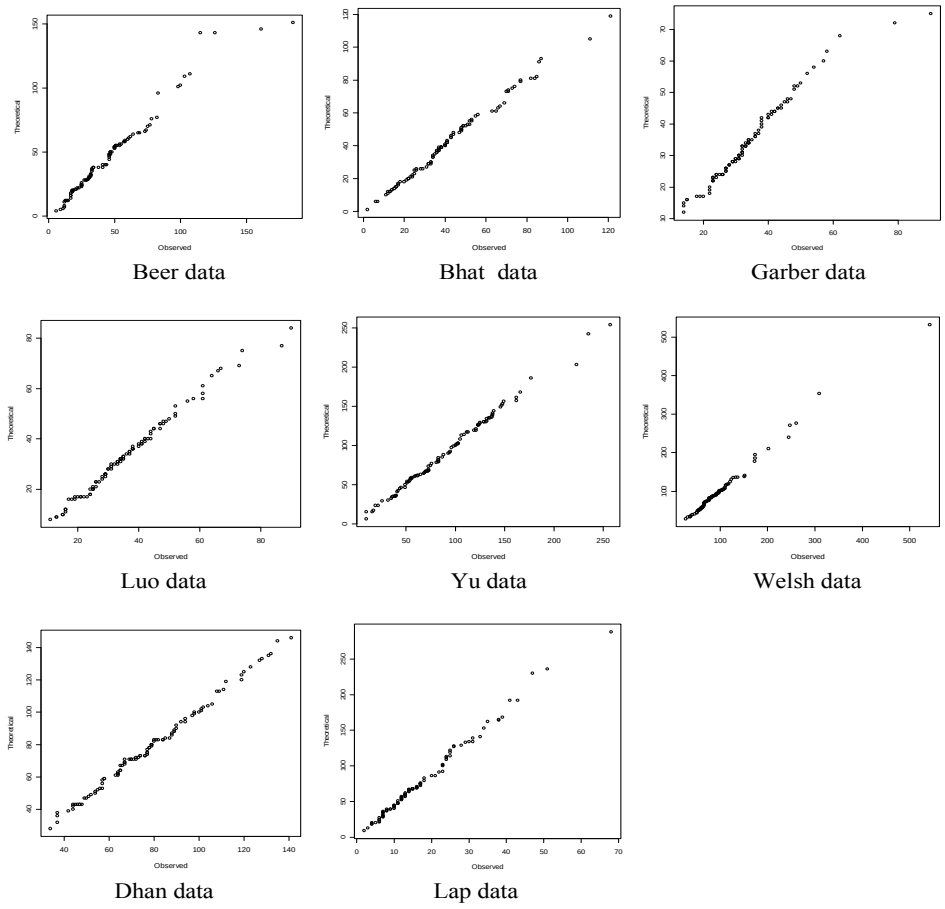

Figure 3.1 QQ-plots of observed cluster size versus theoretical cluster size generated according to multinomial distributions: X axis represents cluster size generated by *K*-means clustering method and Y axis represents cluster size simulated by multinomial distribution.

The simulated results were listed in Table 3.1 (B =1000). The results showed that the threshold δ were higher when K , π , and α increased, whereas G^{tr} decreased. Also cluster sizes after

Table 3.1 Simulated results for estimation of δ and K' (B=1000, N=3)

$n(G^{tr})$	K	π					
		0.2	0.3	0.4	0.5	0.6	0.7
2000	100	1 (100)	1 (100)	1 (100)	15 (91)	20 (63)	25 (43)
2000	125	1 (125)	8 (124)	12 (111)	16 (78)	19 (58)	25 (42)
2000	150	7 (147)	10 (130)	13 (98)	16 (71)	19 (59)	26 (44)
2000	175	8 (156)	10 (130)	13 (91)	15 (76)	19 (61)	26 (45)
2000	200	8 (161)	10 (127)	12 (99)	15 (76)	19 (61)	26 (43)
3000	100	1 (100)	1 (100)	1 (100)	1 (100)	12 (100)	29 (67)
3000	125	1 (125)	1 (125)	1 (125)	12 (125)	21 (98)	27 (61)
3000	150	1 (150)	1 (150)	9 (150)	16 (129)	21 (85)	26 (61)
3000	175	1 (175)	7 (175)	12 (162)	17 (111)	20 (85)	26 (65)
3000	200	6 (200)	10 (187)	13 (154)	16 (114)	20 (82)	27 (64)
4000	100	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	21 (100)
4000	125	1 (125)	1 (125)	1 (125)	1 (125)	1 (125)	30 (89)
4000	150	1 (150)	1 (150)	1 (150)	1 (150)	22 (128)	29 (78)
4000	175	1 (175)	1 (175)	1 (175)	15 (170)	22 (117)	28 (76)
4000	200	1 (200)	1 (200)	11 (198)	17 (161)	22 (105)	27 (79)
5000	100	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)	1 (100)
5000	125	1 (125)	1 (125)	1 (125)	1 (125)	1 (125)	27 (124)
5000	150	1 (150)	1 (150)	1 (150)	1 (150)	1 (150)	31 (109)
5000	175	1 (175)	1 (175)	1 (175)	1 (175)	21 (165)	30 (98)
5000	200	1 (200)	1 (200)	1 (200)	13 (200)	22 (157)	29 (94)

merging, K' were estimated shown in parentheses in Table 3.1. This table provided a reference when the threshold was set up for cluster merging. Meanwhile, based on the simulation results, as shown in Table 3.2, δ for each pair of training and test datasets was estimated, given K = 100, α = 99%, and as indicated in parentheses, π was calculated based on the number of matched intersection genes in both training and test data sets divided by the number of genes in training data sets, and δ was estimated based on the multinomial distribution of cluster size (Table 3.2).

The results showed that gene mismatching between the training datasets and test datasets varied from 36% to 97%. The following evaluations of the MBP and the GBP performances were based on the data mismatching scenarios.

Table 3.2 Common gene cross platform data with (π, δ)

Training	Test							
	Beer	**Bhat**	**Garber**	**Luo**	**Dhan**	**Welsh**	**Yu**	**Lap**
Beer	**4467 (0.00, 1)**	**2493 (0.44,1)**	**1594 (0.64,1)**	-	-	-	-	-
Bhat	**2493 (0.39, 1)**	**4107 (0.00, 1)**	**1493 (0.64, 1)**	-	-	-	-	-
Garber	**1594 (0.53, 1)**	**1493 (0.56, 1)**	**3399 (0.00, 1)**	-	-	-	-	-
Luo	-	-	-	**3673 (0.00, 1)**	**2352 (0.36, 1)**	-	-	-
Dhan	-	-	-	**2352 (0.70, 1)**	**7784 (0.00,1)**	-	-	-
Welsh	-	-	-	-	-	**9494 (0.00,1)**	**2521 (0.73, 1)**	**356 (0.96, 100)**
Yu	-	-	-	-	-	**2521 (0.72, 1)**	**9109 (0.00, 1)**	**295 (0.97, 100)**
Lap	-	-	-	-	-	**356 (0.79, 43)**	**295 (0.83, 34)**	**1735 (0.00, 1)**

3.3 PREDICTION ACCURACIES WITHIN STUDY

The purpose of the analysis is to test the assumption of no significant loss of predictive power regarding to the prediction accuracy using the MBP method within a study. We may or may not

expect an increased prediction accuracy of the MBP. Given $K = 100$, k =10, 20, 30 and δ = 10 for MBP, and selected top differentiated genes of 10, 20, and 30 for the GBP, the prediction accuracies of the MBP versus those of the GBP using three classifiers, LDA, KNN, and SVM across eight cancer datasets were computed using LOOCV, as shown in Figure 3.2, given the condition that if there was at least one gene matched in the test set, i.e. $N \geq 1$. The results showed that the prediction accuracies were classifier, number of selected features, and data dependent; however, there was no evidence showing that either the GBP or the MBP

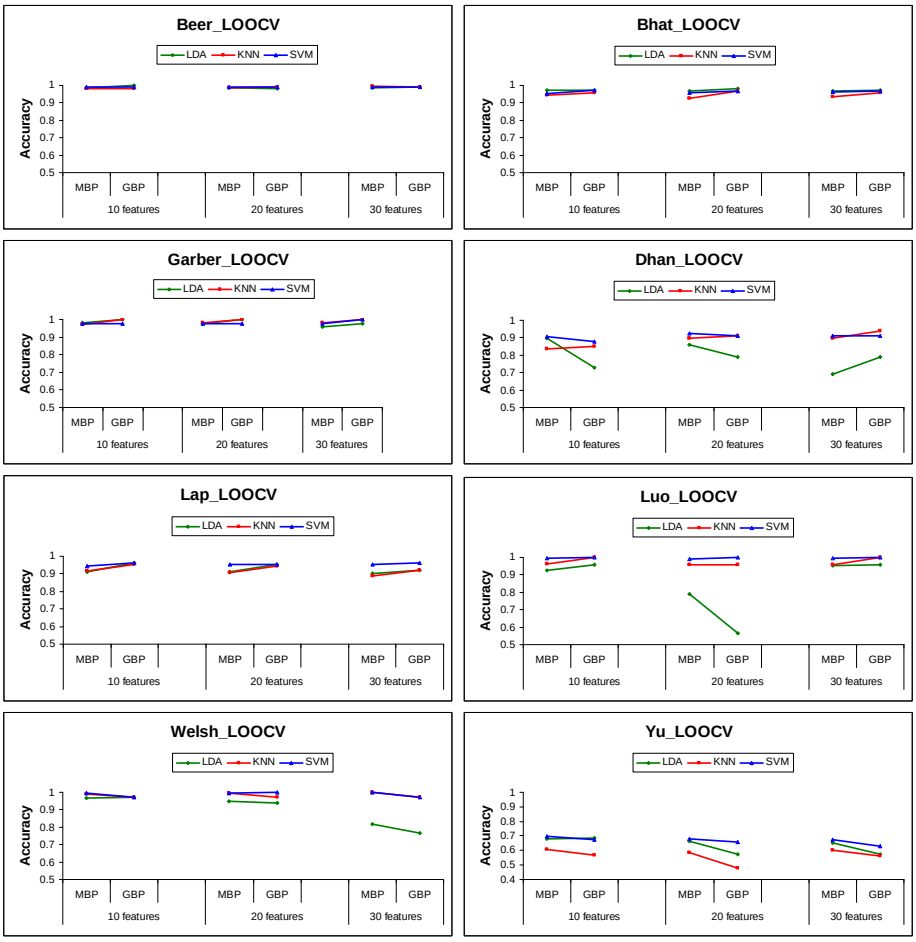

Figure 3.2 Within study prediction accuracies between the MBP and the GBP across eight cancer datasets: The MBP prediction accuracies were compared with those of the GBP within eight datasets using LOOCV approach. The parameters were set for MBP as K=100, k =10, 20, and 30, $N \geq 1$, and δ=10 and the parameters was set for the GBP as q =10, 20, and 30.

method yielded better prediction accuracy one way or another despite fluctuations of the prediction accuracies across different datasets and different classifiers, indicating that the performance of these two methods was indistinguishable. Further, we changed parameters from $N \geq 1$ to $N \geq 3$ and $\delta = 10$ to $\delta = 20$, hoping it would yield more stable results. As shown in Figure 3.3, the results were similar as in Figure 3.2, but with smaller standard deviations after 30 repeated tests. The results indicated that no significant loss of predictive power using the MBP within a study across the eight cancer datasets.

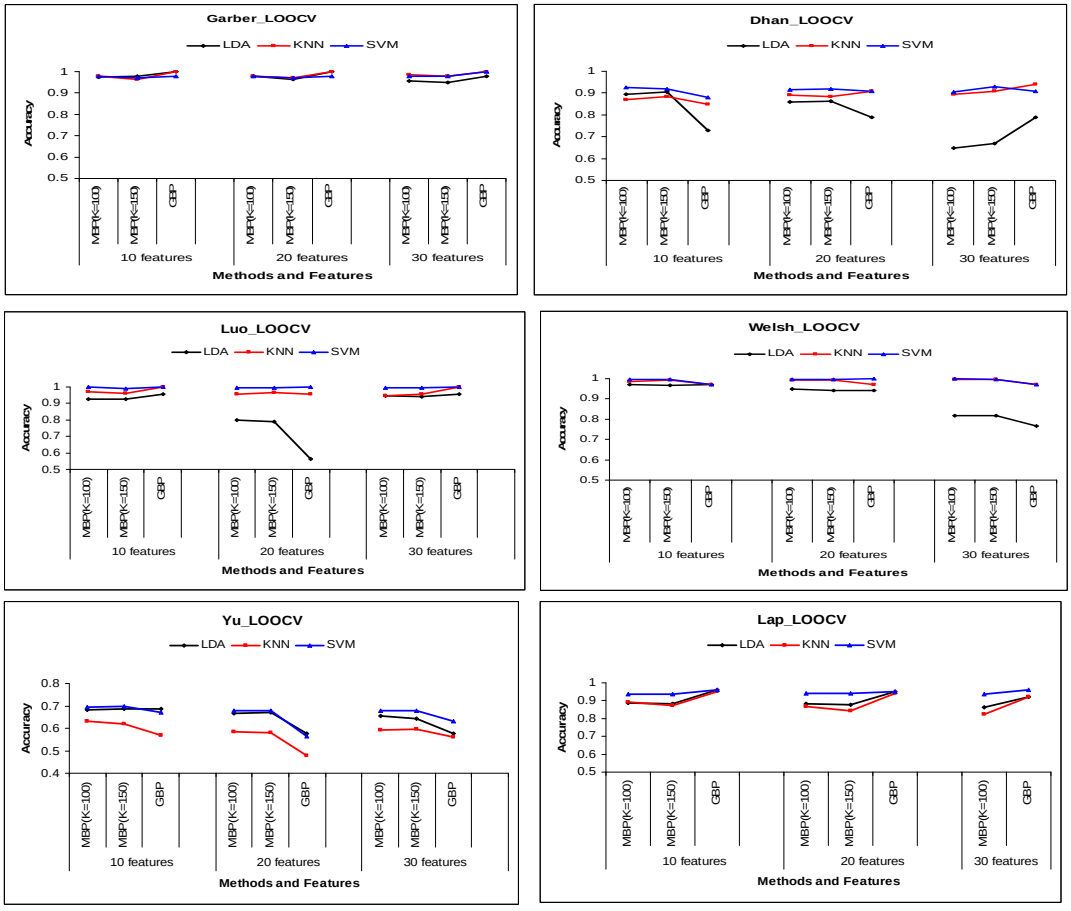

Figure 3.3 Within study prediction accuracies between the MBP and the GBP across six cancer datasets: The MBP prediction accuracies were compared with those of the GBP within six datasets using LOOCV approach. The parameters were set for the MBP as K=100 and 150, k =10, 20, and 30, $N \geq 3$, and δ=10 and the parameters was set for the GBP as q = 10, 20, and 30.

3.4 PREDICTION ACCURACIES INTER-STUDIES

Since prediction accuracies were indistinguishable between the two methods in within studies, the MBP performance was then tested on datasets in inter-studies and compared with the GBP regarding to prediction accuracy and PPI. There were three sets of inter-study evaluations: one set of lung cancer data, including Beer, Bhat, and Garber, and two sets of prostate cancer data, i.e. one set with donors as control including Luo and Dhan and the other set with adjacent tissues as control, including Lap, Welsh, and Yu. The three sets of pair-wise platform comparisons were shown in Table 3.2. Each dataset within the three sets served as training data in turns and tested on the other dataset. The number of genes in each training dataset and the number of common genes to each testing pair dataset were listed in Table 3.2. The gene missing rate varied from 36% to 97% (Table 3.2). Given $K = 100$, k =10, 20, 30 and δ =10 for the MBP and selected top differentiated genes of 10, 20, and 30 for the GBP, the prediction accuracies of the MBP versus those of the GBP using three classifiers, LDA, KNN, and SVM across six pair-wise inter-study cancer datasets were computed, as shown in Figure 3.4. Given the condition of $N \geq 1$, the results showed that the prediction accuracies were classifier, number of selected features, and data dependent; however, there was no evidence showing that neither the GBP nor the MBP method performed better than the other one despite fluctuations of the prediction accuracies across different datasets and different classifiers, indicating that the performance of these two methods was indistinguishable. Note that, the GBP inter-study predictions were based on intersection genes or common genes of pair-wise inter-study datasets, otherwise, the prediction failed due to gene mismatching in the test sets, indicating that the prediction model from the training set is not

entirely independent from the test set and varies greatly when the test data set changes. The MBP method, however, does not have this issue.

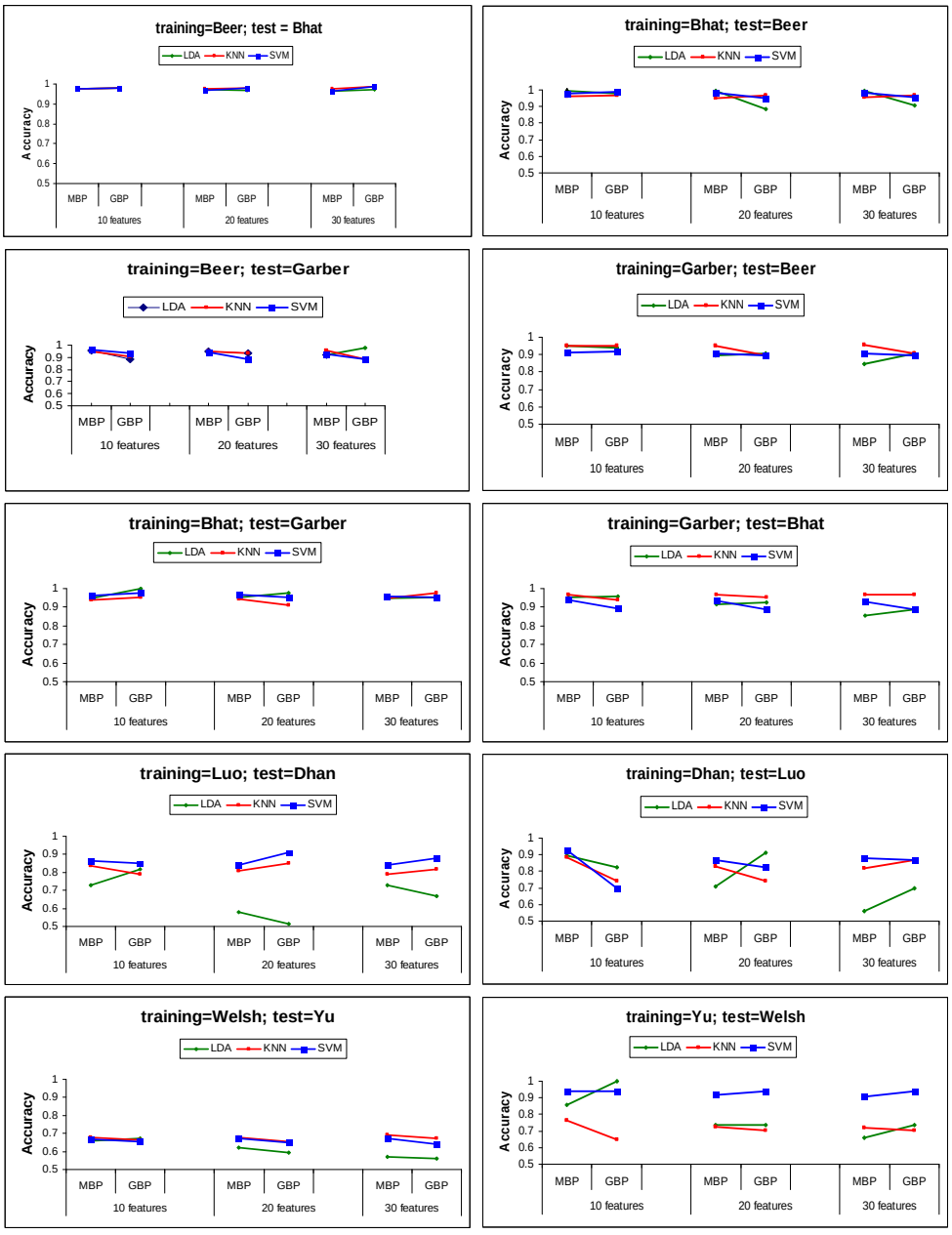

Figure 3.4 Pair-wise inter-study prediction accuracies between the MBP and the GBP: The MBP prediction accuracies were compared with those of the GBP among ten pair-wise inter-study datasets. The parameters were set for the MBP as K=100, k =10, 20, and 30, $N \geq 1$, and δ=10 and the parameters was set for the GBP as q = 10, 20, and 30. Three classifiers were used for the comparison, LDA, KNN, and SVM.

Similarly as within study evaluation, the parameters were changed from $N \geq 1$ to $N \geq 3$ and $\delta =$ 10 to $\delta = 20$ to evaluate the MBP performance, compared with the GBP using intersection genes. As shown in Figure 3.5 and Figure 3.6, the results were similar as in Figure 3.4, but with smaller standard deviations after 100 repeated tests.

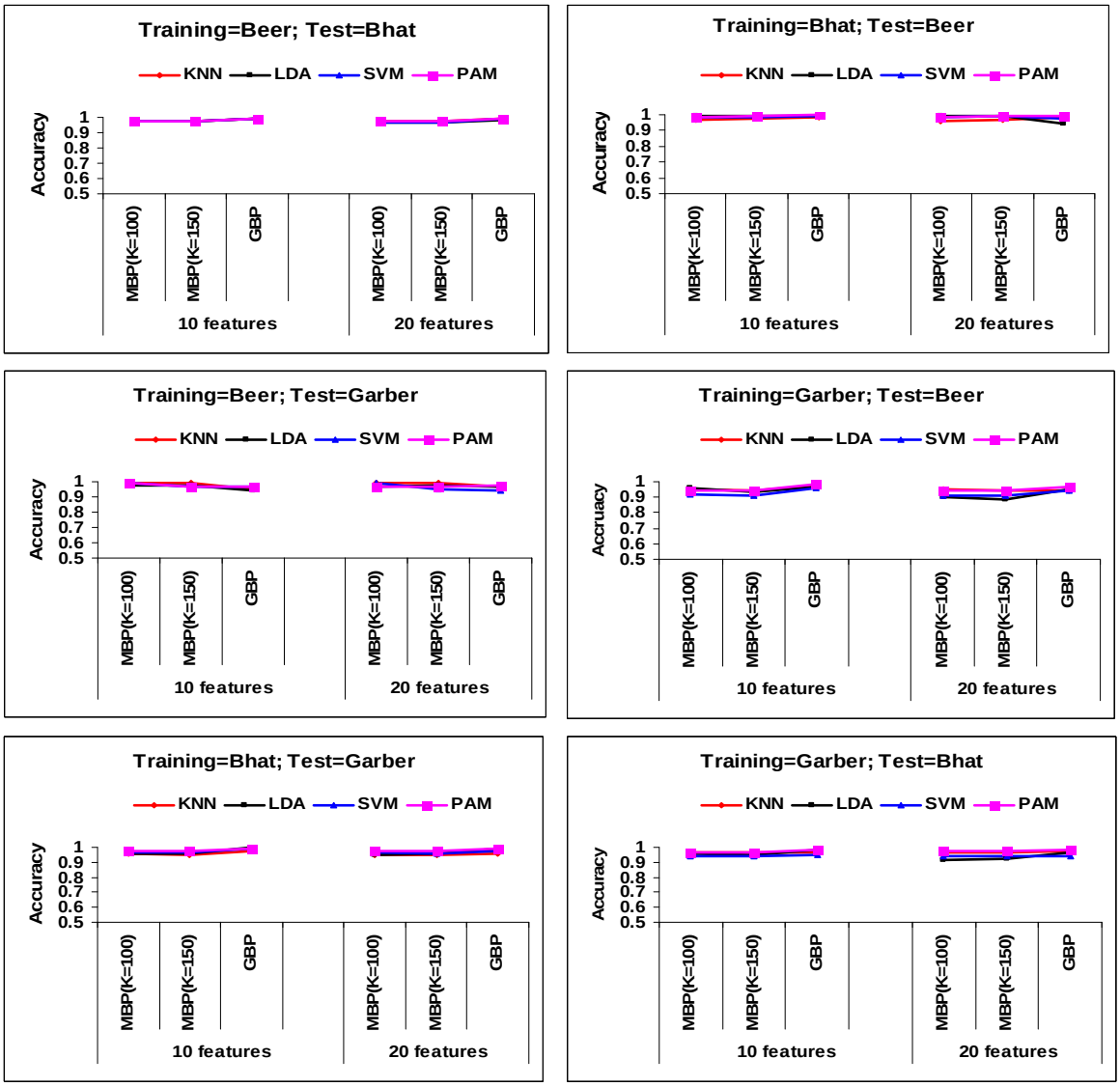

Figure 3.5 Pair-wise inter-study prediction accuracies between the MBP and the GBP: The MBP prediction accuracies were compared with those of the GBP among six pair-wise lung cancer inter-study datasets. The parameters were set for the MBP as K=100 and 150, k =10 and 20, $N \geq 3$, and δ=20 and the parameters was set for the GBP as q =10 and 20. Four classifiers were used for the comparison, LDA, KNN, SVM and PAM.

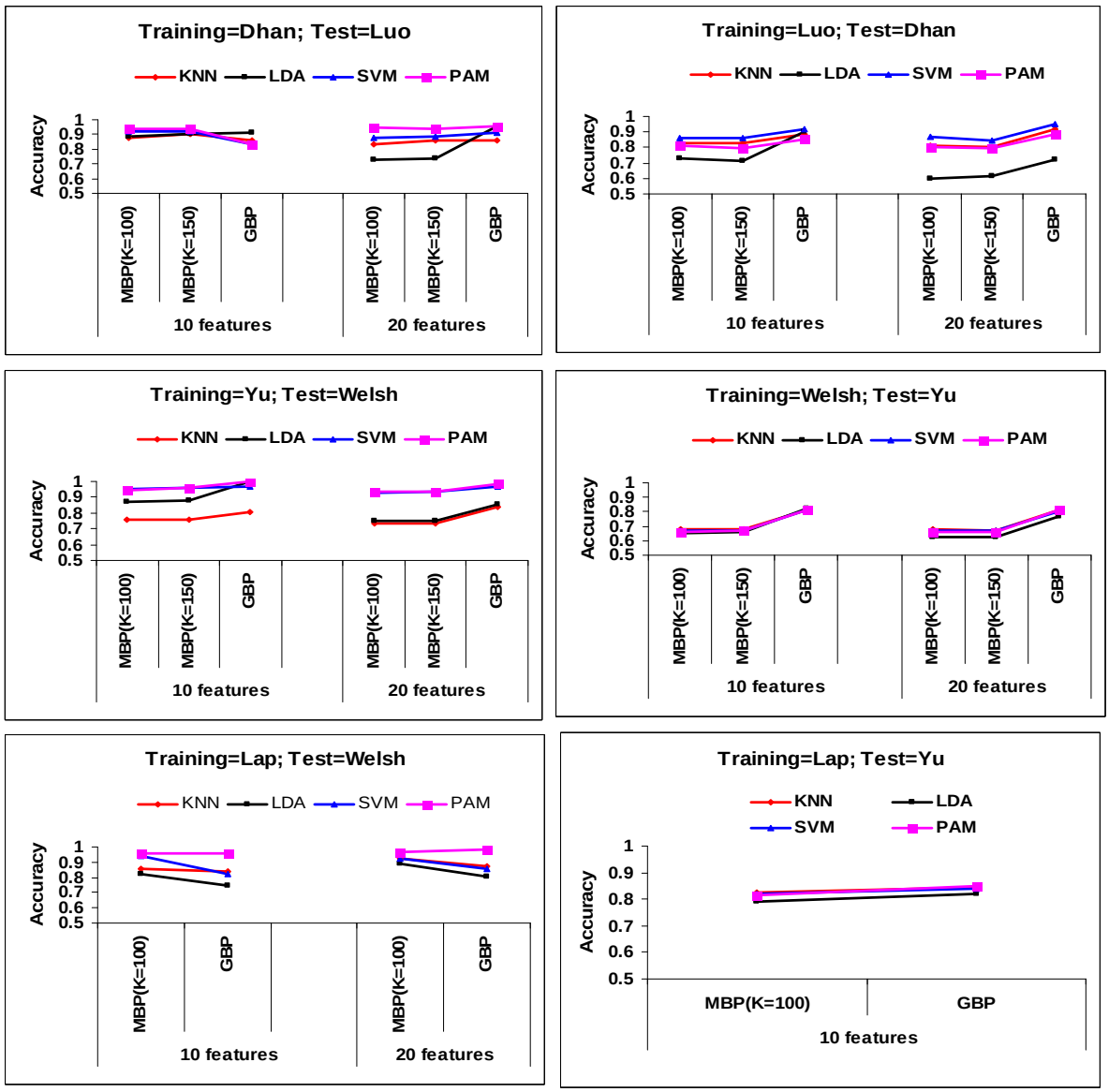

Figure 3.6 Pair-wise inter-study prediction accuracies between the MBP and the GBP: The MBP prediction accuracies were compared with those of the GBP among six pair-wise prostate cancer inter-study datasets. The parameters were set for the MBP as K=100 and 150, k =10 and 20, $N \geq 3$, and δ=20 and the parameters was set for the GBP as q =10 and 20. Four classifiers were used for the comparison, LDA, KNN, SVM and PAM.

Noticed that three lung cancer datasets and two prostate cancer datasets, Luo and Welsh, used in the analyses were unbalanced between the tumor and control groups, we used PPI to evaluate the

MBP performance in addition to prediction accuracy shown in Figures 3.4, 3.5, and 3.6, to avoid inflated prediction accuracy. As shown in Figure 3.7, PPI was generally smaller than the

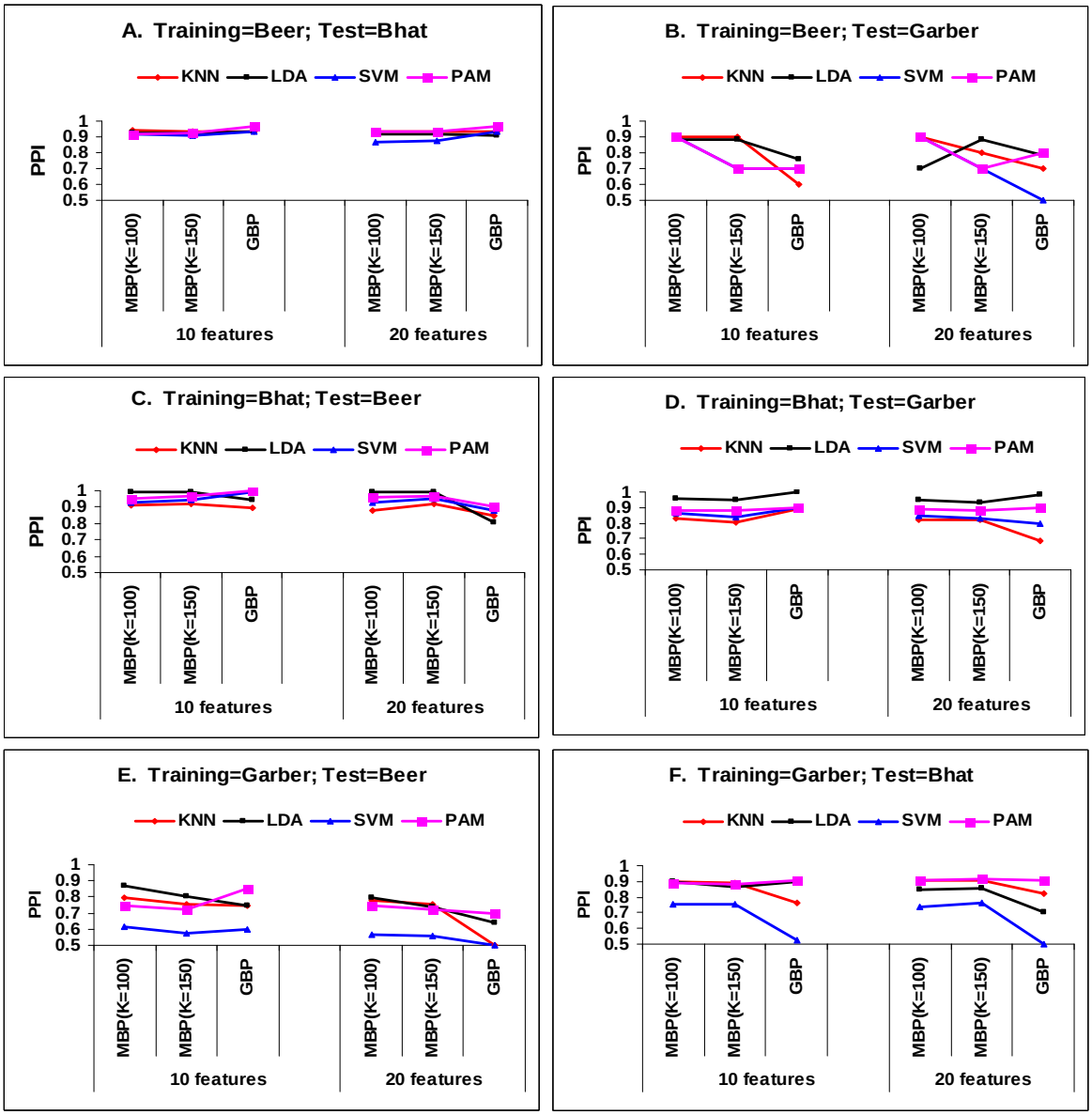

Figure 3.7 Pair-wise inter-study PPI between the MBP and the GBP: The MBP PPI was compared with that of the GBP among six pair-wise lung cancer inter-study datasets. The parameters were set for the MBP as K=100 and 150, k =10 and 20, $N \geq 3$, and δ=20 and the parameters was set for the GBP as q =10 and 20. Four classifiers were used for the comparison, LDA, KNN, SVM and PAM.

prediction accuracy shown in Figure 3.5 and showed greater variations among four classifiers across six pair-wise lung cancer datasets, especially, when Garber was involved in the pair-wise inter-study analyses. The GBP method was more sensitive to the accuracy inflation than that of the MBP method; however, the trend of the predictive power between the MBP and the GBP did not vary much when PPI or the accuracy was used in the performance evaluation. Similar results

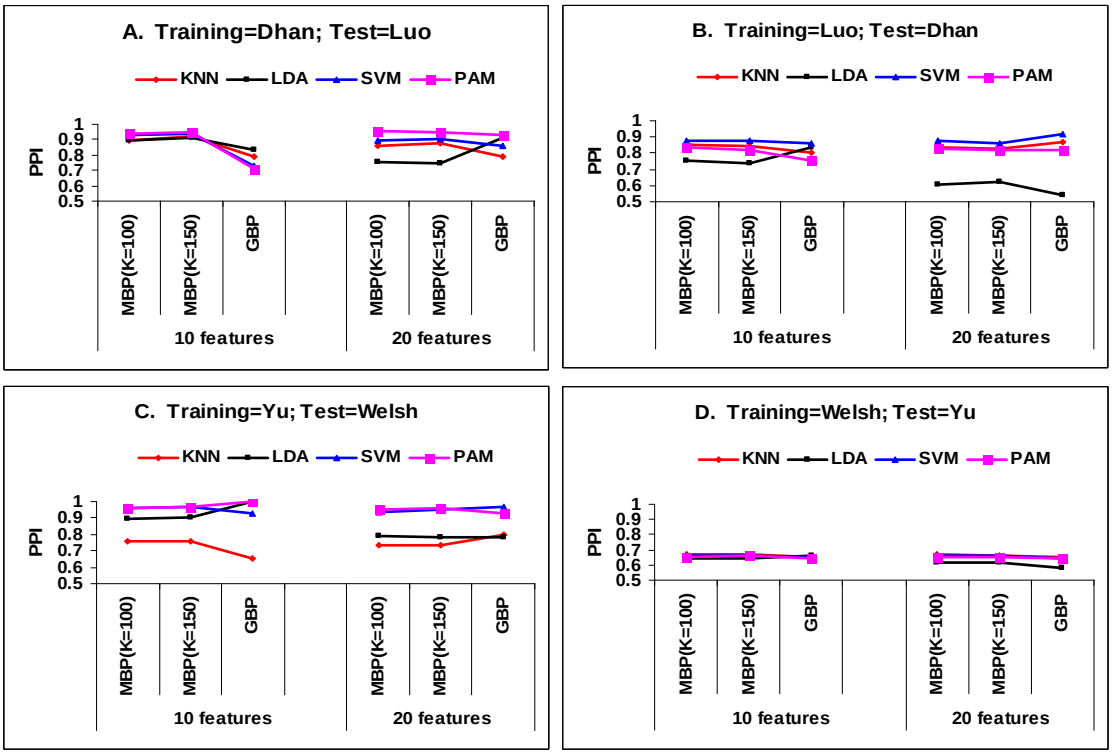

Figure 3.8 Pair-wise inter-study PPI between the MBP and the GBP: The MBP PPI were compared with those of the GBP among four pair-wise prostate cancer inter-study datasets. The parameters were set for the MBP as K=100 and 150, k =10 and 20, $N \geq 3$, and δ=20 and the parameters was set for the GBP as q =10 and 20. Four classifiers were used for the comparison, LDA, KNN, SVM and PAM.

were generated using four prostate cancer datasets. As shown in Figure 3.8, the trend of the predictive power between the MBP and the GBP was consistent using both PPI and the accuracy

across four pair-wise prostate cancer inter-studies. The results suggest that the regular accuracy should be cautiously used when testing data are unbalanced and PPI is robust to the data unbalance. However, the discrepancy between the accuracy and PPI did not affect the results regarding to the comparison of predictive performance between the MBP and the GBP.

3.5 ROBUSTNESS OF THE MBP TO MEASUREMENT VARIABILITY

To determine the stability of the prediction accuracy using the MBP method, white noise was simulated and added to the Luo dataset. The prediction accuracies of the MBP and the GBP were evaluated based on the data with the white noise, using LOOCV approach and LDA, KNN, and SVM classifiers. The noise was added according to $N(0, \sigma)$ at serial proportions of 10%, 20%, 50% and 70% with 0, 0.1σ, 0.5σ and 1.0σ, respectively. The prediction accuracies of the MBP method were quite stable across the noise added to the data (Figure 3.9 panels A-C), whereas the prediction accuracies from the GBP method were unstable when up to 0.5σ of noise was added (Figure 3.9 panel D). As for classifiers, LDA was more robust to noises than the other two classifiers, KNN and SVM. The standard errors were smaller when $\delta = 20$ and $N \geq 3$ were used than those when $\delta = 10$ and $N \geq 1$ were used. However, the pattern of MBP robust to the noise did vary from K and δ. The simulated results supported the hypothesis that the MBP was robust to gene noise.

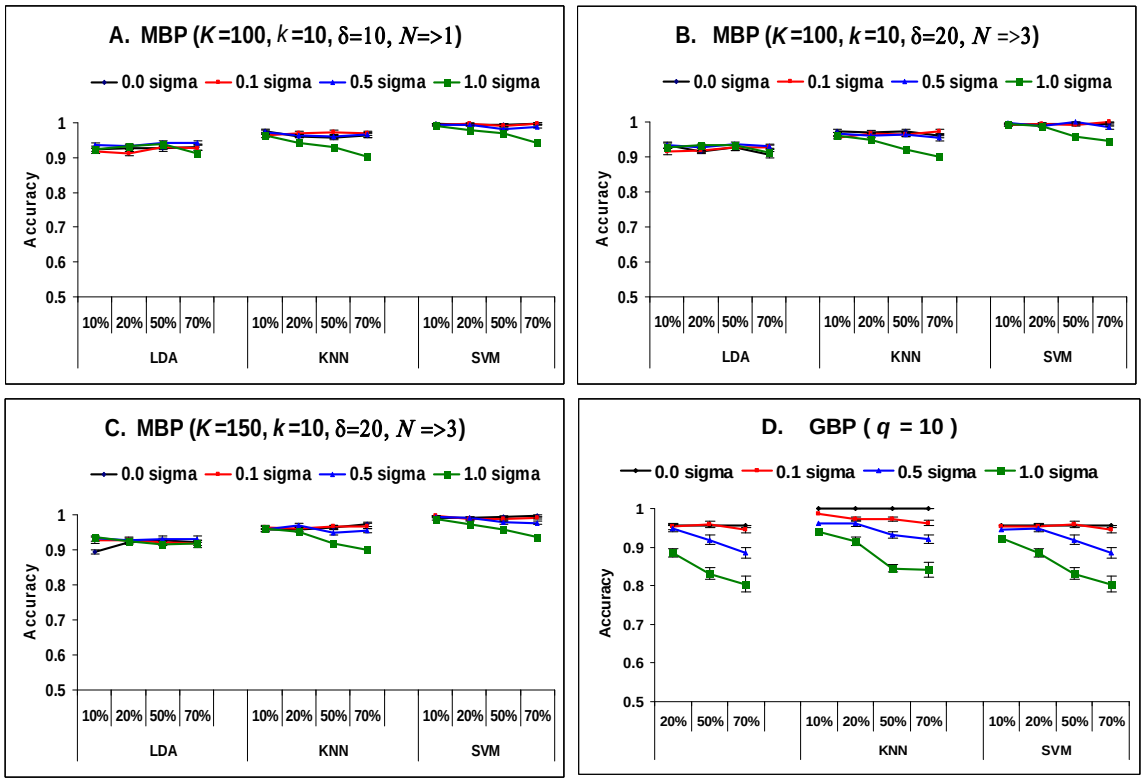

Figure 3.9 Prediction accuracies between the MBP and the GBP after addition of white noise: The white noises with a Gaussian distribution of mean zero and standard deviation of 0, 0.1, 0.5, and 1.0 fold of standard deviation, sigma, derived from the Luo dataset were added to the Luo dataset (3673 genes). The prediction accuracies of the MBP and the GBP were evaluated using LOOCV approach and three classifiers, LDA, KNN, and SVM. The prediction accuracies using the MBP method with gene noise were assessed by selecting K=100, k = 10, $N \geq 1$ and δ =10 (panel A), or K=100, k = 10, $N \geq 3$ and δ =20 (panel B), or K=150, k = 10, $N \geq 3$ and δ =20 (panel C); and the prediction accuracies using GBP method with gene noise were assessed by selecting q =10 (panel D). X axis represents percentage of genes with noise across three classifiers and Y axis represents prediction accuracy performances of different classifiers and amount of noise added. The tests were repeated 30 times and average of accuracies were calculated and represented as mean±SE.

3.6 ROBUSTNESS OF THE MBP TO GENE MISMATCHING

It is common that the pattern of a particular gene behaves differently across platforms due to different probe sequence selection in the platforms. Many genes appearing in one platform may

be missing in another, causing difficulties in applying the GBP method to inter-platform predictions. To evaluate whether the MBP method is robust when genes are missing in the test data, the prediction accuracies were evaluated by splitting an array dataset, Luo, into a training set and a test set and randomly deleting genes from both training and test set at different proportions (π), from 10% to 70% of genes deleted. The procedure was repeated 100 times and the model PSR were counted and tabulated in Table 3.3. The results showed that the MBP method could tolerate genes missing, but the GBP method was very sensitive to the missing genes and the prediction accuracies of the MBP method were stable across the proportions of missing genes (Figure 3.10) using LDA and KNN classifiers, indicating that the borrowed information in the MBP was reliable. The simulated results supported the hypothesis that the MBP was robust to gene missing.

Table 3.3 PSR at gene difference of training and test sets

Method	Training (π)	Test (π)						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
MBP	0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.4	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	0.5	1.00	1.00	1.00	1.00	1.00	1.00	0.97
	0.6	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	0.7	1.00	1.00	1.00	1.00	0.99	1.00	0.93
GBP	0.1	0.33	0.10	0.01	0.00	0.00	0.00	0.00
	0.2	0.40	0.09	0.07	0.01	0.01	0.00	0.00
	0.3	0.33	0.10	0.01	0.01	0.00	0.00	0.00
	0.4	0.43	0.11	0.01	0.01	0.00	0.00	0.00
	0.5	0.42	0.14	0.04	0.01	0.00	0.00	0.00
	0.6	0.36	0.14	0.01	0.01	0.00	0.00	0.00
	0.7	0.30	0.15	0.01	0.01	0.00	0.00	0.00

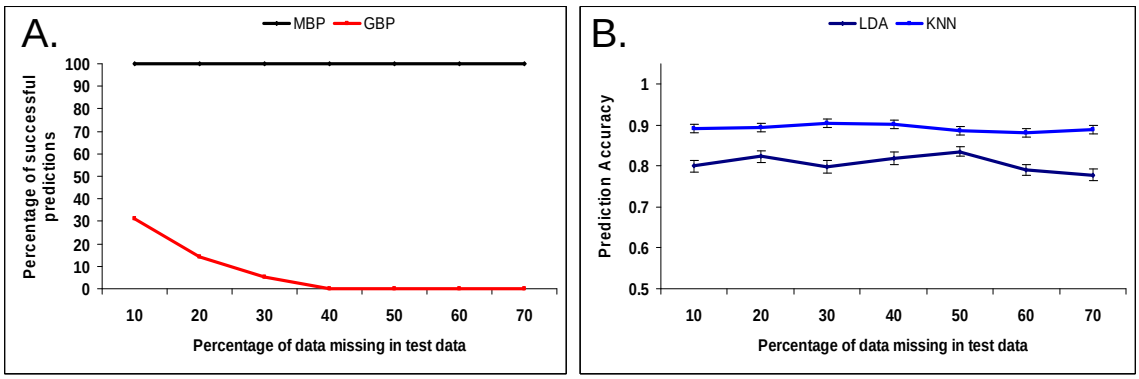

Figure 3.10 The MBP robust to gene missing: Percentage of successful prediction of the MBP versus the GBP using Luo data set with 3673 genes and *K*=100(panel A); Prediction accuracies using the MBP method with gene missing: two classifiers, LDA and KNN were used to assess the prediction accuracies by selecting k = 10 (panel B). And X axis represents percentage of genes deleted from test set given each classifier and Y axis represents performance of different classifiers. The tests were repeated 100 times and average of accuracies were calculated and represented as mean±SE.

3.7 THE MBP PERFORMANCE WITH MULTIVARIATE FEATURE SELECTIONS

From the above results, the MBP showed clear advantages over the GBP method with respect to prediction robustness; however, the feature selections were based on univariate method, i.e. moderate t statistics, to build the prediction models. To modify the feature selection to more sophisticated feature selection methods by considering interaction among the genes, two widely used multivariate feature selection methods, PAM and R-SVM, were used to evaluate prediction performances between the MBP and the GBP, and to assess the prediction performances between the multivariate and univariate feature selections as well. The prediction results using PAM were shown in Figure 3.11 Given $K = 100$ or 150, and $k = 10$, and $\delta = 20$ for the MBP, the average prediction accuracies of twelve cross-platform data analyses were compared between the MBP and the GBP using multivariate PAM (Panels A and B) and the average prediction accuracies

were also compared between univariate and multivariate feature selections(Panels C and D). As for the GBP, the common genes used were in both training and

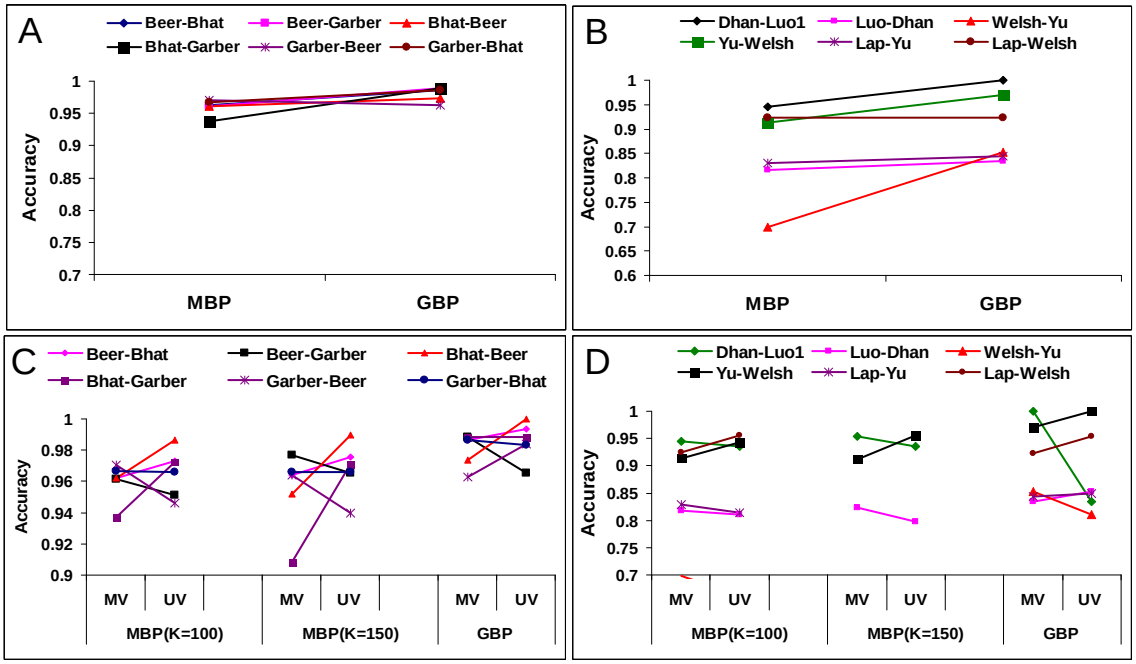

Figure 3.11 The MBP performance using multivariate PAM: In order to select features by considering selection error rate and the correlations among the representative genes of *K* (100 or 150) a multivariate PAM was used to evaluate the MBP performance compared with the GBP; meanwhile, the multivariate and univariate performances were assessed between the MBP and the GBP across three lung cancer datasets and five prostate cancer datasets in inter-study scenarios. Panels A and B show the performances between the MBP and the GBP of six pair-wise inter-studies of lung cancer datasets (panel A) and of six pair-wise inter-studies of prostate cancer datasets (panel B); whereas, panels C and D show the performances between multivariate and univariate selections using the MBP and the GBP of six pair-wise inter-studies of lung cancer datasets (panel C) and of six pair-wise inter-studies of prostate cancer datasets (panel D). The curves in different colors represent average prediction accuracies of 30 repeated tests for six pair-wise lung cancer datasets showing in panel A and panel C and six pair-wise prostate cancer datasets showing in panel B and panel D. Parameters set up for the MBP were K=100 and 150, k =10 for univariate selection, δ =20 and $N \geq 3$, while, univariate selection for the GBP was set up for q =10. MV stands for multivariate selection and UV stands for univariate selection.

test sets known as intersection genes, but predictions failed when all genes in the training sets were used to build prediction models. From Figure 3.11 A-B, the results showed that there was

no significant difference of the prediction accuracies between using the MBP and using the GBP, across six pair-wise lung cancer data tests (Panel A) and six pair-wise prostate cancer data tests (Panel B), and from Figure 3.11 C-D, the prediction performances between the univariate and multivariate feature selection approaches were quite similar when the MBP was used at $K = 100$, whereas, the performances between the two feature selection approaches were data dependent when the GBP or MBP at $K = 150$ was used. Similar results were observed using another

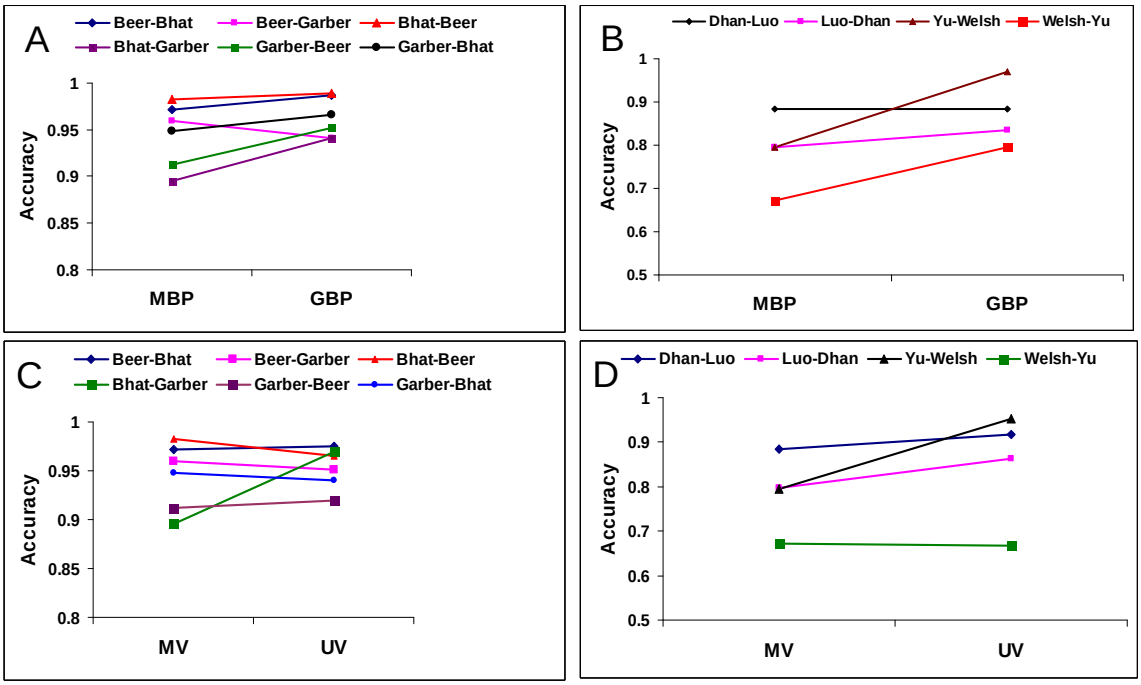

Figure 3.12 The MBP performance using multivariate R-SVM: In order to select features by considering selection error rate and the correlations among the representative genes of *K* (100) a multivariate R-SVM was used to evaluate the MBP performance compared with the GBP; meanwhile, the multivariate and univariate performances were assessed between the MBP and the GBP across three lung cancer datasets and four prostate cancer datasets in inter-study scenarios. Panels A and B show the performances between the MBP and the GBP of six pair-wise inter-studies of lung cancer datasets (panel A) and of four pair-wise inter-studies of prostate cancer datasets (panel B); whereas, panels C and D show the performances between multivariate and univariate selections using the MBP of six pair-wise inter-studies of lung cancer datasets (panel C) and of four pair-wise inter-studies of prostate cancer datasets (panel D). The curves in different colors represent average prediction accuracies of 30 repeated tests for six pair-wise lung cancer datasets showing in panel A and panel C and four pair-wise prostate cancer datasets showing in panel B and panel D. Parameters set up for the MBP were *K*=100, *k* =10 for univariate selection, δ =20 and *N* ≥3, while, univariate selection for the GBP was set up for *q* =10. MV stands for multivariate selection and UV stands for univariate selection.

multivariate feature selection method, R-SVM. As shown in Figure 3.12, the prediction performances of MBP and GBP were data dependent (Panels A and B). As for the performances between the univariate and the multivariate feature selections using R-SVM were quite similar, however, also data dependent (Panels C and D). The results suggest that the prediction performances between the MBP and the GBP are indistinguishable and that there was no evidence that the multivariate feature selection method outperformed the univairate method using both PAM and R-SVM, indicating no significant loss of predictive power using the MBP method. Lai et al reported that the classification performance was indistinguishable between using univariate and multivariate feature selection methods (Lai, Reinders et al. 2006)

3.8 THE MBP PERFORMANCE USING PW-*K*-MEANS CLUSTERING METHOD

One of the factors might affect the MBP performance is the quality of the clusters. Given *K*-means method forces all genes into *K* clusters resulting in outlier genes in certain clusters, PW-*K*-means method was developed to eliminating the gene outliers out of clusters and made the clusters tight by adding penalty weights when genes were assigned to certain clusters, such that the cluster quality would be improved (Tseng 2007). To evaluate whether such a procedure improved the MBP performance, PW-*K*-means clustering was used to replace *K*-means clustering to group genes for the MBP and the results using PW-*K*-means method was compared with those using *K*-means method. As shown in Figure 3.13, the results showed that using PW-*K*-means clustering did not improve the MBP performance given a range of penalty parameter lambda (0.3, 0.5, 0.7, 0.9, 1.1 and 999) compared with using *K*-means method across four

prostate cancer and two ling cancer datasets, indicating that the MBP method was robust enough to overcome gene noise caused cluster outliers. In this section, we also used PPI to evaluate the

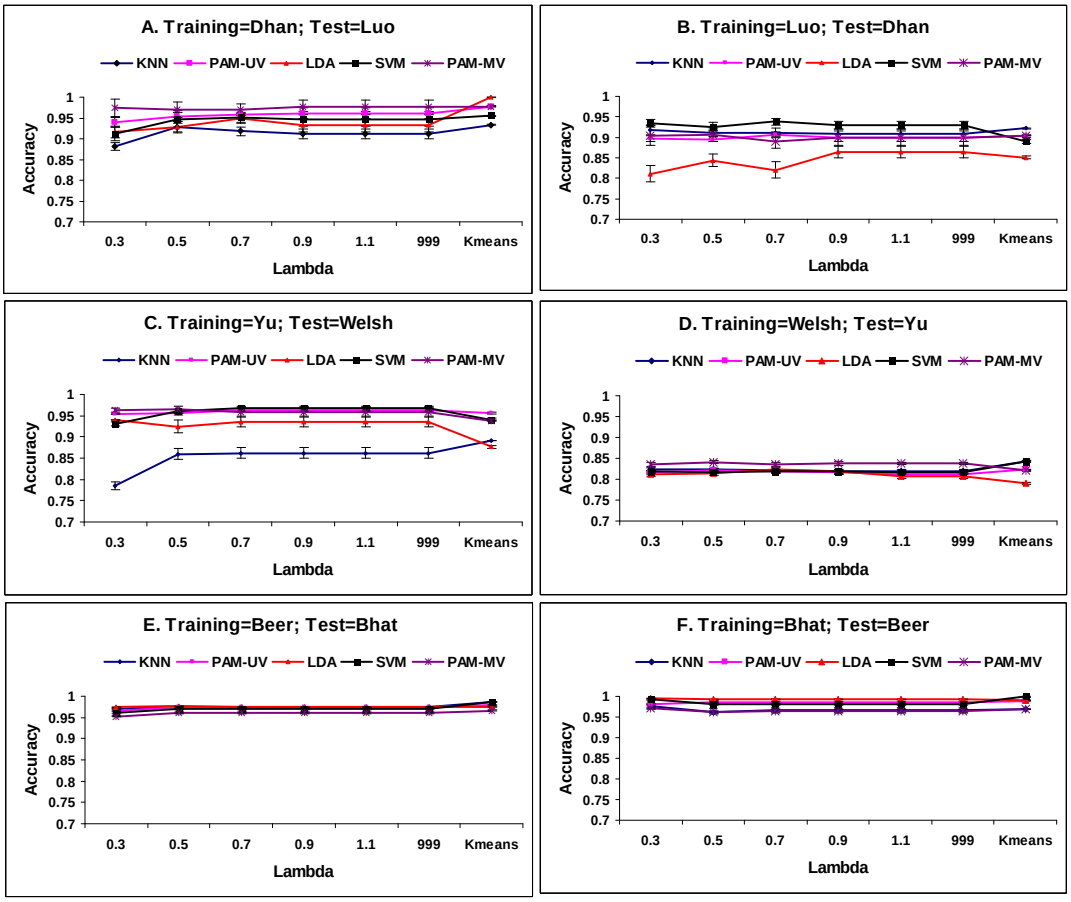

Figure 3.13 The MBP prediction accuracy using PW-*K*-means clustering as gene grouping method: In order to improve quality of gene clusters by eliminating noise genes PW-*K*-means clustering method was used to group genes based on their expression similarities. The elimination of noise genes was determined by a parameter lambda and the smaller lambda would eliminate more noise genes. Panels A-D are the MBP inter-study performances across the four prostate cancer datasets and panels E-F are the MBP inter-study performances across two lung cancer datasets. The curves in different colors represent average prediction accuracies of 30 repeated tests for four univariate classifiers (Dark blue: KNN, Red: LDA, Black: SVM, Magenta: PAM-UV)) and a multivariate classifier (Brown: PAM-MV). Variations from multiple tests are represented as standard errors (error bars). X axis represents a range of lambdas set for PW-*K*-means method and a *K*-means method. Parameters set up for the MBP were K=100, k =10, δ =20 and $N \geq 3$.

MBP performance in addition to prediction accuracy to avoid inflated prediction accuracy. As shown in Figure 3.14, PPI was generally smaller than the prediction accuracy shown in Figure

3.13 across four prostate cancer and two lung cancer datasets and showed greater variations among four classifiers; however, the trend of the predictive power between using the prediction accuracy and PPI did not vary across penalty parameters, lambda used.

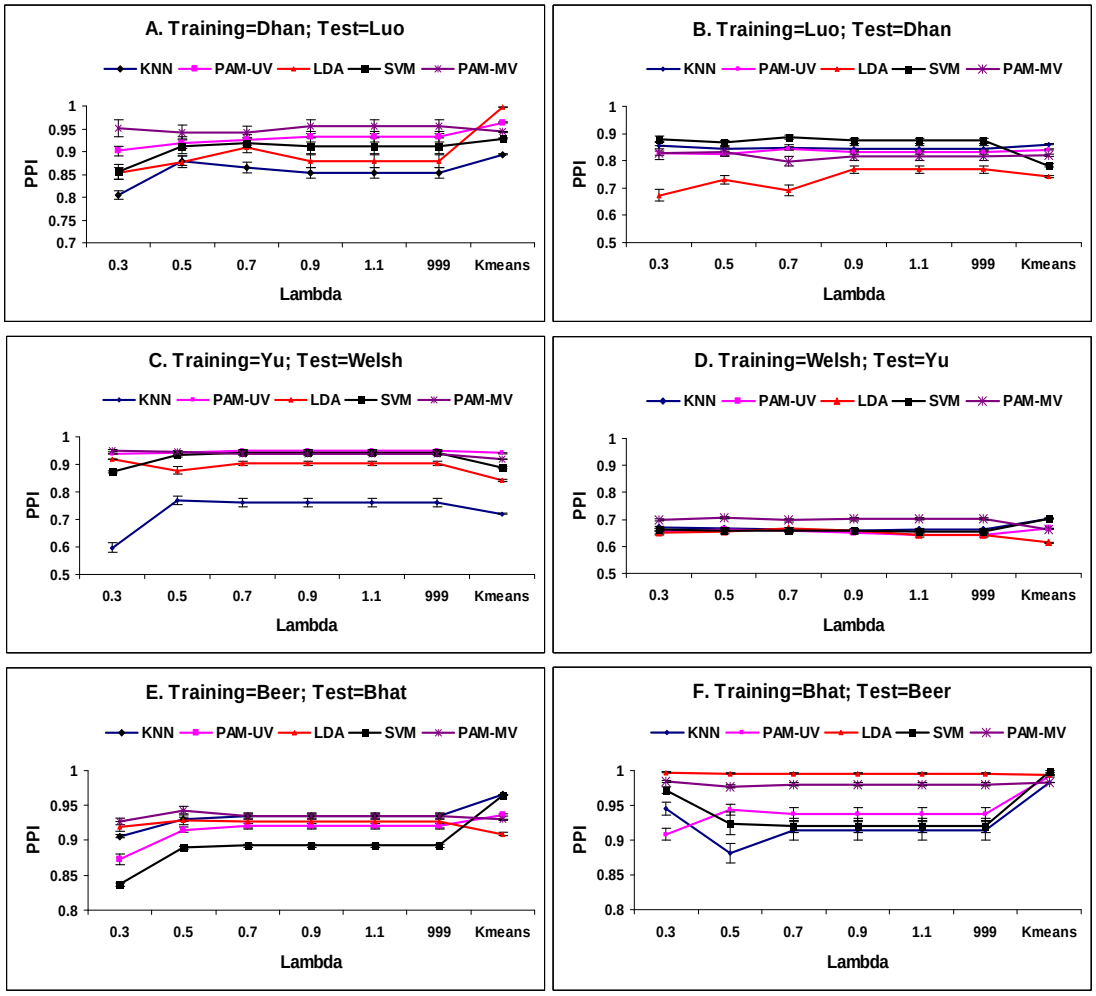

Figure 3.14 The MBP PPI using PW-*K*-means clustering as gene grouping method: In order to improve quality of gene clusters by eliminating noise genes PW-*K*-means clustering method was used to group genes based on their expression similarities. The elimination of noise genes was determined by a parameter lambda and the smaller lambda would eliminate more noise genes. Panels A-D are the MBP inter-study performances across the four prostate cancer datasets and panels E-F are the MBP inter-study performances across two lung cancer datasets. The curves in different colors represent average prediction accuracies of 30 repeated tests for four univariate classifiers (Dark blue: KNN, Red: LDA, Black: SVM, Magenta: PAM-UV)) and a multivariate classifier (Brown: PAM-MV). Variations from multiple tests are represented as standard errors (error bars). X axis represents a range of lambdas set for PW-*K*-means method and a *K*-means method. Parameters set up for the MBP were K=100, k =10, δ =20 and $N \geq 3$.

3.9 THE MBP PERFORMANCE USING MEDIAN GENE VERSUS SAMPLE MEDIAN

To reduce prediction module dimension from $G_{k'}^{tr} \times s$ to $1 \times s$ vector to fit prediction model, we proposed a new metric, a representative gene or a median gene for each module, instead of commonly used sample mean or sample median by computing average point for each sample. The advantage of using the median gene is to identify an actual gene with biological meaning for the MBP method; however, the predictive power of using the median gene needs to be justified by comparing the predictive accuracies between the median gene and the sample median. Table 3.4 shows that the averages of prediction accuracies and standard errors from 100 repeated LOOCV of three lung cancer datasets, respectively. The results demonstrated that the predictive power was equivalent between using the median gene and the sample median using LDA. The averages of prediction accuracy of using the two methods were 0.93 and 0.93, respectively, indicating that using the median gene is feasible.

Table 3.4 Accuracies of median gene and sample median

Data	Median Gene	Sample Median
Beer	**0.988±0.008**	**0.989±0.008**
Bhat	**0.971±0.003**	**0.971±0.003**
Garber	**0.980±0.010**	**0.980±0.011**
Mean	**0.913**	**0.913**

δ =10, K =100, k =10 and mean±SD representing 100 repeats

4.0 CONCLUSIONS AND DISUCUSSIONS

An ideal prediction model would possess high prediction accuracy, model robustness, and model simplicity. To pursue these standards, we developed the MBP method, which borrowed information from genes sharing similar expression patterns. The goals of building the MBP model were to yield robust prediction models by solving gene noise and gene missing problems existing in the GBP and to obtain simple models by using all genes of a dataset to build prediction models without information of test datasets. The results of the current study showed that the prediction accuracies of the MBP method were not worse than those of the GBP method both within-study and inter-studies. However, the MBP method was superior to the GBP method in being both robust to gene noise and missing genes. Therefore, the MBP method would be useful for building a prediction model based on training data without information of test data. Nevertheless, there are three significant factors that need to be addressed for the MBP.

As for model accuracy, it seems based on common sense that the GBP method would have higher prediction accuracies for specific datasets since genes selected into a prediction signature usually have higher differentiated values than those not selected. Thus, one might expect highly differentiated genes to be better than less differentiated genes at predicting patient outcome. However, gene signatures are not always unique. It has been shown that a gene signature may not always be unique for prediction models. It was originally shown in a seminal study that a 70-gene signature could predict breast cancer patient survival (van 't Veer, Dai et al.

2002). The investigators using the same dataset later identified an additional 6 classifiers of genes that performed equally well to the 70-gene signature (Ein-Dor, Kela et al. 2005). Also, disparity in using different gene signatures to predict similar outcomes has been reported (Sorlie, Perou et al. 2001; van 't Veer, Dai et al. 2002; Ramaswamy, Ross et al. 2003), which may lead to unstable predictions. The stability of the MBP method observed in the present study may be the result of grouping genes sharing a similar expression pattern and selecting a gene that can represent the group of genes. It has been postulated that using a cluster average would yield higher prediction accuracy under certain conditions (Park, Hastie et al. 2007). Although in this preliminary analysis, there was no significant difference in prediction accuracy between the GBP method and module based methods, both within study and cross-platform, the prediction accuracy might be affected by preset parameters. Different parameters should be tested in the near future. Even if the MBP method cannot surpass the GBP method in prediction accuracy, the prediction robustness remains its major advantage.

Regarding model simplicity, the clinical utility of a genomic prediction model relies heavily on the model's simplicity and reproducibility. Recent cross-platform analyses used intersection genes across datasets (Bloom, Yang et al. 2004; Bhanot, Alexe et al. 2005; Nilsson, Andersson et al. 2006; Bosotti, Locatelli et al. 2007; Cheadle, Becker et al. 2007), an approach that required information from all datasets involved in the analysis. This approach is good for cross- platform meta-analysis, but it is limited for cross-platform prediction. There are two elements needed for a prediction: 1) a selected gene signature and 2) a prediction model. A prediction is possible only when a test sample or partial test data is available. When intersection genes are used to build a prediction model, the selected prediction signature must be adjusted for new test samples, which would be inconvenient for model application. Furthermore, it may lose

information of training data by only including intersection genes to build the prediction model. Thus, the MBP method would be easier for application.

In the sense of model reproducibility, lack of reproducibility hinders the application of genomic prediction models. Many factors may affect model reproducibility. The MBP method focuses on two factors to increase model reproducibility, missing genes, and gene noise. The robustness of the MBP method toward missing genes was proved by estimating the probability of model failure due to missing genes. The robustness of the method regarding gene noise was assessed by testing on the Luo dataset. This result is a quite common phenomenon in microarray data. Although the MBP method showed model robustness to added noise to the Luo data, the pattern of noise added may not totally mock all data variations. Further study will focus on adding different noises beside normal white noise and high variable data.

Beside the value of the MBP contributing to clinical application, the MBP also added its value to microarray analysis methodology. First, this is the first time we identified that k-means cluster size followed a multinomial distribution and proposed a cluster merging step to avoid model prediction failure due to gene mismatch in microarray cross platform analysis. Second, we used a representative gene with the closest distance to all other genes within a module to summarize the module information, which is an actual gene with its annotation and interpretation rather than using extract information such as average gene expression or gene sample median. Third, the MBP reduced redundant features by summarizing similar gene expression profiles within each module, diminishing model collinearity and adding a novel technique for data reduction. The limitations of the MBP method are a lack of biological correlation information among the genes within each module and no enhancement of prediction accuracy, however, these issues lead to further investigation on the nature of the MBP method.

5.0 FUTURE DIRECTIONS

Based on the results, the MBP method was proved to be plausible for the robust prediction of microarray cross-platform data and the goals of the study were achieved. However, the analyses were not near perfect to implement the validity and adequacy of the method. The further analyses will include the following topics:

To compare the model accuracy and robustness with metagene method Although the other methods using cluster or group gene information instead of using individual gene information for prediction were not designed to achieve the same goal as the module based method was, these methods share the essence of using cluster of genes information. The prediction accuracies and robustness will be compared among these methods using same dataset.

To compare the MBP with biological pathway modules MBP is an empirical data dimension reduction method. Though it can robustly predict patient outcome via inter-study data the mechanism of elements in a module is unknown. Biological pathway tools, such as GO, or KEEG, will be applied to check the biological relationships among the elements.

To test the MBP in other biomarker platforms In this study, the MBP was evaluated in RNA expression microarray platforms. Based on the nature of high dimension data outputs of '-omics', the MBP can be extended to analyze DNA microarray, protein microarray, and metabolite mass spectra data. The MBP will be evaluated using other '-omics' data.

BIBLIOGRAPHY

Beer, D. G., S. L. Kardia, et al. (2002). "Gene-expression profiles predict survival of patients with lung adenocarcinoma." Nat Med **8**(8): 816-24.

Bhanot, G., G. Alexe, et al. (2005). "Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories." Genome Inform **16**(1): 233-44.

Bhattacharjee, A., W. G. Richards, et al. (2001). "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." Proc Natl Acad Sci U S A **98**(24): 13790-5.

Bild, A. H., A. Potti, et al. (2006). "Linking oncogenic pathways with therapeutic opportunities." Nat Rev Cancer **6**(9): 735-41.

Bild, A. H., G. Yao, et al. (2006). "Oncogenic pathway signatures in human cancers as a guide to targeted therapies." Nature **439**(7074): 353-7.

Bloom, G., I. V. Yang, et al. (2004). "Multi-platform, multi-site, microarray-based human tumor classification." Am J Pathol **164**(1): 9-16.

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

Bussey, K. J., D. Kane, et al. (2003). "MatchMiner: a tool for batch navigation among gene and gene product identifiers." Genome Biol **4**(4): R27.

Cantor, C. R., A. Mirzabekov, et al. (1992). "Report on the sequencing by hybridization workshop." Genomics **13**(4): 1378-83.

Cheadle, C., K. G. Becker, et al. (2007). "A rapid method for microarray cross platform comparisons using gene expression signatures." Mol Cell Probes **21**(1): 35-46.

Cristianini, N. and J. Shawe-Taylor (2000). An Introduction to Support Vector Machines and other kernel-based learning methods. New York, Cambridge University Press.

Dasarathy, B. (1991). Nearest neighbor (NN) norms: Nn pattern classification techniques (Unknown Binding), IEEE Computer Society Press Tutorial.

Dhanasekaran, S. M., T. R. Barrette, et al. (2001). "Delineation of prognostic biomarkers in prostate cancer." Nature **412**(6849): 822-6.

Dobbin, K. and R. Simon (2005). "Sample size determination in microarray experiments for class comparison and prognostic classification." Biostatistics **6**(1): 27-38.

Dobbin, K. K., Y. Zhao, et al. (2008). "How large a training set is needed to develop a classifier for microarray data?" Clin Cancer Res **14**(1): 108-14.

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

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics **7**: 179-188.

Garber, M. E., O. G. Troyanskaya, et al. (2001). "Diversity of gene expression in adenocarcinoma of the lung." Proc Natl Acad Sci U S A **98**(24): 13784-9.

Guyon, I., J. Weston, et al. (2002). "Gene Selection for Cancer Classification using Support Vector Machines." Machine Learning **46**: 389-422.

Hartigan, J. A. and M. A. Wong (1979). "A K-means clustering algorithm." *Applied Statistics* **28**: 100-108.

Hsu, C. and C. Lin (2002). "A comparison of methods for multiclass support vector machines." IEEE Transactions on Neural Networks **13**(2): 415-425.

Huang, E., S. H. Cheng, et al. (2003). "Gene expression predictors of breast cancer outcomes." Lancet **361**(9369): 1590-6.

Ioannidis, J. P. (2007). "Is molecular profiling ready for use in clinical decision making?" Oncologist **12**(3): 301-11.

Irizarry, R. A., D. Warren, et al. (2005). "Multiple-laboratory comparison of microarray platforms." Nat Methods **2**(5): 345-50.

Kapp, A. V. and R. Tibshirani (2007). "Are clusters found in one dataset present in another dataset?" Biostatistics **8**(1): 9-31.

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection." Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence **2**(12): 1137-1143.

Kuhn, A., R. Luthi-Carter, et al. (2008). "Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package annotationTools." BMC Bioinformatics **9**(1): 26.

Kulesh, D. A., D. R. Clive, et al. (1987). "Identification of interferon-modulated proliferation-related cDNA sequences." Proc Natl Acad Sci U S A **84**(23): 8453-7.

Lai, C., M. J. Reinders, et al. (2006). "A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets." BMC Bioinformatics **7**: 235.

Lapointe, J., C. Li, et al. (2004). "Gene expression profiling identifies clinically relevant subtypes of prostate cancer." Proc Natl Acad Sci U S A **101**(3): 811-6.

Luo, J., D. J. Duggan, et al. (2001). "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling." Cancer Res **61**(12): 4683-8.

Mardia, K., J. Kent, et al. (1979). Multivariate Analysis. London, Academic Press.

Nilsson, B., A. Andersson, et al. (2006). "Cross-platform classification in microarray-based leukemia diagnostics." Haematologica **91**(6): 821-4.

Paik, S., S. Shak, et al. (2004). "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer." N Engl J Med **351**(27): 2817-26.

Park, M. Y., T. Hastie, et al. (2007). "Averaged gene expressions for regression." Biostatistics **8**(2): 212-27.

Park, P. J., Y. A. Cao, et al. (2004). "Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference." J Biotechnol **112**(3): 225-45.

Pittman, J., E. Huang, 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**(22): 8431-6.

Potti, A., H. K. Dressman, et al. (2006). "Genomic signatures to guide the use of chemotherapeutics." Nat Med **12**(11): 1294-300.

Potti, A., S. Mukherjee, et al. (2006). "A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer." N Engl J Med **355**(6): 570-80.

Poustka, A., T. Pohl, et al. (1986). "Molecular approaches to mammalian genetics." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 131-9.

Pusztai, L. and B. Leyland-Jones (2008). "Promises and caveats of in silico biomarker discovery." Br J Cancer **99**(3): 385-6.

Pusztai, L., C. Mazouni, et al. (2006). "Molecular classification of breast cancer: limitations and potential." Oncologist **11**(8): 868-77.

Ramaswamy, S., K. N. Ross, et al. (2003). "A molecular signature of metastasis in primary solid tumors." Nat Genet **33**(1): 49-54.

Ross, J. S., C. Hatzis, et al. (2008). "Commercialized multigene predictors of clinical outcome for breast cancer." Oncologist **13**(5): 477-93.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Schulze, A. and J. Downward (2001). "Navigating gene expression using microarrays--a technology review." Nat Cell Biol **3**(8): E190-5.

Schulze, A., B. Nicke, et al. (2004). "The transcriptional response to Raf activation is almost completely dependent on Mitogen-activated Protein Kinase Kinase activity and shows a major autocrine component." Mol Biol Cell **15**(7): 3450-63.

Segal, E., N. Friedman, et al. (2005). "From signatures to models: understanding cancer using microarrays." Nat Genet **37 Suppl**: S38-45.

Segal, E., N. Friedman, et al. (2004). "A module map showing conditional activity of expression modules in cancer." Nat Genet **36**(10): 1090-8.

Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-76.

Shi, L., W. Tong, et al. (2005). "Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential." BMC Bioinformatics **6 Suppl 2**: S12.

Shi, L., W. Tong, et al. (2004). "QA/QC: challenges and pitfalls facing the microarray community and regulatory agencies." Expert Rev Mol Diagn **4**(6): 761-77.

Sorlie, T., C. M. Perou, et al. (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proc Natl Acad Sci U S A **98**(19): 10869-74.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J Mol Biol **98**(3): 503-17.

Spang, R., H. Zuzan, et al. (2002). "Prediction and uncertainty in the analysis of gene expression profiles." In Silico Biol **2**(3): 369-81.

Tamayo, P., D. Scanfeld, et al. (2007). "Metagene projection for cross-platform, cross-species characterization of global transcriptional states." Proc Natl Acad Sci U S A **104**(14): 5959-64.

Tan, P. K., T. J. Downey, et al. (2003). "Evaluation of gene expression measurements from commercial microarray platforms." Nucleic Acids Res **31**(19): 5676-84.

Tibshirani, R., T. Hastie, et al. (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proc Natl Acad Sci U S A **99**(10): 6567-72.

Tseng, G. C. (2007). "Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data." Bioinformatics **23**(17): 2247-55.

van 't Veer, L. J., H. Dai, et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer." Nature **415**(6871): 530-6.

van Vliet, M. H., C. N. Klijn, et al. (2007). "Module-based outcome prediction using breast cancer compendia." PLoS ONE **2**(10): e1047.

Welsh, J. B., L. M. Sapinoso, et al. (2001). "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer." Cancer Res **61**(16): 5974-8.

West, M., C. Blanchette, et al. (2001). "Predicting the clinical status of human breast cancer by using gene expression profiles." Proc Natl Acad Sci U S A **98**(20): 11462-7.

West, M., G. S. Ginsburg, et al. (2006). "Embracing the complexity of genomic data for personalized medicine." Genome Res **16**(5): 559-66.

Wong, D. J., D. S. Nuyten, et al. (2008). "Revealing targeted therapy for human cancer by gene module maps." Cancer Res **68**(2): 369-78.

Yu, Y. P., D. Landsittel, et al. (2004). "Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy." J Clin Oncol **22**(14): 2790-9.

Zhang, X., X. Lu, et al. (2006). "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data." BMC Bioinformatics **7**: 197.