MOLECULAR PROFILING IN BREAST CANCER AND TOXICOGENOMICS

Jiangang Liu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics,
Indiana University

December 2010

Accepted by the Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Yaoqi Zhou, Ph.D., Chair

_____
A. Keith Dunker, Ph.D.

Doctoral Committee

_____
Jake Chen, Ph.D.

_____
Vladimir N. Uversky, Ph.D.

_____
Yunlong Liu, Ph.D.

November 5, 2010

_____
Dan S. Li, Ph.D.

Dedicated to my dad who taught me to never give up

ACKNOWLEDGMENTS

This work is carried out at the Center of Computational Biology and Bioinformatics, IUPUI and the Bioinformatics Group at Eli Lilly and Company. I am indebted to a large number of people who have inspired and helped me throughout the years of studies.

I would like to thank my primary advisor Dr. A. Keith Dunker for generously giving me the maximum of freedom to complete this body of work under his guidance and support throughout the past years. I have benefited greatly from his brilliance and excellent combination of knowledge. Except for being a great advisor, Keith also has a fantastic personality. His sense of humor and encouragement for me make it a great pleasure to work with him. The kindness and hospitality atmosphere he and his wife hosted at holiday parties is very happy memory for me, and makes me feel like at home.

I also would like to thank my advisor, supervisor at Lilly, basketball buddy, and a great friend, Dr. Dan Li, for his significant influence on my career development. He is one who leads me into this field, and provides strong support, mentoring, and encouragement to keep me going to reach this point. I have learned so much from him, including how to think analytically, how to do research, and how to write and present research.

Dr. Vladimir N. Uversky played an important role in the aspect of my work on protein disorder. I would like to express my appreciation for his

brilliant and insightful advices and ideas as well as for that he always believes in me, which means a lot to me.

I want to express my sincere appreciation of the efforts and time of the members of my research committee, Drs. Yaoqi Zhou, Jake Chen, and Yunlong Liu, for their strong guidance and encouragement in my research and the dissertation preparation. I have gained a great deal from them for reviewing my research proposal and dissertation and for providing the critical advices, suggestions, comments, and criticism.

My special thanks to Dr. Tao Wei who has been a great colleague and friend to me at Lilly Bioinformatics Group. Without his support, I would not have completed this work. I am very grateful to him for the skills I learned from him and for his insightful ideas and comments. I was often inspired by discussing with him and I really enjoyed working with him.

All my co-authors have tremendous contributions to the work in this dissertation. In particular, Dr. Craig Thomas has brought me the opportunity to work in Toxicology and Toxicogenomics. I have great experiences to work with him and people in his investigative toxicology group at Lilly.

I am also thankful to Dr. Narayanan B. Perumal, my classmate Christopher J. Oldfield, and people in Dr. Keith Dunker's and Dr. Jake Chen's group for their wonderful helps.

I gratefully acknowledge the financial support received from Eli Lilly and Company.

Finally, I want to give my deep thanks to my parents, wife, and children. I am very grateful for their unconditional love and support for me. Particularly, without all that my wife Hongling Xiao has done to help me, I would not be where I am today.

ABSTRACT

Jiangang Liu

MOLECULAR PROFILING IN BREAST CANCER AND TOXICOGENOMICS

This dissertation presents a body of research that attempts to tackle the 'overfitting' problem for gene signature and biomarker development in two different aspects (mechanistically and computationally).

In achievement of a deeper understanding of cancer molecular mechanisms, this study presents new approaches to derive gene signatures for various biological phenotypes, including breast cancer, in the context of well-defined and mechanistically associated biological pathways. We identified the pattern of gene expression in the cell cycle pathway can indeed serve as a powerful biomarker for breast cancer prognosis. We further built a predictive model for prognosis based on the cell cycle gene signature, and found our model to be more accurate than the Amsterdam 70-gene signature when tested with multiple gene expression datasets generated from several patient populations. Aside from demonstrating the effectiveness of dimensionality reduction, phenotypic dissection, and prognostic or diagnostic prediction, this approach also provides an alternative to the current methodology of identifying gene expression markers that links to biological mechanism.

This dissertation also presents the development of a novel feature selection algorithm called Predictive Power Estimate Analysis (PPEA) to

computationally tackle on overfitting. The algorithm iteratively apply a two-way bootstrapping procedure to estimate predictive power of each individual gene, and make it possible to construct a predictive model from a much smaller set of genes with the highest predictive power. Using DrugMatrix™ rat liver data, we identified genomic biomarkers of hepatic specific injury for inflammation, cell death, and bile duct hyperplasia. We demonstrated that the signature genes were mechanistically related to the phenotype the signature intended to predict (e.g. 17 out of top 20 genes for inflammation selected by PPEA were members of NF-kB pathway, which is a key pre-inflammatory pathway for a xenobiotic response). The top 4 gene signature for BDH has been further validated by QPCR in a toxicology lab. This is important because our results suggest that the PPEA model not largely deters the over-fitting problem, but also has the capability to elucidate mechanism(s) of drug action and / or of toxicity.

Yaoqi Zhou, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ALT	Alanine Aminotransferase
BDH	bile duct hyperplasia
cDNA	complementary deoxyribonucleic acid
DILI	drug induced liver injury
DNA	deoxyribonucleic acid
EPV	events per variable
FC	fold change
FDA	U.S. Food and Drug Administration
GEO	Gene Expression Omnibus
GO	gene ontology
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAS5	Affymetrix Microarray Suite 5
PAM	prediction analysis for microarrays
PCA	principal component analysis
PCR	polymerase chain reaction
PPEA	predictive power estimate analysis
QPCR	quantitative polymerase chain reaction
QRT-PCR	quantitative reverse transcription PCR
SNPs	single-nucleotide polymorphism
SVM	support vector machine

All gene symbols can be found at

www.ncbi.nlm.nih.gov/sites/entrez?db=gene.

CHAPTER ONE: INTRODUCTION

1.1. Introduction

For personal medicine, the ability to dissect biological complexity to understand the unique characteristics of the individual patient is the key in developing effective therapeutic strategies. However, the capacity to understand this complexity is often limited by the ability to define relevant phenotypes. There is perhaps no better example of this challenge than that seen in cancer (Nevins et al, 2007). The complexity of the oncogenic process, involving the somatic acquisition of large numbers of mutations coupled with variability in a host's genetic constitution, produces a disease of enormous complexity (Potti & Nevins, 2008). Traditional methods of characterizing tumors, based on gross visual information along with a limited number of biochemical assays, do provide a way to define tumor subgroups with distinct biology, it is clear that these classifications are imprecise, creating heterogeneous groupings of tumors and patients. Numerous examples show that expression profiling, a technology to measure gene expression on a genome-wide scale to identify patterns of gene expression, and gene signatures, patterns that can be dynamic in response to both physiological and patho-physiological processes, can dissect this heterogeneity and complexity. The promise for improvement in treatment decision making also attracted great attention to these gene signatures. In this study we focus on one particular aspect - to develop gene expression signatures or biomarkers that

reflect phenotypes, and to use these signatures as measures of signaling pathway activity and other aspects of biological or pathological endpoints or events. Thus, the aim of this dissertation is to develop, perfect, and validate the methodologies built around this aspect through following two different mechanistic and computational approaches.

1.2. Specific aim I: Derive gene signatures for breast cancer prognosis in the context of known biological pathways.

Modeling by using the high dimensional and high noisy genomic data is prone to over-fitting and often consist of a large number of genes with no obvious functional relevance to the biological effect the model intends to predict, which can make it challenging to interpret the modeling results. Also, there are significantly challenges lie in interpreting the profiling results to gain insights into biological mechanisms. Ultimately, finding gene signatures that can be linked to the molecular mechanisms of cancer development is critical for translating these markers into the clinic. As described in Chapter Three, we attempt to address these two issues mentioned above by developing a novel approach to identify gene signatures mechanistically for cancer prognosis in the context of known biological pathways. Our rationale for this approach is if we attempt to identify gene signatures within well defined pathways, not only does this approach alleviate the dimensionality problem, but the mechanism-based gene signatures should also be more biologically relevant than the signatures derived from the entire human transcriptome. In preliminary

studies, we first tested more than a thousand well-defined signal pathways or functional gene sets from several public or commercial available sources such as Ingenuity, GO, KEGG, Biocarta, and Iconix. Preliminary results demonstrate that several signal pathways have been potentially linked to cell cycle, DNA damage response, DNA repair capacity, mitotic checkpoints, hypoxia, and other tumor micro-environmental factors such as glucose deprivation and oxidative stresses. Motivated by the initial success, we identified the pattern of gene expression in the cell cycle pathway can indeed serve as a powerful biomarker for breast cancer prognosis. We further built a predictive model for prognosis based on the cell cycle gene signature, and found our model to be more accurate than the Amsterdam 70-gene signature when tested with multiple gene expression datasets generated from several patient populations (Liu et al, 2008). To our knowledge, this is the first gene signature that was systematically derived directly from well-defined cancer-associated signal pathways. The result suggests that the roles of this pathway and its interaction with oncogenic networks are key and potentially critical to understand and predict behaviors related to sensitivity to cell proliferation inhibitors. These initial achievements provide strong incentives for further expand, refinement, and validation of the model by integrating genomic data with other biological prior knowledge.

1.3. Specific aim II: Develop and Apply new feature selection algorithm to alternatively tackle the curse-dimensionality issue for biomarker identification from microarray data

Numerous recent studies demonstrated that gene expression signatures not only outperformed traditionally used clinical parameters in outcome prediction, but also contribute to a better understanding of the biological mechanism. However, the gene signature obtained for the same clinical types of patients by different groups differed widely and had only very few genes in common. This lack of agreement raised doubts about the reliability and robustness of the reported predictive signature, and the main source of the problem was shown to be the risk of overfitting. Overfitting that arises when the number of training samples is small and the number of attributes or features (i.e., the genes) is comparatively large. In such a situation, we can easily train a classifier that correctly describes the training data but performs poorly on an independent set of data. It is of paramount importance to reduce the dimensionality of the data by deleting unsuitable features. As described in Chapter Four, we developed a novel algorithm, called PPEA, Predictive Power Estimate Analysis in order to improve the performance of learning algorithms in computational aspect. This algorithm iteratively applies a two-way bootstrapping procedure to manage the number of genes equal or less than the number of samples in each splitting subset using for machine learning, and then assessed the merit of each individual feature by evaluating its strength of class predictability. This gave us the ability

to find a small size of feature subsets with high classification performance. Using DrugMatrix™ rat liver data in our studies, we identified genomic biomarkers of hepatic specific injury for inflammation, cell death, and bile duct hyperplasia. We further demonstrated that the signature genes were mechanistically related to the phenotype the signature intended to predict (e.g. 17 out of top 20 genes for inflammation selected by PPEA were members of NF-kB pathway, which is a key pre-inflammatory pathway for a xenobiotic response). This is important because our preliminary results suggest that the PPEA model not largely deters the over-fitting problem, but also has the capability to elucidate mechanism(s) of drug action and/or of toxicity. Clearly, it will be critical for the further refinement, perfection, and validation of this algorithm to much more hetero-genetic datasets like cancer or other type of diseases.

1.4. Dissertation structure

Currently, whole genome microarrays are frequently used in clinical and preclinical studies that aim for diagnostic or prognostic prediction, phenotypic dissection, and mechanistic understanding. The recurring question when working with whole genome microarray data is how to handle the ubiquitous “overfitting”. Because of the uniqueness of the resulting microarray data whereas the sample size is typically far smaller than the feature size, this situation necessitates dimensionality reduction through gene selection to avoid data overfitting and improve generalization of discriminant (classifier). This

dissertation addresses the issue of overfitting primarily focused on several methodologies in feature selection.

Chapter One contextualizes the dissertation by introducing the way feature selection, concerning with problem of overfitting, and proposing solutions that are discussed throughout this work.

Chapter Two introduces the biological background, basic concept, and existing methods of molecular profiling in breast cancer and toxicogenomics.

Chapter Three describes an alternative way to derive gene signatures for breast cancer prognosis in the context of known biological pathways.

Chapter Four gives details for development of a novel algorithm called PPEA, which extends feature selection ideas from mechanistic approach to computational modeling.

Finally, Chapter Five summarizes the dissertation, and discusses possible limitations and future work.

CHAPTER TWO: LITERATURE REVIEW

2.1. History of breast cancer advancements

Over the past two decades, important medical advances in diagnosis and treatment of breast cancer, revolutionized our understanding of breast cancer (Gauthier-Villars, 1999; Nevins & Potti, 2007). These advances include mammography, surgical improvements, chemotherapy, estrogen-limiting hormone therapy, genetic testing and targeted molecular therapy.

Mammography is now the number one method of breast cancer detection. Although controversies arose from the quality of the randomized trials that evaluated the effectiveness of mammography, mammography screening are credited for raising the 5-year survival rate for localized breast cancer (that hasn't spread from its site of origin) from 80% to 98% since the 1950s (Kriege, 2004; Dershaw, 2005; Berry et al, 2005).

Introduced in the 1940s, chemotherapy can reduce tumor size before surgery, prevent recurrence afterwards and treat cancer that has metastasized, that is, spread beyond its initial location. Although it still produces side effects, including nausea, exhaustion and bone marrow toxicity, chemotherapy is much less harsh today than in years past (Hirsch, 2006).

As pharmaceutical breakthrough emerged, selective Estrogen Receptor Modifiers (SERMs), such as Nolvadex (tamoxifen), and a similar effective drug, Evista (raloxifene), fight cancers that need estrogen to grow by limiting the ability of estrogen to enter the cancer cell. In high-risk women, this class of

drugs was found to reduce recurrence and the development of invasive breast cancer by 50% when taken over a 5-year period (Osborne, 1998). Aromatase inhibitors, a class of medications that includes Arimidex (anastrozole), Aromasin (exemestane) and Femara (letrozole), work by reducing the estrogen available to cancer cells, and have been found to be more effective than tamoxifen in women who are postmenopausal and who have estrogen positive breast cancer (Mokbel, 2002).

As a standard of treatment for both early stage and advanced or metastatic disease, Herceptin (trastuzumab, Genentech) is a classic example in targeted therapy that specifically binds to a particular subtype of breast cancer that has over-expression of the HER2/neu protein on its surface. It destroys the cancer cells, but very little healthy tissue. The validated association of aggressive disease and the over-expression of HER2 in any type or stage of breast cancer have created a $4 billion global market for Herceptin in 2007. Herceptin paired with chemotherapy cuts recurrence of HER2/neu-positive breast cancer by 50% (Hudis, 2007). Still, nearly 50% of HER2-positive patients do not respond to Herceptin, and survival benefits are transient, often lasting under a year. Furthermore, side effects remain a significant problem.

Taking together, there is mounting evidence that adjuvant systemic therapy has resulted in a substantial improvement in both disease-free survival and overall survival of patients with breast cancer.

2.2. Traditionally prognostic factors

Despite significant medical advances have been made, we still lack the ability to accurately predict if an individual patient would benefit from adjuvant therapy. In fact, the majority of women receive treatment unnecessarily for the benefit of a few. The importance of discovering strong prognostic and predictive markers to identify patients at high risk for relapse and aid in the selection of the most appropriate therapy has been long recognized by both cancer researchers and clinicians. Until recently, the only validated prognostic factors for breast cancer have been clinico-pathologic features such as lymph node status, tumor size, histologic grade, proliferative index and age, while hormone receptor and HER2/neu status serve as both prognostic and predictive factors (Stadler & Come, 2009). Traditional clinical risk classification systems like the St. Gallen (Goldhirsch et al, 1998) and National Institute of Health guidelines (Eifel et al, 2000) use these clinical and histopathologic features to develop treatment recommendations for adjuvant systemic therapy. In addition, a computer-based program called Adjuvant! Online, has been developed to help health professionals make estimates of the risk of negative outcome (cancer related mortality or relapse) without systemic adjuvant therapy, estimates of the reduction of these risks afforded by therapy, and risks of side effects of the therapy. These estimates are based on clinico-pathologic information entered about individual patients and their tumors (for example, patient age, tumor size, nodal involvement, histologic grade, etc.) (Ravdin et al, 2001; Olivotto et al, 2005) However, the prognostic prediction of

Adjuvant! is merely based on limited number of prognostic factors (ER, PR and HER2 status, proliferation markers, genomic scores not included), and has limited efficacy adjustments (only for chemotherapy according to age and ER-status in post-menopausal patients) and a short time frame with calculations for 10-year outcome or even shorter (aromatase inhibitors, taxanes).

While useful in predicting outcome, the fundamental flaw for these prediction approaches is that these clinicopathologic features do not fully reflect the biologic complexity of an individual's tumor. More reliable prognostic and predictive models are needed to refine which individual patient would derive benefit from adjuvant systemic therapy.

2.3. Molecular profiling advancements

Since first introduced high-throughput gene expression microarrays used fluorescent labeling instead of radioactively labeled targets hybridized to 46 cDNA probes simultaneously in 1995 (Schena et al, 1995), rapid growth in the technology allowed millions of probes to fit on a 1.28 cm2 chip (CeneChip expression arrays from Affymetrix) enough to cover whole genome of an organism. 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 & Downward, 2001). While

various microarray systems use different chip printing processes, the two most important chip types are one-channel arrays (Affymetrix) and two-channel arrays (spotted arrays). The Affymetrix one-channel arrays are in situ synthesized oligonucleotide arrays that use single fluorescence channel to measure expression level of genes of a sample that build up oligos directly on a slide. Two-channel arrays (spotted arrays) are made by depositing pre-made oligos or cDNAs onto slides and two colors of fluorescence are used to label experiment and control samples before hybridization. Use of this technology to systemically measure gene expression on a global level has evolved from large scale gene mapping and sequencing (Poustka et al, 1986) to transcript level analysis and gene signaling pathway identification (Schena et al, 1995; Schulze et al, 2004). This technology currently has been widely applied for identifying gene expression changes that are reacting to or causing disease promises to significantly enhance our understanding of common disorders. The transcriptome represents the collection of all RNAs produced in a cell or tissue at a defined time in development. Within biological systems it provides the critical link in the flow of information between genes and disease. It enables researchers to rapidly survey thousands of differentially expressed genes, proteins, or metabolites simultaneously, thus employing an integrative approach to identify unique biomarker signatures (Lamb et al, 2006; Nevins & Potti, 2007). In cancer, for example, elucidation of the role of Kras as biomarkers of response to EGFR inhibition in the treatment of colorectal cancer exemplifies the recent remarkable achievements that have been made

based on molecular profiling (Amado et al, 2008). Identification of novel cancer-related targets, gene signatures or biomarkers, and pathways using molecular profiling is redefining our understanding of the nature of malignancy, and is prompting new, more detailed classifications of malignancy that enable the development of specific/personalized cancer therapies (Ioannidis, 2007; Geyer et al, 2009a, 2009b). Traditional classifications involving gross pathologic hallmarks such as stage and grade are now being further augmented by data detailing the over expression of oncogenes, silencing of tumor suppressors and presence of mutant forms of relevant markers. In each of these cases novel targeted agents and various combination strategies are being investigated based on the specific molecular profiles of cancer subtypes (Nevins & Potti, 2007). In breast cancer, for example, classification has been extended to include metastatic sites (bone, brain), expression of HER2, and hormone receptor status. A recently introduced molecular classification is triple negative (ER-, PGR- and HER2-) breast cancer, also known as basal-cell cancer (Perou et al, 2000). Other classifications include inflammatory breast cancer and ductal carcinoma in situ (Yu et al, 2004; Sotiriou et al, 2003, 2006). Different treatment strategies are in development for each of these tumor types. Similarly, most other malignancies are also undergoing reclassification based on specific molecular profiles (Tothill et al, 2008).

As we know, gene signature or biomarkers are measurable and quantifiable cellular characteristics that serve as indicators of normal or pathogenic biological processes. Identification of predictive signature or

biomarkers of cancer from molecular profiling will arm physicians with foreknowledge that will potentially increase efficacy of current treatments, and will provide critical information for designing improved treatment strategies (Potti et al, 2006). Eventually, it aids in the design of treatment regimens tailored specifically to individual patient, and could lead to the identification of potential targets for drug development and for evaluating the efficacy and adverse effect of lead compounds in clinical trials (Gibbs, 2000).

2.4. Development of gene signature for breast cancer prognosis

Numerous studies sought to utilize microarray technology in order to identify gene expression patterns that could be used to distinguish between patients who had the same stage of disease but different responses to treatment and hence different overall clinical outcomes (Potti et al, 2006; Lee & Havaleshko, 2007). For example, a 70-gene expression signature, often referred to as the Amsterdam signature, was developed from gene expression profiles of 117 breast tumors and was strongly predictive of a short interval to distant metastases in patients with tumors that were lymph node negative (van't Veer et al, 2002). The 70-gene signature was further validated in a follow-up study of 295 breast cancer patients (van de Vijver et al, 2002). A custom-designed array chip of the 70-gene-expression profile, known as MammaPrintTM, has recently gained FDA approved. A second large prospective clinical trial called the Microarray for Node-negative Disease may Avoid Chemotherapy (MINDACT) has being conducted in Europe. This trail

will determine if gene-expression profiling using MammaPrint™ can identify women who could be spared chemotherapy without compromising long-term disease outcome (Bogaerts et al, 2006).

Although amount evidence showed that gene-expression-based biomarkers were more powerful predictors of outcome than traditional clinical criteria, there are two major concerns among biologists and physicians regarding gene expression signatures obtained from microarray data as prognosis markers or predictors for drug responses (Massague, 2007). First, gene signatures reported by different studies have little overlap. For example, a subset of 64 genes was identified from gene expression profiling data of 159 population-derived breast cancer patients to give an optimal separation of patients with good and poor outcomes (Pawitan et al, 2005). Only three of the 64 genes were among the 70-gene prognosis signature. In another study, a 76-gene signature was developed from Affymetrix array data of 286 lymph node negative breast cancer patients for risk assessment (Wang et al, 2005). Similarly, upon comparison of this 76-gene signature with the Amsterdam 70-gene signature, only 3 genes overlapped. There are several additional prognostic models with various number of genes derived from microarray gene expression data including the intrinsic subtype model (Perou et al, 2000; Sorlie et al, 2001, 2003), the wound response model (Chang et al, 2005), the recurrence score model (Paik et al, 2004) and the two-gene-ratio model (Ma et al, 2004). The gene overlap between these models is minimal. Fan and colleagues compared five models in a single dataset and found four of the five

models to be concordant in their outcome prediction (Fan et al 2006). While this result suggested that different prognostic gene signatures may track a common set of biological characteristics, the question remains that why there is a lack of consensus gene expression models for prognosis. The van't Veer dataset, for which the 70-gene signature was derived from (van't Veer et al, 2002) was analyzed retrospectively (Ein-Dor et al, 2005). It was found that different genes can be identified as prognosis markers depending on which subset of patient samples is selected as the training dataset (Ein-Dor et al 2005), further casting the doubt on the current methodology of developing prognostic gene signatures from the whole genome transcription profiles. Second, the gene expression signatures for prognosis or drug responses are often difficult to interpret with respect to the underlying biology. Up to 30% of the signature genes have unknown function while the rest of them are associated with various unrelated biological pathways. Ultimately, finding gene signatures that can be linked to the molecular mechanisms of cancer development is critical for translating these markers into the clinic. Recent controversy in deriving gene expression patterns from microarray data to predict whether tumors will respond to chemotherapy (Coombes et al, 2007) is a reflection of these two issues.

In Chapter Three, we attempted to address these two issues by developing a novel approach to identify gene signatures for cancer prognosis in the context of known biological pathways. Our rationale for this approach was if we attempt to identify gene signatures within well defined pathways, not

only does this approach alleviate the dimensionality problem, but the mechanism-based gene signatures should also be more biologically relevant than the signatures derived from the entire human transcriptome. In preliminary studies, we first tested more than a thousand well-defined signal pathways or functional gene sets from several public or commercial available sources such as Ingenuity, GO, KEGG, Biocarta, and Iconix. Preliminary results demonstrate that several signal pathways have been potentially linked to cell cycle, DNA damage response, DNA repair capacity, mitotic checkpoints, hypoxia, and other tumor micro-environmental factors such as glucose deprivation and oxidative stresses. Motivated by the initial success, we identified the pattern of gene expression in the cell cycle pathway can indeed serve as a powerful biomarker for breast cancer prognosis. We further built a predictive model for prognosis based on the cell cycle gene signature, and found our model to be more accurate than the Amsterdam 70-gene signature when tested with multiple gene expression datasets generated from several patient populations (Liu et al, 2008). To our knowledge, this is the first gene signature that was systematically derived directly from well-defined cancer-associated signal pathways. The result suggests that the roles of this pathway and its interaction with oncogenic networks are key and potentially critical to understand and predict behaviors related to sensitivity to cell proliferation inhibitors.

2.5. Biomarker in organ toxicity

Many preclinical candidate compounds do not achieve ultimate regulatory approval because of induced organ toxicity. Because of organ toxicity, up to half of compounds discontinued in drug development are due to drug induced liver injury (DILI) including necrosis, steatosis, cholestasis, proliferation, inflammation, and bile duct hyperplasia (Ozer et al, 2008). It has been well-documented that biomarkers that identify incipient damage that leads to preclinical and clinical toxicities will enable better decision-making during drug development (Ryan et al, 2008.) Particularly valuable are translational biomarkers that bridge preclinical testing species and humans as they can expand the usefulness of the former for detection of human liabilities (Sistare & DeGeorge, 2007).

Currently, serum ALT (Alanine aminotransferase) activity level is the most frequently relied upon laboratory indicator of hepatotoxic effects (Amacher, 1998, 2002). It shows infrequent false negative signals of liver histopathological injury as well as limited false positive signals, and is considered as the gold standard clinical chemistry marker of DILI. However, it does not always correlate well with preclinical histomorphological data, although the overall clinical utility of serum ALT measurements is exceptional. Our own analysis suggested that ALT was not particularly sensitive for early minimal necrosis in rat liver (Liu et al, in preparation). Thus, additional genomic signature or biomarkers are sought to add information to serum ALT

enzymatic signals, especially as bridge biomarkers in early human trials where histopathological data are usually not available.

Although a sole biomarker is appealing as it can be easier to understand, there are few examples in preclinical testing or in clinical practice wherein a single measurement is considered definitive. Multiple markers are required to capture the biological heterogeneity of organs involved, individual variations and disease or toxicity processes (Mendrick, 2008). As described above, the technology of molecular profiling underlying pharmacogenomics and pharmacogenetics enables assessment of thousands of genes expression and single nucleotide polymorphisms (SNPs) simultaneously in each sample. Sophisticated machine learning approaches can be employed to identify meaningful biomarkers and discover the appropriate weight or influence applied to each to generate a final algorithmic conclusion.

To date, cumulated evidences indicated that molecular profiling could achieve following contribution in toxicogenomics: (i) Safety assessment could be improved with the ability to link a chemical-elicited phenotype with gene expression changes (Phenotypic Anchoring) and the potential to identify subtle markers of cellular injury that precipitate overt organ toxicity (Luo et al, 2005). (ii) Biomarkers of toxicity can be identified to monitor drug therapy for evidence of toxicity or therapeutic outcome and predict exposure levels, particularly, can be used as diagnostic tools for traditionally difficult toxico-dynamic monitoring (Euna et al, 2008). (iii) Based on the assumption of that toxicants that elicit similar pathology or disease will elicit a common pattern of gene expression

changes, molecular profiling can be utilized to facilitate high-throughput chemical toxicity (Ganter et al, 2005).

Although there have been only limited reports of success in toxicity detection, the use of gene expression signatures or biomarkers in peripheral blood cells as sentinels of tissue damage or dysfunction due to disease processes is showing great promise in many areas (Baird, 2007; Burczynski & Dorner, 2006; Deng et al, 2006; McHale et al, 2007; Mendrick et al, 2007; Mendrick, 2008). One excellent example is that, Rick Paules and collaborators recently published gene expression signatures in the peripheral blood that predict exposure to harmful levels of acetaminophen in the rat, and reported that these gene based measurements were more accurate than classical clinical pathology and histopathology assessments (Bushel et al, 2007). They translated these rat genes into their human orthologs and found they could separate acetaminophen-intoxicated patients from control humans.

2.6. Current approaches in signature or biomarker discovery

Since the first report of a DNA microarray was published in 1995 (Schena et al, 1995) and the technique became commercially available around the year 2000, various methodologies and applications have been developed and implemented. Currently, unsupervised learning and supervised learning, two basic approaches adopted from machining learning in computer science field have been described for the analysis of DNA microarray data (Golub et al, 1999).

2.6.1. Unsupervised learning

Unsupervised learning involves the discovery of intrinsic properties in a given data set without regard for prior knowledge of the underlying biology. In other words, no assumptions is made about what mechanisms might underlie for a given gene expression profile. As pioneered by Brown, Botstein and colleagues (Eisen et al, 1998), this approach became an effective tool in classifying biological samples into categories that were not previously known to exist. The power of this approach was exemplified in the work of Perou and colleagues, who have used expression patterns to define clinically significant subtypes of human breast cancer (Sorlie et al, 2003; Perou et al, 2000). Now, numerous examples have been reported in which this approach has been used to analyze gene expression data, often uncovering biological complexity that was not previously appreciated or the refinement of tumor classification. Here I only list a few examples such as hierarchical clustering in R (http://cran.r-project.org), Dchip (http://biosun1.harvard.edu/complab/dchip/) and GenePattern (http://www.broadinstitute.org/cancer/software/genepattern/). Typically, the result of unsupervised clustering is displayed in a color-coded matrix called heat map, where samples and genes are sorted according to the results of clustering. The heat map is used to represent the expression values for each gene in each sample and is the basis of many of the published microarray figures.

2.6.2. Supervised learning

By contrast, ‘supervised learning’ strategies do consider existing information and, indeed, use it to guide the analysis of the gene expression data. This approach has been particularly useful in the identification of gene expression patterns that relate to clinically relevant phenotypes such as the ability to predict the potential for recurrence of disease. Mostly lies in following two aspects, the capabilities of the supervised learning are: (i) the ability to specifically drive the analysis to the phenotype of interest, taking advantage of the relevant information as a guide; (ii) the approach to find those gene expression patterns that relate to the phenotype, when the underlying biology relevant to the phenotype is uncertain, or if the clinical outcome reflects multiple components of the subtypes defined by unsupervised analysis. As likes ‘unsupervised learning’, plentiful applications have been developed and published for this approach. One of excellent examples is Prediction Analysis of Microarrays (PAM, http://www-stat.stanford.edu/~tibs/PAM/). However, the selection of a unique list of genes by this approach does not offer sufficient knowledge to understand the biology of a given system, suggests the necessity to incorporate biological knowledge into array analysis.

2.6.3. Gene-set enrichment analysis

Gene-Set Enrichment Analysis (GSEA) is an approach that was developed to go beyond the analysis of patterns on a gene-by-gene basis (Subramanian et al, 2005). Traditionally we focus on genes at the top (up

regulated) and bottom (down-regulated) of a list ranked by some measure of statistical differences or “cutoff” in expression between phenotypes. However, a long list of statistically significant genes without any unified biological theme, and most of time no individual gene meets the threshold or “cutoff” for statistical significance and many genes show subtle differences. Also, single-gene analysis may miss important effects on pathways because biological function or phenomena is orchestrated coordinately by a set of genes in a complex network. Unlike traditional approach, GSEA uses statistical measures of enrichment of annotated gene sets within expression profiles. The value of GSEA and another, similar approach, ‘gene module map’ (Segal et al, 2004), is to attempt to examine the true context by looking at representations of gene sets that might better reflect the underlying biology. Although the ‘no-cutoff’ strategy is the key advantage of GSEA, it is a difficult task to summarize many biological aspects of a gene into one meaningful value when the biological study and genomic platform is complex (i.e. SNP). In many cases, the upstream data processing and comprehensive gene selection statistics cannot be simply avoided or replaced by GSEA (Huang et al, 2009). Some similar ways to do this are either to use commercially available software (Ingenuity Pathway Analysis: http://www.ingenuity.com/; Pathway Studio: http://www.ariadnegenomics.com/) or public online applications (DAVID, http://david.abcc.ncifcrf.gov/).

2.6.4. Connectivity map

One of fundamental challenges for molecular profiling is to make these disease-gene-drug connections. Conceptually, the expression signature as representing a distinct and well defined experimental state that can then be connected to an otherwise unrelated biological system opens the way for applications to inform and understand biological complexity (Lamb et al, 2006). The ability of an expression signature to dissect and connect two states, where the expression signature is the intermediary, is exemplified in the recent studies of Golub and colleagues, which describe a 'Connectivity Map' (Lamb et al, 2006; Hieronymus et al, 2006; Wei et al, 2006). Based on the creation of a large reference library of gene expression signature from cultured human cells perturbed with many chemicals and genetic reagents, this library of signatures is then used as a database that can be queried with expression information for other biological contexts, thereby linking otherwise disparate physiological events. The expression signature is represented as a group of gene identities, not by the actual properties of expression that are defined in the experimental setting, that create its independence of the methodology for determining expression, on other hands, differences between assay platform or methods of measuring actual expression. However, one of the major limitations of this model is that, the signatures for this model were completely built on cell cultures, and cells that grow on plastic in a laboratory are very different from tissues in a whole organism. Which means that effects modulated by specific microenvironments or that involve more than one cell

type are simply inaccessible. Furthermore, perhaps the biggest shortcoming of the current Connectivity Map resource is the complexity of the core dataset (sample space). Both the number and diversity of the small molecules in the collection is extremely low, meaning that the fraction of all possible induced cellular states represented is probably quite small.

2.6.5. Feature selection

The three characteristics of microarray datasets in molecular profiling: noise, large number of genes and relatively small number of samples, make over-fitting a ubiquitous danger for any tasks of model building and selection in molecular profiling (Donoho, 2000; Dudoit et al, 2002). As many pattern recognition techniques were originally not designed to cope with large amounts of irrelevant features, combining them with feature selection techniques has become a necessity in many applications (Guyon & Elisseeff, 2003; Liu et al, 2002; Wang et al, 2005). The objectives of feature selection are manifold, the most important ones can be: (i) to avoid over-fitting and defying the curse of dimensionality to improve prediction performance, i.e. prediction performance in the case of supervised classification and better cluster detection in the case of clustering, (ii) to provide faster and more cost-effective models and (iii) to gain a deeper insight into the underlying processes that generated the data. However, the advantages of feature selection techniques come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity and bias in the modeling

task. Also, there is no guarantee that the optimal parameters for the full feature set are equally optimal for the optimal feature subset.

Since Kira and Rendell (1992) first described a statistical feature selection algorithm called RELIEF that uses instance based learning to assign a relevance weight to each feature, plentiful algorithms and methods have been developed. The methods given here are a selection from many others possible, and mostly based on the chronological literature review. In order to address the problem of irrelevant features and the subset selection problem, John et al (1994) suggested that features selected should depend not only on the features and the target concept, but also on the induction algorithm. Further, they claim that the filter model approach to subset selection should be replaced with the wrapper model. In the same year, Pudil et al (1994) presented a sequential search method, called 'floating', characterized by a dynamically changing number of features included or eliminated at each step. They claimed that this method was computationally more effective than the branch and bound method. Two years later, Koller and Sahami (1996) developed a method for feature subset selection based on information theory: they presented a theoretically justified model for optimal feature selection based on using cross-entropy to minimize the amount of predictive information lost during feature elimination. In 1997, Jain and Zongker considered various feature subset selection algorithms and found that the sequential forward floating selection algorithm, proposed by Pudil et al (1994), dominated the other algorithms tested. Yang and Pedersen (1997) evaluated document

frequency, information gain, mutual information, a Chi-test and term strength, and found information gain and Chi-test to be the most effective. Kohavi and John (1997) introduced wrappers for feature subset selection. Their approach searches for an optimal feature subset tailored to a particular learning algorithm and a particular training set. Yang and Honavar (1998) used a genetic algorithm for feature subset selection.

After the microarray technology had been introduced in 1995, and commercialized around 2000, the motivation for applying feature selection techniques in bioinformatics area has shifted from being an illustrative example to becoming a real prerequisite for gene expression signature development or biomarker discovery. In particular, the small sample size and high dimensional nature of modeling tasks has given rise to a wealth of feature selection techniques being used in many microarray data analyses in both supervised learning (i.e., classification) and unsupervised learning (i.e., clustering) contexts. The obvious need for these dimension reduction methods was realized (Golub et al, 1999; Alon et al, 1999; Ben-Dor et al, 2000; Ross et al, 2000), and soon their application became a de facto standard in the field. Whereas in 2001, the field of microarray analysis was still claimed to be in its infancy, a considerable and valuable effort has since been done to contribute new and adapt known feature selection methodologies (Efron & Tibshirani, 2002).

Currently, three kinds of methods have generally been studied and applied in microarray domain: filter methods (Golub et al, 1999 in early

classification), wrapper methods (Maldonadoa & Weber, 2009), and embedded methods (Guyon et al, 2002). See Saeys et al (2007) for an excellent review on this subject.

The filter model relies on general characteristics of the training data to select predictive features (i.e., features highly correlated to the target class) without any learning algorithm involved. Mostly due to that the output provided by univariate filter feature rankings is intuitive and easy to understand, and the result can be validated by laboratory techniques, the prevalence of these univariate filter techniques has dominated the field (Dudoit et al, 2002; Lee et al, 2005). Although widely accepted and achieved substantial success for its behaviors of fast, scalable, and independent of the classifier, the filter techniques has its limitation in which the feature dependencies and interaction with classifier have been completely ignored.

Conversely, the wrapper model uses the predictive accuracy of a predetermined learning algorithm to give the quality of a selected feature subset, generally producing features better suited to the classification task at hand. Wrapper algorithms use the interactions between feature selection and the learning algorithm by involving the learning algorithm in the feature selection step. A characteristic significance for wrapper methods offers a way to perform a multivariate gene subset selection, incorporating the classifier's preference or bias into the search and thus offering an opportunity to construct more accurate classifiers (Blanco et al, 2004; Jirapech-Umpai & Aitken, 2005; Inza et al, 2004; Li et al, 2004). An interesting hybrid filter-wrapper approach is

introduced recently (Ruiza et al, 2005), which crossing a univariately preordered gene ranking with an incrementally augmenting wrapper method. The disadvantages for this method are computationally intensive, classifier dependent selection, and particularly, high risk of over-fitting.

The embedded method allows the classifier to have the capacity of discarding input unneeded features and thus propose a subset of discriminative genes. Examples include the use of random forests (a classifier that combines many single decision trees) in an embedded way to calculate the importance of each gene (Dı´az-Uriarte & Alvarez de Andre´ s, 2006; Jiang et al, 2004). Another type of embedded feature selection techniques uses the weights of each feature in linear classifiers, such as SVMs (Guyon et al, 2002) and logistic regression (Ma & Huang, 2005). These weights are used to reflect the relevance of each gene in a multivariate way, and thus allow for the removal of genes with very small weights. Partially due to the higher computational complexity and classifier dependent selection of wrapper and to a lesser degree embedded approaches, these techniques have not received as much interest as filter proposals.

To date, no single recommendation in literature is given for methods in either the feature selection or the classification of microarray data (Guyon & Elisseeff, 2003). Each of these techniques has its merits. None of them is superior to others for all microarray data sets. The particular test used depends on the data set under study.

In Chapter Four, I will describe that a new algorithm we developed that try to take the advantage in both filter (variable ranking) and wrapper methods to estimate predictive power of each individual gene. The algorithm iteratively applies a two-way bootstrapping to enforce the sample size larger than the feature size for each subset whereas the predictive power of individual gene is evaluated with an approach called prediction analysis of microarray, PAM (Tibshirani et al, 2002). We showed that the relative predictive power of genes stabilized after a definite number of iterations, which makes it possible to construct a predictive model from a much smaller (the number of genes < the number of samples) set of genes with the highest predictive power. Using DrugMatrix™ rat liver data, we identified genomic biomarkers of hepatic specific injury for inflammation, cell death, and bile duct hyperplasia. We further demonstrated that the inflammation genes selected using PPEA were mechanistically related to the NF-kB pathway and bile duct hyperplasia (BDH) genes related to the oncogenic p53 and ERBB2 pathways. More importantly, our models achieved high sensitivity and specificity when tested with completely independent datasets generated either in Lilly or in external research labs. Top four genes of BDH signature has been successfully implemented for a BDH liability assay with QPCR. Thus, we believe that the PPEA model may partially overcomes the over-fitting problem, and can be used to facilitate genomic biomarker discovery and development for predictive toxicology and to elucidate mechanism(s) of drug action and/or of toxicity.

2.7. References

2.7.1. My publications cited

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK (2006) Intrinsic disorder in transcription factors. Biochemistry 45:6873-6888.

Liu J,Campen A, Huang S, Peng SB, Ye X, Palakal M, Dunker AK, Xia U and Shuyu Li (2008) Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. BMC Medical Genomics 1:39

Liu J, Jolly RA, Thomas CE, Stevens JL, Ryan TP, Watson DE, Searfoss GH, Goldstein KM, Dunker AK, Li D, and Wei T (2010) Identification of Toxicogenomic Biomarkers by Development and Application of a New Mining Algorithm to the DrugMatrix™ Database, In preparation

Liu J, Jolly RA, Thomas CE, Dunker AK, Li D and Wei T (2010) Relating ALT to necrosis in rat liver, in preparation

2.7.2. Other literature cited

Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD (2008) Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol, 26:1626-34.

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A. 96(12):6745-50.

Amacher DE (1998) Serum transaminase elevations as indicators of hepatic injury following the administration of drugs. Regul Toxicol Pharmacol. 27(2):119-30.

Amacher DE (2002) A toxicologist's guide to biomarkers of hepatic response. Hum Exp Toxicol. 21(5):253-62.

Baird AE (2007). Blood genomics in human stroke. Stroke 38:694-698.

Burczynski ME & Dorner AJ (2006) Transcriptional profiling of peripheral blood cells in clinical pharmacogenomic studies. Pharmacogenomics 7:187-202.

Blanco R, Larranaga P, Inza I, Sierra B (2004) Gene selection for cancer classification using wrapper approaches. Int. J. Pattern Recognit. Artif. Intell., 18:1373-1390.

Ben-Dor A, Bruhn L, Friedman N, Nachman I (2000) Tissue classification with gene expression profiles. J Comput Biol. 7(3-4):559-83.

Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clarke L, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JD, Feuer EJ (2005) Cancer Intervention and Surveillance Modeling Network (CISNET) Collaborators. Effect of screening and adjuvant therapy on mortality from breast cancer. N Engl J Med. 353 (17):1784-92.

Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, Harrison JA, Bines J, Mook S, Decker N, Ravdin P, Therasse P, Rutgers E, van 't Veer LJ, Piccart M; TRANSBIG consortium. (2006) Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 3:540-51.

Bushel PR, Heinloth AN, Li J, Huang L, Chou JW, Boorman GA, Malarkey DE, Houle CD, Ward SM, Wilson RE, Fannin RD, Russo MW, Watkins PB, Tennant RW, Paules RS (2007) Blood gene expression signatures predict exposure levels. Proc Natl Acad Sci U S A. 104:18211-6

Cantor CR, Mirzabekov A, Southern E (1992) Report on the sequencing by hybridization workshop. Genomics. 13(4):1378-83.

Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, Rijn M van de, Brown PO, Vijver MJ van de (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 102(10):3738-3743.

Coombes KR, Wang J, Baggerly KA (2007) Microarrays: retracing steps. Nat Med 13(11):1276-1277.

Deng MC, Eisen HJ, Mehra MR, Billingham M, Marboe CC, Berry G, Kobashigawa J, Johnson FL, Starling RC, Murali S, Pauly DF, Baron H, Wohlgemuth JG, Woodward RN, Klingler TM, Walther D, Lal PG, Rosenberg S, Hunt S (2006).Noninvasive discrimination of rejection in cardiac allograft recipients using gene expression profiling. Am. J. Transplant. 6:150-160.

Dı´az-Uriarte R & Alvarez de Andre´s S (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:3.

Dershaw D (2005) Film or Digital Mammographic Screening? New England Journal of Medicine. 353:1846-1847.

Donoho, DL (2000) High-dimensional date analysis: The curses and blessings of dimensionality. Available at http://www-stat.stanford.edu/~donoho/.

Dudoit S, Fridlyand J, Speed, T (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc., 97:77-87.

Eifel P, Axelson JA, Costa J, Crowley J, Curran WJ Jr, Deshler A, Fulton S, Hendricks CB, Kemeny M, Kornblith AB, Louis TA, Markman M, Mayer R, Roter D (2001) National institutes of health consensus development conference statement: adjuvant therapy for breast cancer, November 1-3, 2000. J Natl Cancer Inst 93:979-989.

Efron B, Tibshirani R (2002) Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol. 23(1):70-86.

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95:14863-14868.

Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21(2):171-178.

Euna JW, Ryua SY, Noha JH,Leeb,MJ,Jangc JJ, Ryud JC, Junga KH, Kima JK,Baea HJ, Xiea H, Kima SY,Leea SH, Parka WS, Yooa NJ, Leea JY, Nama SW (2008) Discriminating the molecular basis of hepatotoxicity using the large-scale characteristic molecular signatures of toxicants by expression profiling analysis. Toxicology 249:176-183

Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM: (2006) Concordance among gene-expression based predictors for breast cancer. N Engl J Med 355(6):560-569.

FANTOM Consortium (2009) The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. Nat Genet. 41(5):553-62.

Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, Nguyen P, Nicholson SM, Pham H, Roter AH, Sun D, Tan S, Thode S, Tolley AM, Vladimirova A, Yang J, Zhou Z, Jarnagin K (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol. 119(3):219-44.

Gauthier-Villars, Marion (1999) Genetic Testing for Breast Cancer Predisposition. Surgical Clinics of North America 79:1171-1187.

Geyer FC, Marchiò C, Reis-Filho JS (2009a) The role of molecular analysis in breast cancer. Pathology, 41:77-88

Geyer FC, & Reis-Filho JS (2009b) Microarray-based Gene Expression Profiling as a Clinical Tool for Breast Cancer Management: Are We There Yet? International Journal of Surgical Pathology, 17:285-302

Gibbs JB (2000) Mechanism-Based Target Identification and Drug Discovery in Cancer Research, Science, 287 (17):1969-1973

Goldhirsch A, Glick JH, Gelber RD, Senn HJ (1998) Meeting highlights: international consensus panel on the treatment of primary breast cancer. J Natl Cancer Inst 90:1601-1608.

Golub TR, Slonim DK, Tamayo P, Huard C, et al (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286:531-537.

Guyon I,Weston J, Barnhill S, V Vapnik (2002) Gene selection for cancer classification using support vector machines. Mach. Learn., 46:389-422.

Guyon I & Elisseeff A (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3:1157-1182

Hieronymus H, Lamb J, Ross KN, Peng XP, Clement C, Rodina A, Nieto M, Du J, Stegmaier K, Raj SM, Maloney KN, Clardy J, Hahn WC, Chiosis G, Golub TR (2006) Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. Cancer Cell. 10(5):349-51.

Hirsch J (2006) An anniversary for cancer chemotherapy. JAMA 296 (12):1518-20

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, NAR, 37:1-13

Hudis CA (2007) Trastuzumab--mechanism of action and use in clinical practice. N Engl J Med. 357(1):39-51.

Inza I, Larrañaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med. 31(2):91-103.

Ioannidis, J P (2007) Is Molecular Profiling Ready for Use in Clinical Decision Making? The Oncologist 12:301-311.

Jiang & Simon (2007) A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Stat Med 26:5320-34

Jirapech-Umpai T & Aitken S (2005) Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6:148.

John GH, Kohavi R, K Peleger (1994) Irrelevant features and the subset selection problem. In: William W. COHEN & Haym HIRSH, eds. Machine Learning: Proceedings of the Eleventh International Conference. San Francisco, CA: Morgan Kaufmann Publishers, pp.121-129.

Kira K & Rendell LA (1992) A practical approach to feature selection. In: Derek H. SLEEMAN & Peter EDWARDS, eds. ML92: Proceedings of the Ninth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp.249-256.

Kriege M (2004) Efficacy of MRI and Mammography for Breast Cancer Screening in Women with a Familial or Genetic Predisposition. New England Journal of Medicine 351:427-437

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease, Science 313:1929-1935

Lee JW, Lee JB, M Park, SH Song (2005) An extensive comparison of recent classification tools applied to microarray data. Comput. Stat. and Data Anal., 48:869-885.

Lee JK & DM Havaleshko (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proc Natl Acad Sci U S A 104(32):13086-91.

Li J, Zhu X, JY Chen (2009) Building Disease-specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. PLoS Computational Biology, 5(7):e1000450.

Li T, Zhang C, Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics. 20(15):2429-37.

Liu H, Li J, Wong L (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform. 13:51-60.

Luo W, Fan W,Xie H, Jing L, Ricicki E, Vouros P, Zhao LP, Zarbl H (2005) Phenotypic Anchoring of Global Gene Expression Profiles Induced by N-Hydroxy-4-acetylaminobiphenyl and Benzo(a)pyrene Diol Epoxide Reveals Correlations between Expression Profiles and Mechanism of Toxicity. Chem. Res. Toxicol. 18:619-629

Kohavi R & John G (1997) Wrapper for feature subset selection. Artificial Intelligence, 97:273-324

Ma S & Huang J (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics, 21:4356-4362.

Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC: (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 5:607-616.

Maldonadoa S & Weber R (2009). A wrapper method for feature selection using Support Vector Machines. Information Sciences 179 (13):2208-2217

Massague, J. (2007). Sorting out breast-cancer gene signatures. N Engl J Med 356(3):294-7.

McHale CM, Zhang L, Hubbard AE, Zhao X, Baccarelli A, Pesatori AC, Smith MT, Landi MT (2007) Microarray analysis of gene expression in peripheral blood mononuclear cells from dioxin-exposed human subjects. Toxicology 229:101-113.

Mendrick DL & Daniels KK (2007) From the bench to the clinic and back again: translational biomarker discovery using in silico mining of pharmacogenomic data. Biomarkers Med. 1:319-333.

Mendrick DL (2008) Genomic and genetic biomarkers of toxicity, Toxicology 245:175-181

Mokbel K (2002) The evolving role of aromatase inhibitors in breast cancer. Int J Clin Oncol 7 (5):279-83.

Nevins JR & Potti A (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. Nat. Rev. Genetics. 8:601-609

Olivotto IA, Bajdik CD, Ravdin PM, Speers CH, Coldman AJ, Norris BD, Davis GJ, Chia SK, Gelmon KA (2005) Population-based validation of the prognostic model ADJUVANT! for early breast cancer. J Clin Oncol. 23:2716-25.

Osborne CK (1998) Tamoxifen in the Treatment of Breast Cancer. New England Journal of Medicine. 339:1609-1618.

Ozer J, Ratner M, Shaw M, Bailey W, Schomaker S (2008) The current state of serum biomarkers of hepatotoxicity. 1: Toxicology. 245(3):194-205.

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N: (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817-2826.

Pawitan Y, Bjöhle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Holmberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedrén S, Bergh J (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7(6):R953-64.

Perou CM, Sorlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: (2000) Molecular portraits of human breast tumours. Nature 406:747-752.

Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR (2006) Genomic signatures to guide the use of chemotherapeutics. Nature Medicine, 12(11):1294-1300

Potti,A. & J. R. Nevins,(2008) Utilization of genomic signatures to direct use of primary chemotherapy. Current Opinion in Genetics & Development 18:62-67

Poustka A, Pohl T, Barlow DP, Zehetner G, Craig A, Michiels F, Ehrich E, Frischauf AM, Lehrach H (1986) Molecular approaches to mammalian genetics. Cold Spring Harb Symp Quant Biol. 51 Pt 1:131-9.

Pudil P, Novovicova J, J Kittler (1994) Floating search methods in feature selection. Pattern Recognition Letters, 15(11):1119-1125.

Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N, Parker HL(2001) Computer program to assist in making decisions about adjuvant therapy forwomen with early breast cancer. J Clin Oncol 19:980-91.

Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat. Genet., 24:227-235.

Ruiza R, JC Riquelmea, JS Aguilar-Ruizb (2005) Incremental wrapper-based gene selection from microarray data for cancer classification. Pattern Recognit., 39:2383-2392.

Ryan T P, Stevens JL, Thomas CE (2008) Strategic applications of toxicogenomics in early drug discovery. Current Opinion in Pharmacology 8:1-7

Saeys Y, Inza I, Larra˜naga P (2007) A review of feature selection techniques in bioinformatics, Bioinformatics, 23:2507-2517.

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

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

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

Sistare FD & DeGeorge JJ (2007) Preclinical predictors of clinical safety: opportunities for improvement. Clin. Pharmacol. Ther. 82:210-214.

Smith GJ, DeLuca CM, Yong LC (1984) A model of bile duct hyperplasia in the rat induced by diethylnitrosamine and selective cytotoxicity. Pathology.16(4):396-400.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98(19):10869-10874.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA, 100(14):8418-8423.

Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris AL, Liu ET (2003) Breast cancer classifi cation and prognosis based on gene expression profi les from a population-based study. Proc Natl Acad Sci U S A 100:10393-8.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, Desmedt C, Larsimont D, Cardoso F, Peterse H, Nuyten D, Buyse M, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M (2006) Gene Expression Profi ling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. J Natl Cancer Inst 98:262-72

Stadler ZK & Come SE (2009) Review of gene-expression profiling and its clinical use in breast cancer, Critical Reviews in Oncology/Hematology 69:1-11

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA, 99(10):6567-6572.

Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, et al (2008) Novel Molecular Subtypes of Serous and Endometrioid Ovarian. Clin Cancer Res. 14(16):5198-5208.

van't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530-536.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 347(25):1999-2009.

Wang H, Parrish A., Smith RK, Vrbsky S (2005) Improved variable and value ranking techniques for mining categorical traffic accident data, Expert Systems with Applications 29 (4):795-806.

Wei G, Twomey D, Lamb J, Schlis K, Agarwal J, Stam RW, Opferman JT, Sallan SE, den Boer ML, Pieters R, Golub TR, Armstrong SA (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. Cancer Cell. 10(5):349-51.

Yang Y & JO Pedersen (1997) A comparative study of feature selection in text categorization. In: ICML ’97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp.412-420.

Yu K, Lee CH, Tan PH, Tan P (2004) Conservation of Breast Cancer Molecular Subtypes and Transcriptional Patterns of Tumor Progression Across Distinct Ethnic Populations. Clinical Cancer Research, 10:5508-5517.

CHAPTER THREE: IDENTIFICATION OF A GENE SIGNATURE IN CELL CYCLE PATHWAY FOR BREAST CANCER PROGNOSIS USING GENE EXPRESSION PROFILING DATA (Paper I, Liu et al 2008)

3.1. Abstract

Background: Numerous studies have used microarrays to identify gene signatures for predicting cancer patient clinical outcome and responses to chemotherapy. However, the potential impact of gene expression profiling in cancer diagnosis, prognosis and development of personalized treatment may not be fully exploited due to the lack of consensus gene signatures and poor understanding of the underlying molecular mechanisms.

Methods: We developed a novel approach to derive gene signatures for breast cancer prognosis in the context of known biological pathways. Using unsupervised methods, cancer patients were separated into distinct groups based on gene expression patterns in one of the following pathways: apoptosis, cell cycle, angiogenesis, metastasis, p53, DNA repair, and several receptor-mediated signaling pathways including chemokines, EGF, FGF, HIF, MAP kinase, JAK and NF-κB. The survival probabilities were then compared between the patient groups to determine if differential gene expression in a specific pathway is correlated with differential survival.

Results: Our results revealed expression of cell cycle genes is strongly predictive of breast cancer outcomes. We further confirmed this observation by building a cell cycle gene signature model using supervised methods.

Validated in multiple independent datasets, the cell cycle gene signature is a more accurate predictor for breast cancer clinical outcome than the previously identified Amsterdam 70-gene signature that has been developed into a FDA approved clinical test MammaPrint®.

Conclusion: Taken together, the gene expression signature model we developed from well defined pathways is not only a consistently powerful prognosticator but also mechanistically linked to cancer biology. Our approach provides an alternative to the current methodology of identifying gene expression markers for cancer prognosis and drug responses using the whole genome gene expression data.

3.2. Background

DNA microarray technology has created a new paradigm for understanding cancer biology by simultaneous meas-urement of tens of thousands of genes in malignant or normal cells. Gene expression profiles have been utilized to identify gene signatures for cancer diagnosis and prognosis (Quackenbush, 2006). Motivated by the lack of accurate outcome prediction with the best clinical predictors of metastasis including lymph-node status and histological grade, numerous studies sought to utilize microarray technology in order to identify gene expression patterns that could be used to distinguish between patients who had the same stage of disease but different responses to treatment and hence different overall clinical outcomes. For example, a 70-gene expression signature, often referred to as the Amsterdam

signature, was developed from gene expres-sion profiles of 117 breast tumors and was strongly predictive of a short interval to distant metastases in patients with tumors that were lymph node negative (van't Veer et al, 2002). The 70-gene signature was further validated in a follow-up study of 295 breast cancer patients (Vijer et al, 2002). These studies showed that gene-expression-based biomarkers were more powerful predictors of outcome than traditional clinical criteria. Recently, microarray-based gene expression signatures have also been developed to predict patient responses to therapeutic agents (Lee et al, 2007; Potti et al, 2006).However, there are two major concerns among biologists and physicians regarding gene expression signatures obtained from microarray data as prognosis markers or predictors for drug responses (Massague, 2007). First, gene signatures reported by different studies have little overlap. For exam-ple, a subset of 64 genes was identified from gene expression profiling data of 159 population-derived breast cancer patients to give an optimal separation of patients with good and poor outcomes (Pawitan et al, 2005). Only three of the 64 genes were among the 70-gene prognosis signature (van't Veer et al, 2002). In another study, a 76-gene signature was developed from Affymetrix array data of 286 lymph node negative breast cancer patients for risk assessment (Wang et al, 2005). Similarly, upon comparison of this 76-gene signature with the Amsterdam 70-gene signature, only 3 genes overlapped. There are several additional prognostic models with various numbers of genes derived from microarray gene expression data including the intrinsic subtype model (Perou et al, 2000;

Sorlie et al, 2001, 2003), the wound response model (Chang et al, 2005), the recurrence score model (Paik et al, 2004) and the two-gene-ratio model (Ma et al, 2004). The gene overlap between these models is minimal. Fan and colleagues compared five models in a single dataset and found four of the five models to be concordant in their outcome prediction (Fan et al, 2006). While this result suggested that different prognostic gene signatures may track a common set of biological characteristics, the question remains that why there is a lack of consensus gene expression models for prognosis. The van't Veer dataset, for which the 70-gene signature was derived from (van't Veer et al, 2002), was analyzed retrospectively (Ein-Dor et al, 2005). It was found that different genes can be identified as prognosis markers depending on which subset of patient samples is selected as the training dataset (Ein-Dor et al, 2005), further casting the doubt on the current methodology of developing prognostic gene signatures from the whole genome tran-scription profiles. Second, the gene expression signatures for prognosis or drug responses are often difficult to inter-pret with respect to the underlying biology. Up to 30% of the signature genes have unknown function while the rest of them are associated with various unrelated biological pathways. Ultimately, finding gene signatures that can be linked to the molecular mechanisms of cancer development is critical for translating these markers into the clinic. Recent controversy in deriving gene expression patterns from microarray data to predict whether tumors will respond to chemotherapy (Coombes et al, 2007) is a reflection of these two issues.

In this report, we attempted to address the above mentioned two issues by developing a novel approach to identify gene signatures for cancer prognosis in the context of known biological pathways. Due to the nature of high dimensional data spaces in microarray studies where the number of measurements (> 10,000 mRNA transcripts) is greatly higher than the number of samples, data overfitting is an inevitable issue (Clarke et al, 2008). Therefore, our rationale was if we attempt to identify gene signatures within well defined pathways, not only does this approach alleviate the dimensionality problem, but the mechanism-based gene signatures should also be more biologically relevant than the signatures derived from the entire human transcriptome. Unsupervised hierarchical clustering analysis was first used to divide cancer patients into separate groups based on expression patterns of genes in a known pathway. Patient survival in the different groups was then compared. If a specific pathway plays a critical role in tumor progression and metastasis, patients with distinct gene expression patterns in the pathway may have very different clinical outcomes. The results presented here indicate that the pattern of gene expression in the cell cycle pathway can indeed serve as a powerful biomarker for breast cancer prognosis. We further built a predictive model for prognosis based on the cell cycle gene signature and found our model to be more accurate than the Amsterdam 70-gene signature when tested with multiple gene expression datasets generated from several patient populations.

3.3. Methods

3.3.1. Data source

Five different gene expression profiling datasets on breast cancers were analyzed in this study. Multiple datasets were used to demonstrate repeatability of the analysis. Specific details on each dataset are summarized in Table 1. For each gene expression dataset, 20 molecular pathways were analyzed. The 20 pathways were assembled from the Ingenuity Pathway databases http://www.ingenuity.com/ and the SuperArray cancer pathway array annotations http:// www.superarray.com/home.php. The list of 20 pathways and genes within each pathway are provided in additional files (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2551605/bin/1755-8794-1-39-S1.doc).

3.3.2. Data preprocessing

For each array study based on Affymetrix oligonucleotide platforms, we downloaded the .CEL files and generated gene expression values using the Affymetrix MAS5 algo-rithm with trimmed mean values normalized to 500. A trimmed mean is the average value after removing the lowest 2% and the highest 2% of all expression values on the array. Prior to analysis, each data set was preprocessed with a log2 transformation and subsequently expression of each gene was standardized using median-centering. Data transformation and standardization were performed using scripts written in the R statistical programming language. When a gene is represented by multiple probe sets on

Affymetrix oligonucleotide arrays, the average expression value was used for further analysis.

Table 1: Breast cancer gene expression profiling datasets analyzed in this study

Reference	Study summary	Sample Size	Microarray platform	Data download	How dataset was used in this study
Van de Vijver et al.	Demonstrated that a 70- gene expression signature is a more powerful predictor for outcome than standard clinical and histological criteria in 295 primary breast cancer patients	295	Inkjet Oligo	http://www.rii.com/publications/2002/nejm.html	Initial unsupervised analysis to identify outcome associated pathways
Wang et al.	Developed a 76-gene 286 signature to predict distant metastasis using gene expression profiling data in 286 node negative primary breast cancer tumors	286	U133A	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2034	Initial unsupervised analysis to identify outcome associated pathways; Training dataset to build prognostic gene signature models.
Miller et al.	Identified a 32-gene signature from 251 primary breast cancers to distinguish p53-mutant and wild-type tumors and to predict prognosis.	251	U133A	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3494	Initial unsupervised analysis to identify outcome associated pathways; Independent dataset for validating the prognostic gene signature models.
Pawitan et al.	Identified a subset of 64 genes from gene expression profiles in 159 primary breast cancers that give an optimal separation of good and poor outcomes.	159	U133A	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1456	Initial unsupervised analysis to identify outcome associated pathways; Independent dataset for validating the prognostic gene signature models.
Bild et al.	Developed gene expression signatures for oncogenic pathways and demonstrated these signatures are predictive of clinical outcomes in lung, breast and ovarian cancers.	171	U95Av2	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3143	Initial unsupervised analysis to identify outcome associated pathways.

3.3.3. Hierarchical clustering

Each pathway specific data set was analyzed by hierarchical average-linkage clustering. The clustering was performed using Gene Cluster 3.0 http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/ or using R pro-grams. The resulting numerical output was used by Java Treeview v1.1 http://jtreeview.sourceforge.net/ to generate the associated heatmaps and clustering dendrograms.

3.3.4. Kaplan-Meier survival analysis

In addition to gene expression data, clinical information for each primary tumor sample is provided by the authors in each array study we analyzed (Table 1). The clinical data included survival and/or relapse time and censoring status. Using the available clinical outcome data, Kaplan-Meier analysis was performed on the patient groups defined by the hierarchical clustering analysis. An outcome curve for each cluster was produced using GraphPad Prism 4. The associated p-values generated from log-rank test in Kaplan-Meier analysis was used to represent the statistical significance of differential survival probabilities between the two patient groups.

3.3.5. Supervised learning analysis

The PAM (Prediction Analysis for Microarray) algorithm (Tibshirani et al, 2002) was used as the classification method. The analysis was implemented in the R programming language. A 10-fold cross validation was used by

dividing the training dataset into 10 approximately equal-sized groups. The model was fitted on the 90% of the samples and tested on the remaining 10%. The procedure was repeated 10 times so each of the 10 groups was used as the testing samples and contributed to the overall error rate. The amount of shrinkage was chosen to minimize the error rate.

3.4. Results

3.4.1. Gene expression profiling datasets and the analyzed pathways

Although there are dozens of breast cancer microarray studies, the available datasets that we could utilize in our study are limited. First, to ensure statistical power, we selected datasets with at least 100 patient samples. In addition, both gene expression data and patient clinical data such as survival time and status needed to be available. To obviate fundamental difference inherent in different array platforms, we focused mainly on gene expression data based on Affymetrix oligonucleotide arrays, particularly more advanced platforms such as U95Av2 or U133 series. We also included the 295-sample dataset that served as the basis for the development and validation of the original Amsterdam 70-gene prognostic signature (Vijer et al, 2002). As indicated in Table 1, five datasets on primary breast tumors were analyzed.

The datasets in Table 1 were analyzed using 20 molecular pathways that were compiled from Ingenuity Pathway databases http://www.ingenuity.com/ and the SuperArray cancer pathway array

Table 2: Gene expression in specific pathways as prognosis markers

Pathways	Van de Vijver	Wang	Miller	Pawitan	Bild
The 70-gene signature	5.1E-07*	0.0059*	0.00020*	0.00049*	0.038*
Angiogenesis	0.069	0.3	0.12	0.0023*	0.711
Apoptosis	0.5	0.23	0.0017*	0.19	0.055
Breast cancer	3.2E-08*	0.0035*	2.4E-04*	4.7E-05*	0.050*
Chemokines	0.16	0.28	0.064	0.00045*	0.64
Cell Cycle	9.9E-09*	0.0035*	0.0017*	9.5E-05*	0.037*
DNA damage	2.2E-05*	0.055	0.036*	0.0062*	0.2
EGF	3.5E-06*	0.25	0.0049*	0.00099*	0.013*
FGF	4.9E-06*	0.033*	0.0047*	2.1E-06*	0.14
G1_S	0.0014*	0.00098*	0.0037*	0.0027*	0.21
G2_M	0.1	0.08	3.5E-04*	0.016*	0.19
HIF	0.0035*	0.030*	0.19	0.44	0.011*
JAK	0.67	0.37	0.061	0.084	0.029*
MAPK	0.0069*	0.94	0.0059*	0.25	0.76
Metastasis	0.35	0.015*	2.9E-04*	0.00037*	0.44
NER	0.92	0.8	0.27	0.16	0.64
NF-κB	0.88	0.91	0.49	0.47	0.11
p38	0.078	0.35	0.84	0.054	0.077
p53	9.2E-06*	0.066	0.0065*	5.9E-06*	0.013*
DNA Repair	1.7E-08*	0.0076*	0.047*	0.023*	0.22
Cell surface signaling	0.045*	0.13	0.025*	4.9E-05*	0.55

The numbers represent the log-rank test P values in Kaplan-Meier analysis in two patient groups defined by hierarchical clustering. *P < 0.05.

annotations http://www.superar ray.com/home.php. These pathways are involved in cancer development by directly regulating angiogenesis or metastasis processes, by regulating cell cycle, apoptosis, DNA repair, or by mediating cell signaling events (Table 2). The genes in each pathway were assembled manually from literature information as of February 2007. In

addition, we included the Amsterdam 70-gene signature as a control in our analysis. We also included a breast cancer gene set that contains 264 genes as known molecular markers in the prognosis and diagnosis of breast cancer. These genes were derived from literature as well as from previous microarray studies (van't Veer, 2002; Vijver et al, 2002;Paik et al, 2004; Hu et al, 2003). The 70-gene signature is a subset of the 264 breast cancer gene model. Listed in additional file 1 are the pathway names and genes associated with each pathway (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2551605/bin/1755-8794-1-39-S1.doc).

3.4.2. Overall analysis strategy

Illustrated in Figure 1 is a flow chart describing the overall analysis. For each dataset, we first extracted expression data of genes involved in a specific pathway, followed by an unsupervised two-way hierarchical clustering analysis. If the hierarchical clustering analysis resulted in several distinct patient groups, then patient outcome in these distinct groups were compared using the Kaplan-Meier analysis. Our rationale is that if a specific pathway plays a critical role in tumor progression and metastasis, patients with distinct gene expression patterns in the pathway may have very different clinical outcome. This process was repeated for each of the 20 pathways we assembled.

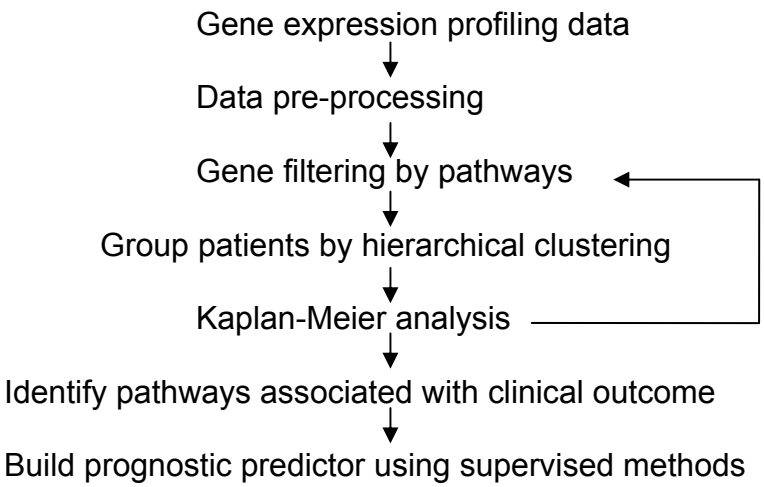


Figure 1: Analysis strategy. Hierarchical clustering using gene expression in specific pathways is followed by Kaplan-Meier survival analysis. The pathways exhibiting strong correlation between gene expression and clinical outcome were further examined using supervised methods to build predict models.

The five datasets in Table 1 were analyzed as demonstrated in Figure 1 for the 20 pathways. For each hierarchical clustering, cancer patients were separated into two distinct groups that Kaplan-Meier analysis was applied to. Summarized in Table 2 are the log-rank test P values of the Kaplan-Meier survival analysis. A P-value of less than 0.05 suggests that the two patient clusters have significantly differential survival probabilities.

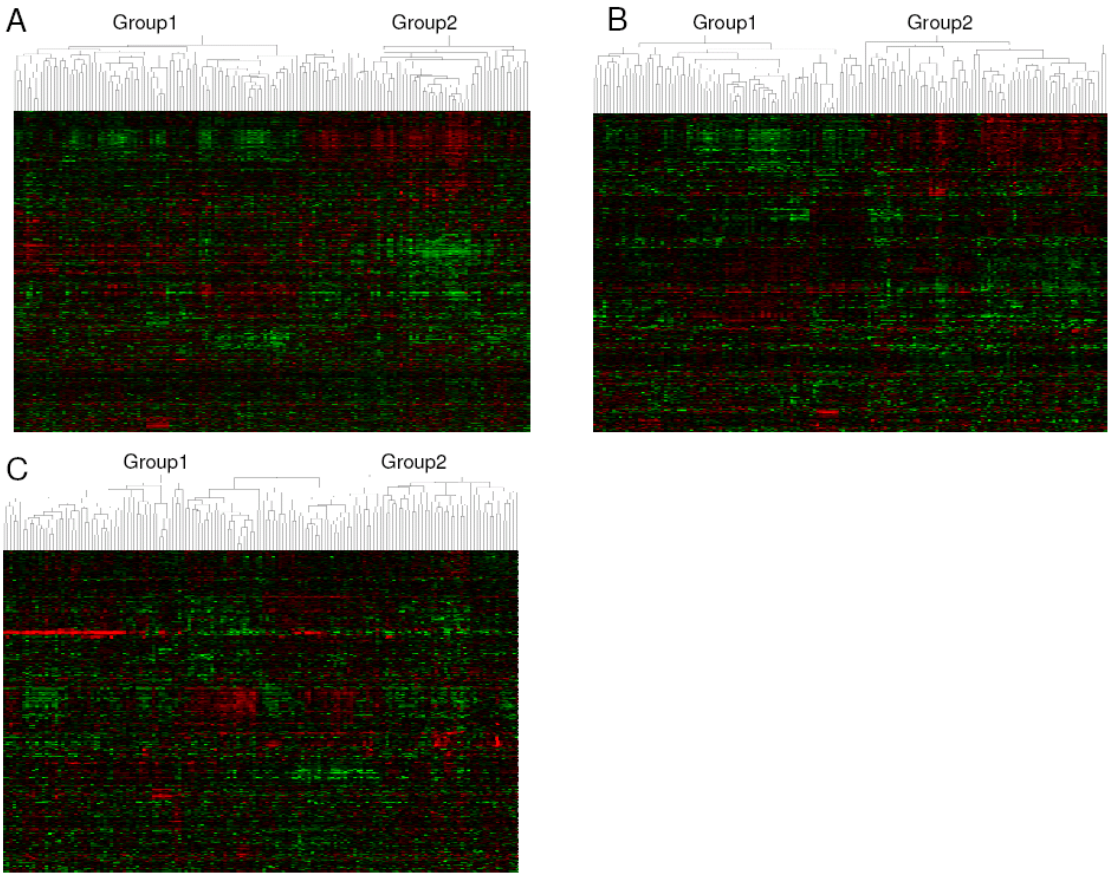


Figure 2: Hierarchical clustering heatmap of breast cancers based on expression of genes in breast cancer gene marker set (A), cell cycle pathway (B), and NF-κB pathway (C). The dendrograms indicated that patients are clustered into two groups (Group1 and Group2) according to their expression patterns of the specified gene set.

3.4.3. Identify pathways with gene expressions correlated with clinical outcome using unsupervised clustering

We first tested the Amsterdam 70-gene signature and the breast cancer gene set including 264 genes as known molecular markers in the prognosis

and diagnosis of breast cancer. Our goal was to examine if patients with differential expression patterns of these markers exhibited distinct survival probabilities as one would expect. This is a proof-of-concept test and served as the positive control in our study. As demonstrated in Table 2, there is indeed a significant difference in clinical outcome between the two patient groups with distinct expression patterns of genes in the 70-gene signature or in the 264 breast cancer gene set. This result is reproducible in all of the five datasets ($P < 0.05$). We would like to emphasize that the five array datasets we analyzed were generated from different patient cohorts that included a total of 1,162 breast tumor samples. Figure 2A depicts a heatmap of the breast cancer gene marker expressions in 159 samples of one dataset (Pawitan et al, 2005). The column dendrogram revealed these 159 patients were clustered into two groups with opposite expression patterns. The two groups exhibited a markedly different survival as revealed by the Kaplan-Meier analysis (Figure 3A).

We next investigated if gene sets based on any of the well known pathways (see additional file 1) could be used as cancer prognosis markers. As shown in Table 2, breast cancer patients with differential gene expressions in cell cycle had significantly different clinical outcome shown in all of the five datasets ($P < 0.05$), suggesting that the cell cycle pathway may be functionally important in breast cancer progression and that the genes in this pathway could be used as prognosis markers. EGF, FGF, G1-S and p53 pathways exhibited significant correlation between gene expression and survival in 4

datasets. This is somewhat expected given that G1-S transition is a part of the cell cycle pathway and significant roles of EGF, FGF and p53 pathway genes in regulating cell cycle. Figure 2B illustrates in one breast cancer array study (Pawitan et al, 2005), tumor samples can be separated into two groups with distinct expression patterns of cell cycle genes, and the two groups had significantly different survival probabilities (Figure 3B). In contrast, patients with distinct expression patterns of genes in the NF-κB pathway (Figure 2C) have similar outcomes (Figure 3C).

3.4.4. Confirm prognostic gene signatures in cell cycle pathway using supervised classification

Next we applied the PAM (Prediction Analysis for Microarray) method (19), a supervised learning algorithm to confirm the predictive powers of cell cycle pathway genes for breast cancer clinical outcome, and to build a gene signature prognostic model. We used the Wang study (Wang et al, 2005) as the training dataset to build a classification model from the Amsterdam 70-gene set, the breast cancer marker gene set and the cell cycle pathway gene set, respectively, using the PAM algorithm. The models were fitted on 90% of the samples and tested on the remaining 10%. Each patient in the 10% testing samples was classified into the good or the poor prognosis groups based on the model developed using the training data. The procedure was repeated 10 times so each of the 10 groups was used as the testing sam-ples and contributed to the overall prediction accuracy. Kaplan-Meier analysis of the

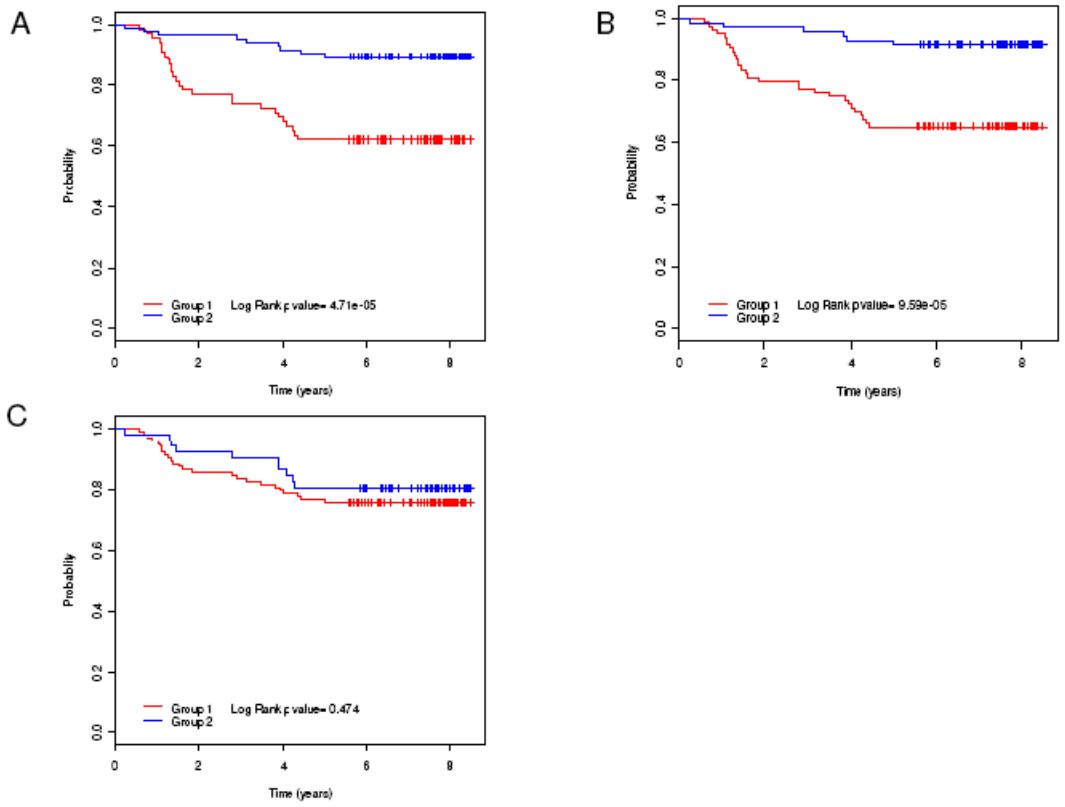


Figure 3: Kaplan-Meier survival analysis of breast cancer patient groups defined by the hierarchical clustering analysis shown in Figure 2 for breast cancer gene marker set (A), cell cycle pathway (B), and NF-κB pathway (C).

predicted good and poor prognostic groups was performed to assess the predictive power of the models. We further carried out independent validation in two other datasets based on the same Affymetrix array platforms U133A (Table 1). The van de Vijver dataset (3) and the Bild dataset (Bild et al, 2006) were based on completely different microarray platforms, an InkJet oligonucleotide array and Affymetrix U95Av2 array respectively, and therefore were omitted in independent validation analysis due to technical reasons (for

example, many genes in the prognostic models built on the Affyme-trix U133A arrays are not represented on the InkJet oligo-nucleotide array and Affymetrix U95Av2 array). The patient samples in the two validation datasets (Pawitan et al 2005; Miller et al, 2005) were classified into the good and poor prognostic groups respectively using the models developed from the Wang study (Wang et al, 2005), subsequently followed by Kaplan-Meier analysis. The significance of differential survival probabilities between the two groups, represented by log-rank test P values in the Kaplan-Meier analysis, were recorded as shown in Table 3. Both the cell cycle signature we developed and the previously identified breast cancer gene signature performed well as prognostic biomarkers in training dataset and two independent validation datasets. However, the 70-gene Amsterdam signature was less accurate, particularly when evaluated using independent datasets. A set of 26 gene transcripts in the cell cycle pathway exhibited expression elevations greater than 2 fold in the poor prognosis groups in our training dataset (Table 4) and most of these genes have well documented roles in cancer development.

We also randomly selected 232 genes, the number of genes used in the breast cancer gene set signature, to build prediction models and the random models were similarly assessed in the training dataset and two independent datasets as described above. This random testing was repeated 100 times and the P-values in the Kaplan-Meier analysis were the average of the 100 experiments. Interestingly, the classification models based on randomly selected genes performed exceptionally well in the training dataset

using the10-fold cross validation procedure (Table 3), suggesting if one uses a large number of genes to build a prediction model, some of the randomly chosen genes will be differentially expressed between the good and poor prognosis groups by chance and therefore provide prognostic values. However, when analyzed in independent datasets of different patient cohorts, the models with random genes did not show predictive power (Table 3), demonstrating that microarray based gene expression predictors must be tested through multiple independent datasets to validate their robustness, a practice that has failed to be recognized by most published studies in the literature.

Table 3: Evaluation of cell cycle gene expression signature as breast cancer prognosis markers by supervised methods

		Dataset		
Gene signature model	Number of genes used in the classification model	Training and testing: Wang dataset	Independent validation: Miller dataset	Independent validation: Pawitan dataset
The 70-gene signature	51	6.10E-05	0.057	0.051
Breast cancer	232	2.60E-09	0.0012	0.0019
Cell cycle	108	1.40E-06	0.005	0.0046
Random	232	1.80E-13	0.14	0.52

The numbers represent the log-rank test P values in Kaplan-Meier analysis in the good and poor prognosis groups predicted by the Amsterdam 70 gene signature, the breast cancer gene set, the cell cycle gene set classifier, and the randomly selected gene set respectively.

3.5. Discussion

Our analysis demonstrated that differential expression of genes in the cell cycle pathway is associated with differential patient outcome in breast cancers, suggesting that cell cycle regulation may be one of the most important factors contributing to breast cancer progression. In fact, cell proliferation markers have been extensively investigated for their prognostic values (Colozza et al, 2005; van Diest et al, 2004). A literature search has revealed expressions of many cell cycle related genes are correlated with breast cancer progression and patient survival as individual outcome predictors. Cyclins bind and activate cyclin-dependent kinases to drive cell cycle progression. The prognostic role of cyclins has been retrospectively assessed in numerous studies. For example, measurement of cyclin E by Western Blot and immuno-histochemistry in 395 breast cancer patients showed that higher level of total cyclin E is strongly correlated with poor outcome (Keyomarsi et al, 2002). Cyclin A, B and D also appeared to be strong prognostic markers in some studies (Kuhling et al, 2003; Peters et al, 2004; Suzuki et al, 2007). CDC25A is a protein tyrosine-threonine phosphatase and regulates G1-S and G2-M transitions. Over-expression of CDC25A is associated with poor prognosis in breast cancers (Evans, 2000). Several independent reports demonstrated that high level E2F1 expression correlates with reduced disease-free survival in node-negative breast cancer patients (Baldini et al, 2006; Han et al, 2003; Vuaroqueaux et al, 2007). Ki-67 (MKI67) antigen induces chromatin condensation and is a well known cell

proliferation marker. A recent review summarized that Ki-67 expression assayed by IHC showed prognostic values in 15 studies where a total of more than 5000 tumor samples were analyzed (van Diest et al, 2004). While these cell cycle related genes have been individually linked to breast cancer outcome, the multi-gene signature we applied in our analysis may provide a more accurate predictor, and more importantly these genes are mechanistically implicated in breast cancer progression. A close examination of gene identities in the cell cycle pathway, the Amsterdam 70-gene signature, and the control breast cancer gene signature revealed that the Amsterdam signature only included one cell cycle gene (cyclin E2). In contrast, the 232-gene breast cancer signature and the 108-gene cell cycle pathway have a 25-gene overlap including several cyclins (cyclin B1, B2, D1, E1, E2), cyclin-dependent kinases (CDK2, CDK4), tumor suppressors p53 and RB1, and the proliferation marker Ki-67, suggesting that predictive power of the control breast cancer signature may be due to the presence of these cell cycle related genes.

Adjuvant therapy and hormonal treatment of breast cancer patients have been demonstrated to improve survival. However, these treatment regimens are costly and could have serious side effects, therefore, should only be recommended to high risk patients. Traditional prognostic factors such as lymph node status, tumor diameter and histological grades do not accurately predict clinical behaviors of the breast tumors and as a result, patients can be over-treated or under-treated depending on the clinicpathological guidelines.

Identification of additional prognostic markers is important for clinicians to select the most appropriate systemic treatments for individual patients according to their risks of relapse or death. Cell proliferation is a key feature of breast tumor progression and has been widely evaluated as a prognosis factor. Although many proliferation markers have been established as robust prognosticators, they have not been applied in clinic due to various technical barriers. For example, 3H-thymidine labeling index (TLI) was one of the first methods developed to evaluate proliferative activity through measuring 3H-thymidine uptake by tumor cells undergoing DNA synthesis (Lloveras et al, 1991; Meyer & Connor 1977; Waldman et al, 1991). However, it has never been adopted as a standard prognostic marker because the experiment requires fresh tumor tissue and a complex and time consuming radioactive assay for in vivo administration of labeled substances. Measurement of DNA content by flow cytometry has provided a reliable approach to determine tumor cell proliferative activity represented by S-phase fraction (SPF) (Hedley et al, 1987), but the lack of standardized procedure to prepare and analyze tumor samples precluded use of this method as a routine assay (Baldetorp et al, 1995). Application of proliferation antigen Ki-67 is hampered as the Ki-67 monoclonal antibody could only be used on fresh or frozen tissue since fixation greatly reduced immunostaining (Urruticoechea et al, 2005). The predictive power of abovementioned cell cycle regulators such as cyclins has not yet proved definitive since in some studies the correlation between protein level and clinical outcome is not significant (Colozza et al, 2005). The

Amsterdam 70-gene expression signature as breast cancer prognosis marker has been validated in follow-up studies (Bueno-de-Mesquita et al, 2007; Buyse et al, 2006), and a clinical assay MammaPrint® has recently been cleared by FDA. However, the two issues associated with the current gene expression signature markers for prognosis, i.e. the lack of a consensus gene set and the difficulty to understand underlying mechanisms, may prevent them from being widely accepted. The cell cycle gene signature we identified in this study has provided a prognostic gene expression marker that not only performed better than the Amsterdam 70-gene signature but is also mechanistically linked to breast cancer progression.

There have been recent reports to incorporate biological pathway information into classification models by using a network analysis approach (Chuang et al, 2007) or to identify functional gene sets from various sources including Gene Ontology to distinguish two different biological phenotypes (Eichler et al, 2007; Subramanian et al,2005). In this study, we assembled 20 pathways that are known to be involved in cancer development and progres-sion, and then extracted expression data of genes only in these pathways in order to identify a mechanistic gene sig-nature biomarker for breast cancer prognosis. We first selected pathways according to their classification powers based on unsupervised analysis, followed by building prognostic gene signature models using the standard supervised methods. The signature developed after pre-selecting relevant pathways should be more reliable and generally applicable as demonstrated by our validation when

applied to multiple independent datasets. This is not surprising since the signature is derived from the cell cycle pathway and it has been well documented that cell cycle control plays a critical role in determining breast cancer outcomes.

We also recognize the limitation of our study. While the cell cycle gene signature derived from a training dataset (Wang et al, 2005) performed well in prognosis prediction in two independent validation datasets (Pawitan et al, 2005; Miller et al, 2005), we did not specifically examine how stable the signature is by building multiple signatures in different datasets in the context of cell cycle pathway and then comparing these signatures for the extent of overlap. We reasoned that there could be significant overlap simply due to a much smaller gene set that we started with in signature model building. Furthermore, we did not attempt to understand the cell cycle signature at the individual gene level to interpret the role of each gene in disease progression based on the numerical coefficients in the signature model because these numerical parameters are heavily impacted by technical variations. Nevertheless, our pathway oriented approach and the analysis results strongly suggest a critical role of the cell cycle pathway in breast cancer progression, which is also consistent with what has been known from a rich collection of literature information.

Table 4: Expression of cell cycle genes in breast cancers

Symbol	ID	Fold*	Description
BIRC5	332	4.25	Baculoviral IAP repeat-containing 5, antiapoptotic cell cycle regulator, expression in many cancers is associated with poor prognosis and mediates cancer cell resistance to taxol and radiation; rat Birc5 is upregulated in response to acute pancreatitis
BRCA2	675	2.13	Breast cancer 2 early onset, a transcription coactivator that binds to RAD51 and TP53, regulates cell proliferation, cell cycle progression, and DNA repair; mutations in the corresponding gene are associated with Fanconi anemia and multiple cancers
CCNA2	890	3.11	Cyclin A2, a cyclin-dependent protein kinase regulator, promotes G2/M transition, progression through cell cycle, cell proliferation, and phosphorylation of proteins; upregulated in male germ cell tumors and testicular tumors
CCNB1	891	2.43	Cyclin B1, complexes with CDC2 to promote nuclear membrane and Golgi disassembly, chromosome condensation, and microtubule reorganization, aberrant expression is associated with multiple neoplasms, increased expression correlates with Alzheimer disease
CCNB2	9133	3.28	Cyclin B2, a CDC2 kinase-associated cyclin that is involved in Golgi apparatus disassembly, may function in p53 (TP53)-mediated cell cycle arrest at the G2/M transition, may mediate cell cycle arrest and is overexpressed in nonendometrioid carcinomas
CCNE1	898	3.01	Cyclin E1, a CDK and histone deacetylase regulator, regulates mitotic G1-S phase transition and promotes cell proliferation, involved in peptidyl-threonine phosphorylation and aging, aberrant mRNA and protein expression is associated with several cancers
CCNE2	9134	2.75	Cyclin E2, a cyclin-dependent protein kinase regulator that binds CDK2 and CDK3, regulates cell cycle checkpoint; mRNA upregulation correlates with breast and lung cancer, mouse Ccne2 is overexpressed in TPA-induced carcinomas and fore stomach cancers
CDC2	983	2.87	Cell division cycle control protein 2, a cyclin-dependent protein kinase that acts in DNA damage checkpoint, inhibits apoptosis and EGFR signaling, expression is increased in Alzheimer disease, viremia associated with HIV infection, and various cancers
CDC20	991	3.72	Cell division cycle 20, a mitotic checkpoint protein and transcriptional repressor, activates the mitotically phosphorylated form of the anaphase promoting complex as well as the mitotic spindle checkpoint, overexpressed in gastric cancer
CDC25A	993	2.7	Cell division cycle 25A, protein tyrosine-threonine phosphatase, regulates G1-S and G2-M phase transitions, functions in apoptosis and oxidative stress response, activity increases in Alzheimer's disease neurons, overexpressed in many cancers
CDC45L	8318	4.9	Cell division cycle 45 like, associates with ORC2L, MCM7, and POLA2, predicted to be involved in the initiation of DNA replication; corresponding gene is located in a chromosomal region frequently deleted in DiGeorge syndrome
CDC6	990	2.47	Cell division cycle 6, involved in DNA replication initiation, may regulate DNA licensing, pre-replicative complex formation and cell proliferation, upregulated in cervical intraepithelial neoplasia and cervical cancer, downregulated in prostate cancer
CDKN2A	1029	2.13	Cyclin dependent kinase inhibitor 2A, interacts with CDK4 and CDK6, involved in aging, anoikis, and cell cycle arrest, regulates transcription factor activity and cell proliferation, aberrantly expressed in psoriasis and several types of cancer
CHEK1	1111	2.54	Checkpoint homolog 1 (S. pombe), protein kinase, required for mitotic G2 checkpoint in response to radition-induced DNA damage, inhibits mitotic entry after DNA damage via mechanism involving CDC25, alternative form is associated with lung cancer
CKS1B	1163	2.08	CDC28 protein kinase regulatory subunit 1B, essential for SKP2-mediated ubiquitination of CDKN1A and CDKN1B, regulate cell cycle progression, aberrant protein expression is associated with several cancers
CKS2	1164	2.27	CDC28 protein kinase regulatory subunit 2, a protein that binds p34 CDC2 and may regulate cell cycle progression, upregulated in pancreatic cancer cell lines

E2F1	1869	2.39	E2F transcription factor 1, inhibits cell proliferation, aberrant expression correlates with several neoplasms and Alzheimer disease associated with Down syndrome; knockout of mouse E2f1 is associated with early onset of diabetes and Sjogren's syndrome
GTSE1	51512	2.61	G-2 and S-phase expressed 1, a cell cycle-regulated and microtubule-associated protein that acts in nuclear-cytoplasmic shuttling of p53 (TP53), may play a role in DNA-damage induced apoptosis through regulation of p53 function during S and G(2) phases
KPNA2	3838	2.18	Karyopherin alpha 2, an NLS binding protein that acts in the nuclear transport of proteins and may play a role in V(D)J recombination, upregulated in breast cancer; human KPNA2 gene map position correlates with fetal growth retardation
MAD2L1	4085	3	MAD2 mitotic arrest deficient-like 1 (yeast), mitotic spindle checkpoint complex component, inhibits anaphase-promoting complex activation, binds MAD1L1, altered expression is linked to several cancers and adult T-cell leukemia
MCM2	4171	2.88	Mini chromosome maintenance deficient 2, binds chromatin, regulates the onset of DNA replication, inhibits the helicase activity of the MCM 4,6,7 complex, expression is altered and is prognostic in a number of cancers
MCM4	4173	2.82	Minichromosome maintenance deficient 4, forms a single stranded ATP-dependent DNA helicase with MCM6 and MCM7, may monitor sites of unreplicated DNA, displacement from replicated chromatin may ensure that DNA is only replicated once per cell cycle
MCM5	4174	2.39	Mini chromosome maintenance deficient 5, transcriptional coactivator that interacts with STAT1, enhances IFNG -induced and STAT1 -dependent transactivation, localizes to unreplicated chromatin, upregulated in anaplastic thyroid carcinoma
MCM6	4175	2.15	MCM6 minichromosome maintenance deficient 6, a component of the heterohexameric MCM complex that has ATP-dependent DNA helicase activity, acts in DNA replication initiation, upregulated in mantle cell lymphoma
MKI67	4288	2.43	Ki-67 antigen, induces chromatin compaction, acts in cell proliferation, expression is altered in neoplasms including osteosarcoma and prostate, breast and esophageal cancer; gene is mutated in cervical, colon and lung carcinoma cell lines
RAD51	5888	2.13	RAD51 homolog, a DNA binding ATPase that acts in apoptosis, cell proliferation, p53-mediated DNA damage response, and double-strand break repair via homologous recombination, aberrant expression correlates with bloom syndrome and several neoplasms

*The fold changes represent the ratio of expression in the poor prognosis group over that in the good prognosis group in the training dataset (Wang et al, 2005).

3.6. Conclusion

Post-genomic technologies have provided a new paradigm in developing tailored therapeutic strategies for treating complex diseases. One notable example is the development of gene expression signatures based on microarray data to predict prognosis and responses to chemotherapy in cancers (Potti et al, 2006). Several studies have revealed that multiplex gene

expression markers are more powerful in predicting clinical outcomes than the traditional clinical criteria. However, the promise of applying these gene signature biomarkers in clinic is hampered because the underlying biology of gene signatures in cancer development is not well understood. Furthermore, different studies often report different gene expression predictors for the same cancer type and as a result, many biologists and physicians remain skeptical of the gene signature concept. In this study, we developed a novel approach to derive gene expression signatures for cancer prognosis in the context of known biological pathways. Our analysis not only generated mechanism based gene signature predictors, but also shed light on the role of different molecular pathways in cancer development. To our knowledge, the current study is the first effort to integrate gene expression profiling data and well known pathway information to develop pathway specific gene expression signatures for cancer prognosis, and our approach will likely provide a new direction in the Oncogenomics field to develop gene signature biomarkers. The predictive power of the cell cycle gene signature for breast cancer prognosis as demonstrated in the present study warrants further investigation such as prospective clinical trials to explore its utility in clinic. Moreover, the methodology we developed could be utilized to identify gene signature biomarkers to guide clinical development of novel cancer therapeutic agents.

Note added in proof

While this manuscript was in preparation, using a com-pletely different approach, Mosley and Keri described a similar observation that cell cycle genes dictate the power of breast cancer prognostic gene list (Mosley & Keri, 2008).

Availability and requirements

http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/ cluster/: providing open source clustering software.

http://www.superarray.com/home.php: providing path-way focused arrays.

http://www.ingenuity.com/: providing pathway analysis tools for interpretation of genomics data.

3.7. References

Baldetorp B, Bendahl PO, Ferno M, Alanen K, Delle U, Falkmer U, Hansson-Aggesjo B, Hockenstrom T, Lindgren A, Mossberg L, et al.(1995) Reproducibility in DNA flow cytometric analysis of breast cancer: comparison of 12 laboratories' results for 67 sample homogenates. Cytometry 22(2):115-127.

Baldini E, Camerini A, Sgambato A, Prochilo T, Capodanno A, Pas-qualetti F, Orlandini C, Resta L, Bevilacqua G, Collecchi P (2006) Cyclin A and E2F1 overexpression correlate with reduced disease-free survival in node-negative breast cancer patients. Anticancer Res 26(6B):4415-4421.

Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Har-pole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439(7074):353-357.

Bueno-de-Mesquita JM, van Harten WH, Retel VP, van't Veer LJ, van Dam FS, Karsenberg K, Douma KF, van Tinteren H, Peterse JL, Wesseling J, Wu TS, Atsma D, Rutgers EJ, Brink G, Floore AN, Glas AM, Roumen RM, Bellot FE, van Krimpen C, Rodenhuis S, Vijver MJ van de, Linn SC (2007) Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: a prospective community-based feasibility study (RASTER). Lancet Oncol 8(12):1079-1087.

Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assig-nies MS, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98(17):1183-1192.

Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, Rijn M van de, Brown PO, Vijver MJ van de (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 102(10):3738-3743.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140.

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Can-cer 8(1):37-49.

Colozza M, Azambuja E, Cardoso F, Sotiriou C, Larsimont D, Piccart MJ (2005) Proliferative markers as prognostic and predictive tools in early breast cancer: where are we now? Ann Oncol 16(11):1723-1739.

Coombes KR, Wang J, Baggerly KA (2007) Microarrays: retracing steps. Nat Med 13(11):1276-1277.

Eichler GS, Reimers M, Kane D, Weinstein JN (2007) The LeFE algorithm: embracing the complexity of gene expression in the interpretation of microarray data. Genome Biol 8(9):R187.

Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21(2):171-178.

Evans KL (2000) Overexpression of CDC25A associated with poor prognosis in breast cancer. Mol Med Today 6(12):459.

Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355(6):560-569.

Han S, Park K, Bae BN, Kim KH, Kim HJ, Kim YD, Kim HY (2003) E2F1 expression is related with the poor survival of lymph node-positive breast cancer patients treated with fluorouracil, doxorubicin and cyclophosphamide. Breast Cancer Res Treat 82(1):11-16.

Hedley DW, Rugg CA, Gelber RD (1987) Association of DNA index and S-phase fraction with prognosis of nodes positive early breast cancer. Cancer Res 47(17):4729-4735.

Hu Y, Hines LM, Weng H, Zuo D, Rivera M, Richardson A, LaBaer J (2003) Analysis of genomic and proteomic data using advanced literature mining. J Proteome Res 2(4):405-412.

Keyomarsi K, Tucker SL, Buchholz TA, Callister M, Ding Y, Horto-bagyi GN, Bedrosian I, Knickerbocker C, Toyofuku W, Lowe M, Herliczek TW, Bacus SS (2002) Cyclin E and survival in patients with breast cancer. N Engl J Med 347(20):1566-1575.

Kuhling H, Alm P, Olsson H, Ferno M, Baldetorp B, Parwaresch R, Rudolph P (2003) Expression of cyclins E, A, and B, and prognosis in lymph node-negative breast cancer. J Pathol 199(4):424-431.

Lee JK, Havaleshko DM, Cho H, Weinstein JN, Kaldjian EP, Karpovich J, Grimshaw A, Theodorescu D (2007) A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. Proc Natl Acad Sci USA 104(32):13086-13091.

Lloveras B, Edgerton S, Thor AD (1991) Evaluation of in vitro bromodeoxyuridine labeling of breast carcinomas with the use of a commercial kit. Am J Clin Pathol 95(1):41-47.

Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, et al (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 5(6):607-616.

Massague J (2007) Sorting out breast-cancer gene signatures. N Engl J Med 356(3):294-297.

Meyer JS & Connor RE (1977) In vitro labeling of solid tissues with tritiated thymidine for autoradiographic detection of S-phase nuclei. Stain Technol 52(4):185-195.

Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proc Natl Acad Sci USA 102(38):13550-13555.

Mosley JD & Keri RA (2008) Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. BMC Med Genom-ics 1(1):11.

Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351(27):2817-2826.

Pawitan Y, Bjohle J, Amler L, Borg AL, Egyhazi S, Hall P, Han X, Hol-mberg L, Huang F, Klaar S, Liu ET, Miller L, Nordgren H, Ploner A, Sandelin K, Shaw PM, Smeds J, Skoog L, Wedren S, Bergh J (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Res 7(6):R953-964.

Perou CM, Sorlie T, Eisen MB, Rijn M van de, Jeffrey SS, Rees CA, Pol-lack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D (2000) Molecular portraits of human breast tumours. Nature 406(6797):747-752.

Peters MG, Vidal Mdel C, Gimenez L, Mauro L, Armanasco E, Cresta C, Bal de Kier Joffe E, Puricelli L (2004) Prognostic value of cell cycle regulator molecules in surgically resected stage I and II breast cancer. Oncol Rep 12(5):1143-1150.

Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR (2006) Genomic signatures to guide the use of chemotherapeutics. Nat Med 12(11):1294-1300.

Quackenbush J (2006) Microarray analysis and tumor classification. N Engl J Med 354(23):2463-2472.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn M van de, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98(19):10869-10874.

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lonning PE, Brown PO, Borresen-Dale AL, Botstein D (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100(14):8418-8423.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gil-lette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545-15550.

Suzuki T, Urano T, Miki Y, Moriya T, Akahira J, Ishida T, Horie K, Inoue S, Sasano H (2007) Nuclear cyclin B1 in human breast carcinoma as a potent prognostic factor. Cancer Sci 98(5):644-651.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA 99(10):6567-6572.

Urruticoechea A, Smith IE, Dowsett M (2005) Proliferation marker Ki-67 in early breast cancer. J Clin Oncol 23(28):7212-7220.

van Diest PJ, Wall E van der, Baak JP (2004) Prognostic value of proliferation in invasive breast cancer: a review. J Clin Pathol 57(7):675-681.

van't Veer LJ, Dai H, Vijver MJ van de, He YD, Hart AA, Mao M, Peterse HL, Kooy K van der, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530-536.

Vijver MJ van de, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Velde T van der, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347(25):1999-2009.

Vuaroqueaux V, Urban P, Labuhn M, Delorenzi M, Wirapati P, Benz CC, Flury R, Dieterich H, Spyratos F, Eppenberger U, Eppenberger-Castori S (2007) Low E2F1 transcript levels are a strong determinant of favorable breast cancer outcome. Breast Cancer Res 9(3):R33.

Waldman FM, Chew K, Ljung BM, Goodson W, Hom J, Duarte LA, Smith HS, Mayall B (1991) A comparison between bromodeoxyuridine and 3H thymidine labeling in human breast tumors. Mod Pathol 4(6):718-722.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460):671-679.

CHAPTER FOUR: PPEA - A NEW FEATURE SELECTION ALGORITHM FOR IDENTIFICATION OF TOXICOGENOMIC BIOMARKERS IN HEPATOTOXICITY (PAPER II)

4.1. Abstract

Toxicogenomics promises to aid in predicting adverse effects, understanding the mechanisms of drug action or toxicity, and uncovering unexpected or secondary pharmacology. However, modeling adverse effects by using the high dimensional and high noise genomic data is prone to over-fitting. Models constructed in this way often consist of a large number of genes with no obvious functional relevance to the biological effect the model intends to predict, which can make it challenging to interpret the modeling results. To address these issues, we developed a novel algorithm, Predictive Power Estimate Algorithm (PPEA). Let $M_{p\times n}$ be the expression data matrix for *p* genes and *n* samples. The algorithm iteratively applies a two-way bootstrapping over *p* and *n* such that the sample size is larger than the gene number for each subset. The predictive power of individual gene is estimated from using the prediction analysis of microarray (PAM). We showed that the relative predictive power of each individual gene was stabilized after a limited number of iterations. Thus, it is possible to construct a predictive model from a much smaller set of genes than the sample size with the highest predictive power. Applying PPEA to DrugMatrix™ rat liver data, we evaluated and ranked predictive power of individual genes as biomarkers of hepatic injury for

inflammation, bile duct hyperplasia and cell death. We further demonstrated that the top ranked genes were functionally related to the molecular mechanisms of the pathologies. More importantly, the models constructed from a small number of the top ranked genes achieved high sensitivity and specificity when tested with completely independent datasets. Thus, we believe that the PPEA model can overcome the over-fitting problem and be used to facilitate genomic biomarker discovery and development for predictive toxicology and to elucidate mechanisms of drug action and/or of toxicity.

4.2. Background

Many preclinical candidate compounds do not achieve ultimate regulatory approval because of organ toxicity. Up to half of these compounds that are discontinued because of organ toxicity are due to DILI including necrosis, steatosis, cholestasis, proliferation, inflammation, and bile duct hyperplasia (BDH) (Ozer et al, 2008). It has been well-documented that biomarkers that identify incipient damage that leads to preclinical and clinical toxicities will enable better decision-making during drug development (Ryan et al, 2008). Particularly valuable are translational biomarkers that bridge preclinical testing species and humans as they can expand the usefulness of the former for detection of human liabilities (Sistare & DeGeorge, 2007).

Although a sole biomarker is appealing as it can be easier to understand, there are few examples in preclinical testing or in clinical practice wherein a single measurement is considered a definitive. Multiple markers are

required to capture the biological heterogeneity of organs involved, individual variations and disease or toxicity processes (Mendrick, 2008). The microarray technology allows us to observe and assess thousands of genes expression simultaneously in each sample. Machine learning algorithms can be applied to identify gene signatures or biomarkers. Numerous recent studies have demonstrated that gene expression signatures not only outperform traditionally used clinical parameters in toxicity or disease outcome prediction, but also contribute to a better understanding of the biological mechanism (Fielden et al, 2005; Luo et al, 2005; Bushel et al, 2007; Zidek et al, 2007; Euna et al, 2008). However, the gene signatures obtained for the same biological phenotype by different researchers differ widely and have only very few genes in common (Fan et al, 2006; Liu et al 2008). This lack of congruence raises doubts about the reliability and robustness of the reported predictive signatures and is believed to partially result from over-fitting (Dessi & Pes, 2009). Over-fitting can raise when the number of training samples is small and the number of genes relatively large, since in such a situation we can easily obtain a classifier that correctly describes the training data but performs poorly on an independent set of data.

The over-fitting has been closely examined by several studies (Sima & Dougherty 2008; Dougherty et al, 2009). Two studies in logistic and Cox regression shows increasing bias and variability, unreliable confidence interval coverage, and problems with model convergence as events per variable (EPV) declined below 10 and especially below five, leading to the rule of thumb that

logistic and Cox models should be used with a minimum of 10 EPV (Vittinghoff & McCulloch, 2007; Peduzzi et al, 1996). Therefore, feature selection is commonly performed before sample classification, and is even attempted to alleviate the above stated problem. Although numerous reports for feature selection have been published and some techniques have been claimed better than others (Guyon et al, 2002; Zhang et al, 2006; Dess & Pes, 2009; see Saeys et al, 2007 for a comprehensive review). To date, no single recommendation in literature is given for methods in either the feature selection or the classification of microarray data (Guyon & Elisseeff, 2003; Saeys et al, 2007).

Feature selection algorithms mainly fall into two broad categories, the filter model or the wrapper model (Das, 2001; Kohavi & John, 1997). The widely accepted filter techniques are single-feature based, and demonstrated to be effective for improving sample classification accuracy. Some of them are statistical tests (t-test, F-test) (Bø et al, 2002), non-parametric tests like TNoM (Ben-Dor et al, 2000), S2N ratio (signal to noise ratio) (Golub et al, 1999) etc. However, these methods have its limitation in which the interaction with classifier and feature dependencies has been completely ignored. However, the interactions between genes are important for very numerous-if not all-biological functions (Barabási & Oltvai, 2004; Gavin et al, 2006). Although the *wrapper* methods use the interactions between features, perform multivariate gene subset selection, and incorporate the classifier's preference or bias into the search and thus offer an opportunity to construct more accurate classifiers,

the disadvantages are computationally intensive, classifier dependent selection, and particularly, high risk of over-fitting (Saeys et al, 2007). In the present study, we have developed a new method, Predictive Power Estimate Algorithm (PPEA), to evaluate and rank the relative predictive power of individual genes. By applying PPEA to DrugMatrix toxicogenomic database, we identified and validated three small sets of genes highly predictive of and functionally related to liver inflammation, bile duct hyperplasia (BDH) and cell death respectively. Furthermore we developed and validated a RT-PCR assay as a genomic biomarker to predict BDH.

4.3. Materials and methods

4.3.1. The PPEA algorithm

Let $M_{P\times N}$ be the expression data matrix of *P* genes as rows and *N* samples as columns, among which N_1 samples are labeled as $T_1, T_2, \dots T_{N1}$ for toxicity class and N_2 samples labeled as $NT_1, NT_2, \dots NT_{N2}$ for non-toxicity class. Thus, $N = N_1 + N_2$. Let α be a predetermined threshold of acceptable classification error rate and β be the arbitrarily defined sample split ratio to construct training and testing sample sets. Let K be the total number of iterations and *k* be the k^{th} iteration $(k = 1,2, \dots, K)$. Let $E^k_{P\times 4}$ be the performance matrix in the k^{th} iteration consisting of *P* rows, each of which is identified by the genes g_i $(i = 1,2, \dots, P)$ in the data matrix $M_{P\times N}$, and four columns corresponding respectively to T_i^k as the total number of times g_i is sampled in the k^{th} iteration, S_i^k as the total number of times g_i selected in the

predictive model in the k^{th} iteration, $P_i^k = S_i^k / T_i^k$ as an estimate of predictive power of g_i in the k^{th} iteration, and $R_i^k \subset (1,2,\dots,P)$ as the rank order of g_i based on its predictive power P_i^k. Genes with larger P_i^k are more predictive than those with smaller P_i^k and thus ranked higher. Let K be the total number of iterations. At the initiation of the algorithm, $E_{P\times 4}^0 = 0$. For each iteration $k = 1,2,\dots,K$, execute following steps.

Step 1: Apply two-way bootstrapping to the $M_{P\times N}$to obtain a bootstrapping sample matrix $S_{p\times n}^k$ consisting of *p* genes, g_j $(j = 1,2,\dots,p)$, randomly drawn from *P* genes, n_1 samples from N_1 samples of toxicity class and n_2 samples from N_2 samples of non-toxicity class such that $n_1/N_1 = \beta$, $n_2/N_2 = \beta$, $n = n_1 + n_2$ and $p < n$. *n* is the sample size of training sample set while $(N - n)$ is the sample size of testing sample set.

Step 2: Apply Prediction Analysis of Microarray (PAM) to the bootstrapping sample matrix $S_{p\times n}^k$ to perform sample classification using the nearest shrunken centroid method (Tibshirani et al 2002). To build a predictive PAM model, ten-fold cross validation was performed to find out the optimal classifier performance which minimizes classification errors for the training set $S_{p\times n}^k$ Based on the ten-fold cross validation, a threshold Δ^k was varied in search of the optimal classifier performance. The Δ^k is chosen when the lowest classification errors achieved with the fewest genes $g_1, g_2, \dots, g_l \; where \; l \leq p$. The resultant PAM model in the current k^{th} iteration

$$m^k = f(g_1, g_2, \dots, g_l) \qquad l \leq p \qquad (1)$$

is subsequently tested using the $(N-n)$ testing samples. Let *e* be the error rate of the model when tested with the testing samples and estimated by (2).

$$e^k = \frac{false\ positives + false\ negatives}{N-n} \qquad (2)$$

In cases where cross validation errors are greater than α for all possible Δ^k value, i.e., no acceptable PAM model can be constructed from genes $g_1, g_2, \dots, g_l\ where\ l \le p$ for training samples, the independent model test using $(N-n)$ testing samples described above is omitted and the execution proceeds to Step 3b described below.

Step 3a: If $e^k \le \alpha$, *i.e.,* the estimated error rate of the model tested with $(N-n)$ samples is less than the predetermined threshold, the model is deemed to be predictive. The performance matrix $E_{P\times 4}$ is updated as follows. Each gene, g_j $(j = 1,2,\dots,p)$, in the boostrapping samples $S^k_{p\times n}$ is mapped to g_i $(i = 1,2,\dots,P)$ in $E_{P\times 4}$, T_i^k, S_i^k, and P_i^k are updated sequentially as follows.

$$T_i^k = T_i^{k-1} + 1$$

$$S_i^k = \begin{cases} S_i^{k-1} + 1 & if\ g_i \in (g_1, g_2, \dots, g_l) \\ S_i^{k-1} & if\ g_i \overline{\in} (g_1, g_2, \dots, g_l) \end{cases}$$

$$P_i^k = S_i^k / T_i^k$$

Step 3b: On the contrary, if $e^k > \alpha$, *i.e.,* the estimated error rate of the model tested with $(N-n)$ samples is larger than the predefined threshold, the model is deemed to be not predictive for independent testing samples thus over-fitting. T_i^k, S_i^k, and P_i^k in the performance matrix $E_{P\times 4}$ are updated sequentially as follows.

$$T_i^k = T_i^{k-1} + 1$$

$$S_i^k = S_i^{k-1}$$

$$P_i^k = S_i^k / T_i^k$$

Sort P_i^k decreasingly, *i.e.*, $P_{g_{i_1}}^k \geq P_{g_{i_2}}^k \geq \cdots \geq P_{g_{i_P}}^k$, a rank order of genes in term of their predictive power is given as

$$R^k = 1,2,\ldots,P$$

Stop criterion. The rank order R^k is evaluated periodically, say every 10000 iterations, by computing Spearman correlation coefficient between the current rank R^k and the previous rank R^{k-1}, *i.e.*,

$$\rho = 1 - 6\sum_{i=1}^{P} \frac{(R_i^k - R_i^{k-1})^2}{P(P^2-1)}$$

The algorithm stops if $\rho > 0.99$, *i.e.,* the rank may be stabilized even if $k < K$.

For a better visualization of this methodology, each step of the algorithm is illustrated in Figure 1, and the R code is attached as Appendix E.

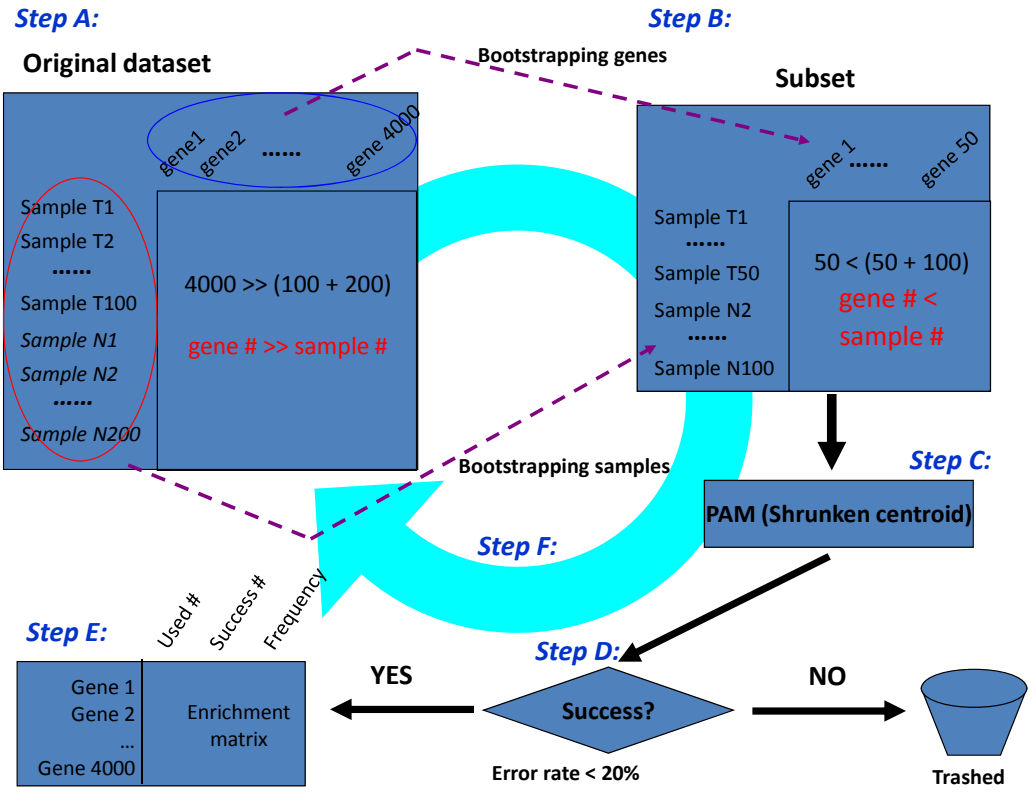


Figure 1: The architecture and workflow of PPEA

4.3.2. Data preprocessing

All the animal studies as well as the array analysis were performed by Entlos™ and the integrated toxicogenomic data had been stored and organized in DrugMatrix® database (Ganter et al, 2005). In the present study, all the array data were downloaded from DrugMatrix®. The array data in RU-1 platform with the detailed animal study information was previously described in detail (Natsoulis et al, 2008) and could be publicly accessible from Gene Expression Omnibus (GEO accession no. GSE8858). For the signal intensity data generated by one-channel oligonucleotide microarrays, we pre-processed the data so each gene has zero mean value and unit variance. In all classification cases, a 60/40 split-sample procedure was applied and the performance was reported as the average of the test results for 1000 random partitions of the data.

Table 1 lists the overview of the positive class compounds and experiments for a given phenotype used in this study. The definition for positive or negative class is similar as previously described (Natsoulis et al, 2008). The positive class was usually defined as the set of samples sharing a particular property for a given phenotype, while the negative class was often defined as the remainder of the sample space. A portion of the sample space was sometimes excluded when the true phenotype might not be known for some samples because they were not assayed, or they were assayed but assay values were missing or uncertain. In other words, the positive and negative classes were assigned as the extremes of this distribution and the

intermediate samples were excluded. In present study, the P value of ridit score significance for a given phenotype is less or equal to 0.01, and percentage of incidence for each experimental group has to be 100 (3/3 animals for each experiment group) for the positive class, and P great or equal to 0.5, and percentage of incidence has to be 0 (0/3 animals) for the negative class. This had the advantage of training neither for nor against samples with intermediate values.

4.3.3. Performance assessment

To compare our proposed techniques to the existing ones in terms of consistency with the existing techniques and performance, we constructed 6 different classifiers for the performance evaluation. Six different colors show in Figure 4 denotes the corresponding feature selection as follows:

(i) The classifier 'Top', 5 genes selected from the gene set with the highest predictive power ranked by PPEA, is labeled as 'red'. This is our desired signature for a given phenotype.

(ii) The classifier 'Bottom', 5 genes selected from the probe set with the lowest predictive power ranked by PPEA, is labeled as 'green'. Using this set of genes as one of baseline models to see the contrast of discriminative power compared to the 'Top' signature.

(ii) The classifier ‘Random’, 5 genes are chosen at random from the whole pre- filtered gene list as well the class labels are shafted randomly, is colored as ‘black’.

(iv) The classifier ‘Whole’, a whole pre-filtered microarray probe set, is labeled as ‘blue’.

(v) The classifier ‘Single’, 5 genes selected using only a single iteration of PAM classification, is labeled as ‘light green’.

(vi) The classifier ‘Iconix’, a corresponding gene signature from DrugMatrix™, is labeled as ‘pink’.

The choice of different colors is a useful heuristic we adopted for revealing the feature selection by different methods.

4.3.4. Functional analysis

The identified top 20 gene sets were subjected to GO analysis by Ingenuity (http://www.ingenuity.com/) and DAVID (database for annotation, visualization, and integrated discovery; http://apps1.niaid.nih.gov/david/) using Fisher’s exact test.

4.3.5. Validation

4.3.5.1. Validated with independent datasets

The microarray data in rat whole genome 230 plus 2 platform (RG230-2) for rat livers treated with Lilly compounds at various doses and times is retrieved from DrugMatrix™. The corresponding histopathology data was

collected and summarized thorough the curation of Lilly toxicology report, and will be used for the validation.

Table 1: Summary description of datasets

Histopath Group	Histopathology Names	# Treatments	# cmpds	Compound Names
Inflammation	PERIPORTAL, INFLAMMATORY CELL INFILTRATE, MIXED CELL	45	15	HARRINGTONIN;LOMUSTINE;AFLATOXIN B1;ETHANOL;TESTOSTERONE;LIPOPOLYSACC;CARVEDILOL;CERIVASTATIN;4,4';CARMUSTINE;KETOCONAZOLE;METHAPYRILEN;DOXAPRAM;1;N
	NONZONAL, INFLAMMATORY CELL INFILTRATE, MIXED CELL	7	6	STAVUDINE;DOXAPRAM;KETOCONAZOLE;VECURONIUM B;THIOACETAMID;GALLAMINE TR
	CENTRILOBULAR, INFLAMMATORY CELL INFILTRATE, MIXED CELL	31	13	ISONIAZID;2,3,7,8;BETA;CLOFIBRIC AC;LIPOPOLYSACC;ACETAMINOPHE;ALPHA;N;AFLATOXIN B1;3;CARBON TETRA;AMINOSALICYL;THIOACETAMID
BDH	BILE DUCT HYPERPLASIA	34	9	LOMUSTINE;N;METHAPYRILEN;2;VINBLASTINE;CARMUSTINE;CARVEDILOL;4,4';1
Necrosis	HEPATOCYTE, NONZONAL, NECROSIS, APOPTOTIC	27	11	CLOTRIMAZOLE;FENOFIBRATE;ATORVASTATIN;METHAPYRILEN;PIRINIXIC AC;BEZAFIBRATE;SIMVASTATIN;FLUVASTATIN;VINBLASTINE;AFLATOXIN B1;LOVASTATIN
	HEPATOCYTE, CENTRILOBULAR, NECROSIS, ONCOCYTIC	10	4	AMINOSALICYL;CARBON TETRA;THIOACETAMID;N NATEGLINIDE;ALLYL
	HEPATOCYTE, NONZONAL, NECROSIS, ONCOCYTIC	21	10	ALCOHO;PRALIDOXIME ;MANGANESE (I;1;VECURONIUM B;DOXAPRAM;EPIRUBICIN;CLOTRIMAZOLE;HYDROCORTISO

4.3.5.2. Validated with QPCR

The overview of the compounds used for QPCR validation in this study is listed in Appendix B. A total of 18 experiments (7 BDH positive and 11 BDH negative experiments) were available in the database when the present analysis was performed. RNA was extracted from livers of total 48 of animals (rats). 18 animals were observed Histopathologically BDH positive and 30

animals BDH negative. Quantitative real-time PCR (qRT-PCR) was performed for measuring the expression of the four genes selected from top 20 of BDH signature based on the fold change of expression.

4.3.6. Statistical analysis

The statistical analyses were performed with the R statistical package, release 2.9 (http://www.r-project.org/). All genes were log-transformed and all p-values are 2-sided.

For the extraction of a predictive gene set, we selected the top 5 of the most discriminative genes ranked by PPEA. Sensitivity, specificity, positive and negative predictive value (PPV and NPV respectively) were calculated and presented with their 95% confidence interval (CI). These genes were analyzed with quantitative RT-PCR.

4.4. Results

4.4.1. Dataset preparation

The overview of datasets from DrugMatrix™ for three phenotypes in liver, inflammation, necrosis, and bile duct hyperplasia (BDH), used in this study is summarized in Table 1 and is briefly outlined here. We divided each phenotype measurement by severity and incidence, and attempted to find patterns of gene expression changes that were able to classify the phenotype. Samples were separated into two classes, those that share a given phenotype (positive class: the P value of ridit score significance for a given phenotype is

less or equal to 0.01, and percentage of incidence is 100 for each experimental group) and those that do not (negative class: P >=0.5, and percentage of incidence = 0) (Table 1). The positive and negative classes were assigned as the extremes and the intermediate samples were excluded. This way allows us to identify the most positive and negative compounds responsible for each phenotype.

4.4.2. Data preprocessing

We started from data sets that were from RU1 (Agilent) platform with 8565 probes without any additional normalization procedure. These datasets have already been normalized in DrugMatrix database. 4231 transcriptionally informative genes passed the filter with Fold Change >1.5, P value <=0.05, and intensity >=2 (range from 1 to 5 for a whole chip, as defined in DrugMatrix).

4.4.3. Overview of PPEA

The idea behind PPEA is very simple. In order to avoid over-fitting, the basic idea of the proposed method is to enforce the feature size inversely smaller than the sample size in the splitting subset for each iterative classification to identify the informative features. This can be achieved by the following algorithm.

The workflow for PPEA is shown schematically in Figure 1, and is described more formally in the Materials and Methods section. Here, we

summarize the basic elements conceptually. For each iteration, both the samples and candidate genes from original dataset are independently bootstrapped (re-sampled with replacement) to form a small subset, wherein the number of genes is reversely equal or less than the number of samples (step A to step B in Figure 1). Then, PPEA builds a PAM classification for this small subset to assess the merit of each feature (gene) by evaluating their strength of class predictability in a multivariate model (step C). If this PAM model performance pass a certain pre-defined threshold (less than 20% error rate in this study), the features (genes) will be ranked and archived in a performance metric called predictive power enrichment matrix (step D). In other words, if the candidate genes in this subset comprise biologically relevant genes or gene combinations, then PAM is expected to perform and assign higher importance rank to at least some of the genes. Otherwise, the failed subset of features will be discarded. The distribution or stabilization status of gene ranks based on their frequency in predictive power enrichment matrix is then evaluated periodically. The iteration process will be terminated if a predefined number of iterations is reached or when the feature rank correlation (Spearman correlation) between two consecutive evaluations extends to a plateau, whereas the feature rank stability cannot be further improved. It should be noted that the schema for feature search in this algorithm is heuristics and suboptimal as it does not exhaustively search in the space of all possible combinations. The choices of pre-defined threshold and the events per variable (EPV) for the number of features to be selected in the

each iteration are purely ad hoc. Although different settings of these parameters may affect the results, we have observed that, for most cases when the two classes can be reasonably separated with the expression data, the classification performances achieved with different settings were very close to each other, and the majority of features ranked at the top positions were also very stable.

Figure 2 illustrates the recapitulation of sampling distribution by two-way bootstrapping with replacement from the original feature set at 3 different numbers of iterations, 20k, 100k, and 200k in the predictive power estimate matrix. As displayed in Figure 2a, a uniformed distribution for the total number of a gene sampled randomly has been achieved at all levels of iteration of sampling. It indicates every feature in original set has equal opportunity to be selected during this two-way bootstrapping process. Figure 2b shows that the number of successful classification involved for each gene. Apparently, each gene differentiates itself based on their abilities for the classification, whereas the highly successful genes are displayed on the right side, and the poorly successful one is located on left side of the table after sorting by its frequency. Figure 2c demonstrates that the relative success rate that a gene was used in successful modeling, as computed as R=S/T, is mostly similar (overlapped) for each gene at three different iteration checkpoints, and an intuitive metric to rank the predictive power of each feature (gene).

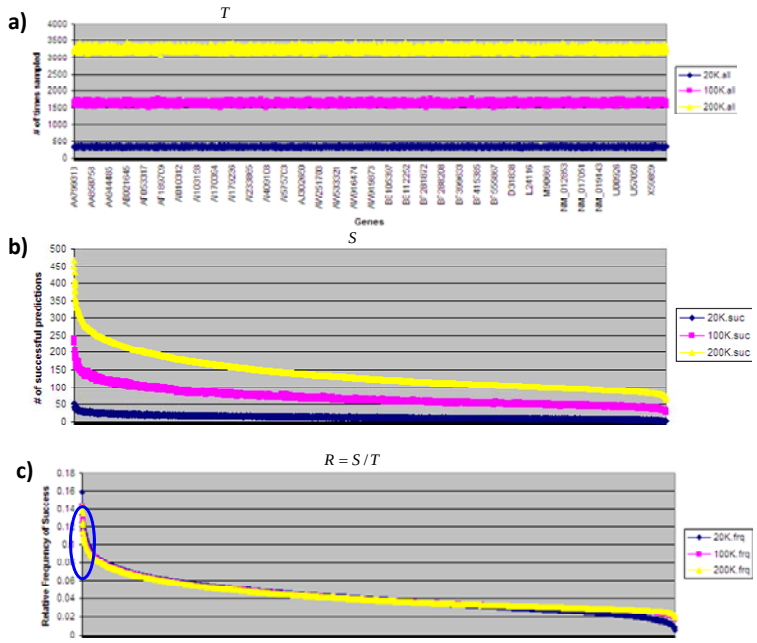


Figure 2: Analysis of sampling distribution in the predictive power estimate matrix. Two-way bootstrapping is performed at 3 different numbers of iterations, 20k, 100k, and 200k. (a) A random number generator with a uniform distribution was used so that each of 4000 features (genes) had equal chances to be sampled. Y axis is the total # of times a gene was sampled, represented as T. (b) A prediction was called a success if sensitivity > 70% AND specificity > 80%. Y axis is the total # of times a gene was included in a successful modeling, denoted as S. (c) Y axis is the Relative Success Rate that a gene was used in successful modeling, computed as R = S/T, which is a metric to measure predictive power of the gene.

In order to give a complete description of the gene rank transition, Figure 3 details the rank shifting at each checkpoint of the iteration for top 10 genes. The result suggests that the rank for 9 of top 10 genes starts to stabilize at early 80k iteration, and tend to be complete plateaued after 220k iterating. It indicates that the absorption and fusion for useful information from these genes becomes satiated after certain number of iteration.

a)

	20K	40K	60K	80K	100K	120K	140K	160K	180K	200K	220K	240K	260K	280K	300K	320K
NM_017207	23	4	2	3	3	2	2	2	1	1	1	1	1	1	1	1
AF251305	28	3	3	1	1	1	1	1	2	2	2	2	2	2	2	2
AI227885	1	1	1	2	2	3	4	4	4	4	4	4	3	3	4	3
AF169636	17	7	6	5	4	4	3	3	3	3	3	3	4	4	3	4
Y00480	15	16	7	7	5	5	5	5	5	5	5	6	6	5	5	5
BE109691	9	21	13	8	7	6	6	7	7	6	6	5	5	6	6	6
AW143130	5.5	2	4	4	6	7	7	6	6	7	7	7	7	7	7	7
AW915705	2	5	5	6	8	8	9	9	9	8	8	8	8	8	8	8
NM_012959	8	6	8	13	14	12	13	11	10	10	10	10	11	9	9	9
BF282961	16	14	31	30	32	29	28	23	20	15	11	12	12	12	11	10

b)

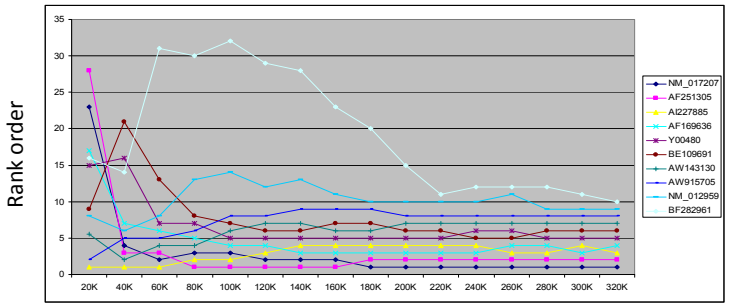


Figure 3: Example of top 10 genes Rank shifting at each checkpoint of the iteration: (a) shows the index of the 10 top-ranked genes (i.e., features) become stabilized when the iteration of splitting reached 280,000. As we can see, the rank for 8 out of 10 genes is consistent as early as the iteration reach to 8,000. (b) A plot for the same data as shown in (a) for an intuitive observation.

4.4.4. Performance assessment and comparison

Our goal is to select a very small subset of features with the maximum discriminatory power between the two classes. Since the feature dimension is large and the sample size is small, there are usually many combinations of features that can give a very small or zero error on the training data. Therefore, the "minimal error" criterion for feature selection on this high dimensional data structure in the original dataset cannot work. The PPEA breaks the original dataset into many small subsets with new low dimensionality, and then evaluates the features based on their predictive merit. We expect this new gene selection algorithm could yield very small sets of genes (often smaller than alternative methods) while preserving high predictive accuracy. Using DrugMatrix™ rat liver data, we first applied this PPEA algorithm to construct the predictive power enrichment matrix for three hepatic specific injuries, inflammation, cell death, and bile duct hyperplasia. Then we used top 5 ranked genes from each list as a classifier for model validation and performance assessment, where the threshold of 5 is chosen based on a common practice in microarray studies, and the practical number of genes can be handled in 96-well plate format in a quantitative Polymerase Chain Reaction (qPCR) assay. Later is a common approach for wet lab validation and biomarker assay implementation. The PPEA is an exploratory method, and the optimal (or biologically reasonable) number of the candidate genes in a signature would depend on the particular data sets with complex trade-off. For presentation clarity and space limitation, we use only BDH

signature here as an example to elucidate the performance and validation of this newly developed algorithm. The summary of the result for BDH, cell death, and inflammation signature performance assessment and comparison is presented in Table 2.

Recall that we construct 6 different classifiers for the performance assessment to verify that the achievable advantage of feature selection with our PPEA algorithm does not occur by chance only. First of all, in most of cases, the accuracy performance show that the “TOP” 5 genes signature generated from PPEA, has significant performance advance comparing to other 5 classifiers in term of error rate (Figure 4a), sensitivity (Figure 4b), and specificity (Figure 4c). The unsupervised hierarchical clustering result suggests that there is clear separation between positive and negative class of toxicity for BDH based on their top-20 gene expression (see the heatmap in Figure 4d). Next, we highlight some of interesting points as follows:

i) Of the six classifiers, overall speaking, the classifier “Top” appeared to be the best, and the classifier ‘Bottom’ the worst. Recall that the classifier ‘Bottom’ is generated from the bottom 5 genes of feature set ranked by PPEA. As shown in Table 2, the sensitivity, specificity, and error rate for ‘Top’ is 95.3%, 94.5%, and 5.5%, respectively. In contrast, the classifier “Bottom” has only 45% and 44.5% for sensitivity and specificity, and 55% for error rate, which is 10-fold higher than “Top” has. This significant difference demonstrates the effectiveness and power of PPEA ranking.

ii) We also observe that the other two baseline models, ‘Whole’ and ‘Iconix’, have good performance. However, the number of features for ‘Whole’ and ‘Iconix’ classifier are 4231 and 66, respectively, which are much larger than “Top” has. As we know, medical doctors and biologists like a small number of features to separate two classes of samples. Manually examining a large amount of features is tedious and sometimes impossible. As shown in Table 2, a small number of most discriminatory features are capable to distinguish the two classes well. Otherwise, even with more number of features, the distinction would not be necessarily become better and even prone to be over-fitted.

iii) Although the ‘Single’ classifier, which selected with single iteration of PAM classification, yields a satisfactory performance, its ability for class discrimination is still behind the ‘Top’ classifier. As shown in Table 2, the P value of the Student t-test are all significant (<0.05) except the sensitivity for ‘Cell Death’ signature. This result indicates that PPEA has the capability to enrich the predictive power.

iv) We notify that the performance for the classifier consisted with randomly selected probes only is not the worst as we expected among these six classifies. This may due to following several reasons: (a) The features are randomly selected from a ‘pre-filtered’ set, which about half of the original features have been filtered out, and each of the rest is expect to have some degree of predictive

As we know, gene signature or biomarkers are measurable and quantifiable cellular characteristics that serve as indicators of normal or pathogenic biological processes. If the biomarkers truly reflect this alternation to a therapeutic intervention, they should have following properties: (i) functional relevance to what they intend to predict, and (ii) highly predictive in independent tests.

Figure 4 (a)

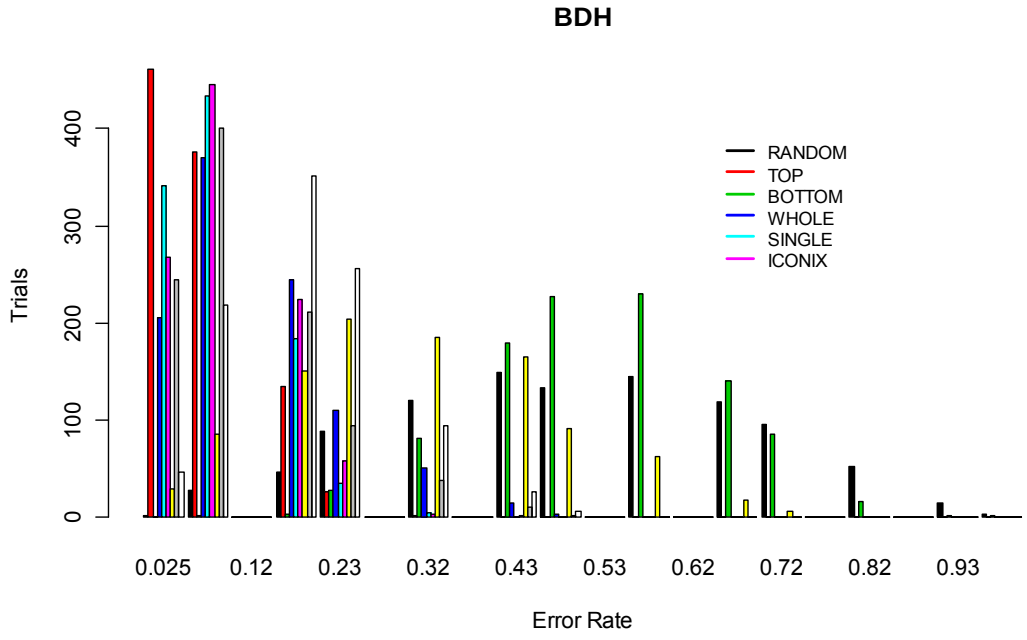


Figure 4 (b)

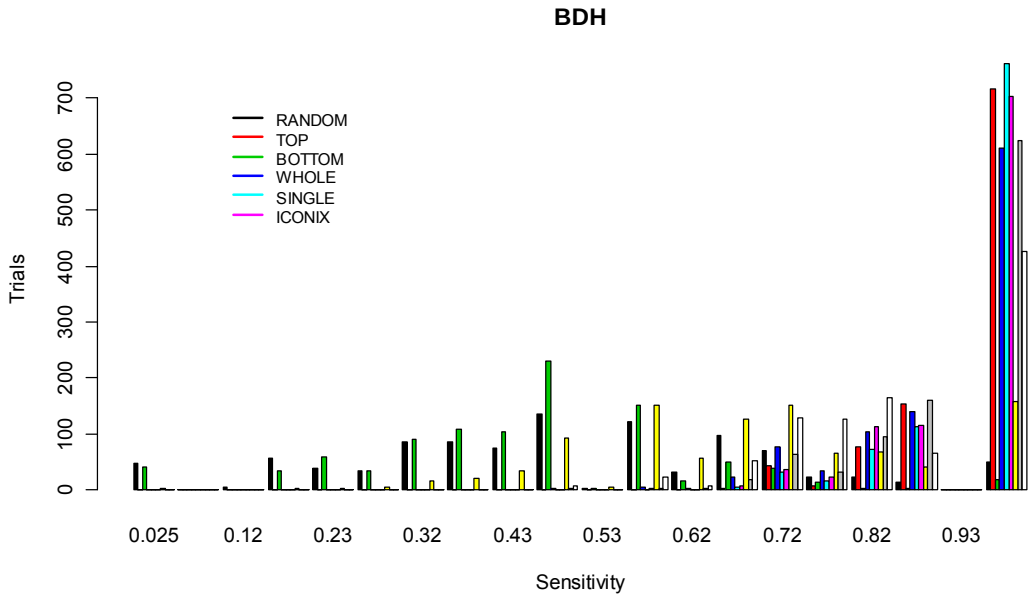


Figure 4 (c)

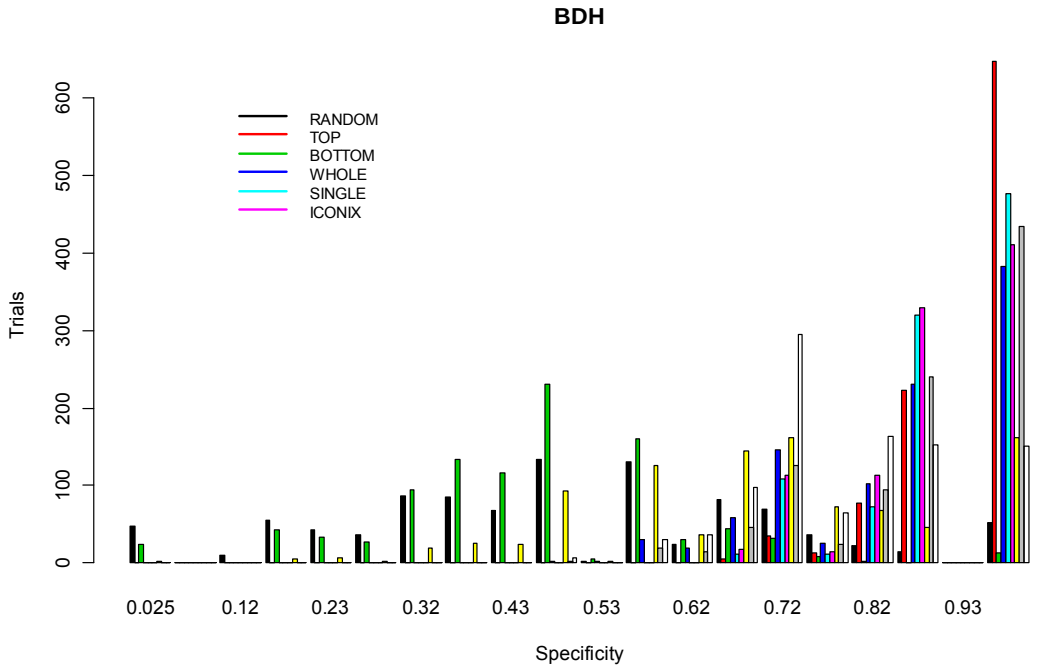


Figure 4 (d)

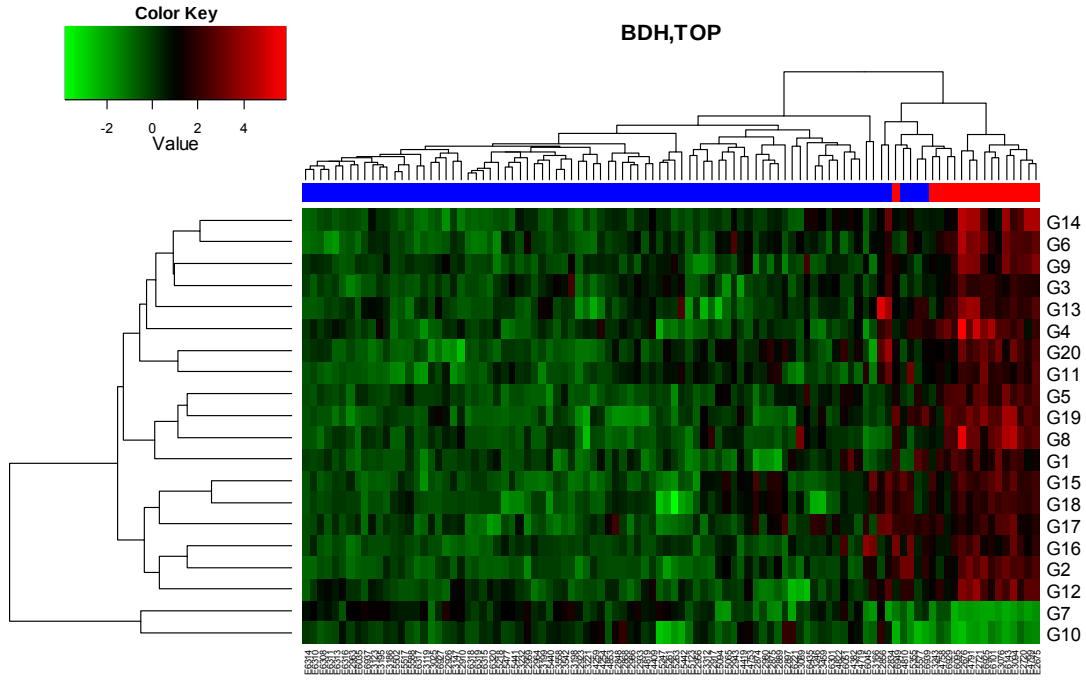


Figure 4: Performance assessment for BDH signature in term of error (a), sensitivity (b), specificity (c), and hierarchical clustering (d): Six subsets of genes have been used for this performance comparison, and denoted as following coloring schema in panel (a - c): 'black' for 5 genes selected by randomization from the whole pre-selected gene list with shafting class labeling, 'red' for 5 genes selected from the gene set with the highest predictive power ranked by PPEA, 'green' for 5 probes selected from the probe set with the lowest predictive power ranked by PPEA, 'blue' for a whole pre-selected microarray probe set, 'light green' for 5 probes selected using only a single iteration of PAM classification, and 'pink' for a corresponding gene signature from Iconix. (d) A heatmap for the expression of top 20 genes selected from the gene set ranked by PPEA. The positive and negative class of toxicity for BDH is labeled in 'red' and 'blue' at the side bar, respectively.

4.4.5. Functional relevance analysis

To ascribe that PPEA is capable of identifying the gene signature truly functional relevant to what they intend to predict, we further show that the signature for BDH is partially associated (12 out of top 20 genes) and may mechanistically related to the oncogenic p53 and ERBB2 pathways (Figure 5). BDH may be part of a general xenobiotic reaction of the liver but is most important as a purely cholangiolar proliferation, usually as a result of exposure to carcinogenic compounds such as Phomopsin (Peterson, 1990). It has been

Table 2: Performance and comparison of six different signatures

		Rate (%)			P-value, comparing to 'TOP'		
	Signature	Cell Death	INFLA	BDH	Cell Death	INFLA	BDH
Error	TOP	20.5%	20.9%	5.5%	NA	NA	NA
	BOTTOM	54.2%	52.8%	55.0%	0	0	0
	WHOLE	23.9%	25.7%	12.5%	1.13E-30	4.42E-76	3.26E-71
	SINGLE	21.2%	23.9%	9.5%	0.0103051	6E-34	3.98E-30
	ICONIX	23.0%	26.3%	10.0%	5.44E-17	1.19E-92	1.56E-38
	RANDOM	51.2%	50.4%	48.1%	0	0	0
sensitivity	TOP	85.1%	84.4%	95.3%	NA	NA	NA
	BOTTOM	45.8%	47.2%	45.0%	0	0	0
	WHOLE	82.3%	81.7%	91.7%	6.69E-13	3.06E-16	1.26E-18
	SINGLE	84.6%	82.0%	91.3%	0.1591644	2.58E-14	1.94E-23
	ICONIX	81.7%	80.5%	93.4%	2.17E-18	1.32E-32	3.55E-07
	RANDOM	49.4%	49.7%	48.1%	0	0	0
specificity	TOP	75.9%	75.5%	94.5%	NA	NA	NA
	BOTTOM	45.6%	47.2%	44.5%	0	0	0
	WHOLE	72.3%	69.9%	85.3%	1.17E-37	9.52E-109	4.49E-93
	SINGLE	75.1%	72.2%	90.9%	0.0009121	9.05E-44	1.11E-20
	ICONIX	73.9%	69.6%	88.1%	1.39E-12	1.52E-120	2.09E-57
	RANDOM	47.6%	49.3%	51.0%	0	0	0

reported that cholangiocellular carcinoma frequently developed from bile duct hyperplasia (Kurashina et al, 1988), although nodular and biliary hyperplasias could not be unequivocally accepted as pre-neoplastic lesions (Smith et al, 1984). A typical histopathological observation for BDH is shown in Figure 5c (Figure 5b is a normal control). The proliferation of bile duct consists of the cells that were often hypertrophic and hyperplastic, which often associated with inflammatory cell infiltrates, edema or even periportal fibrosis. This observation supports the hypothesis that genes for BDH share specific pathways involved with biological mechanism of pre-neoplastic lesions. All

these over-represented biological pathways are closely relevant to cancer- or cancer-liked associated cell hyperplasia. Similarly, we found strong enrichment of that 17 out of top 20 genes for “inflammation” signature identified by PPEA were members of NF-kB pathway, which was a key pre-inflammatory pathway for a xenobiotic response (See Appendix A for more detail.)

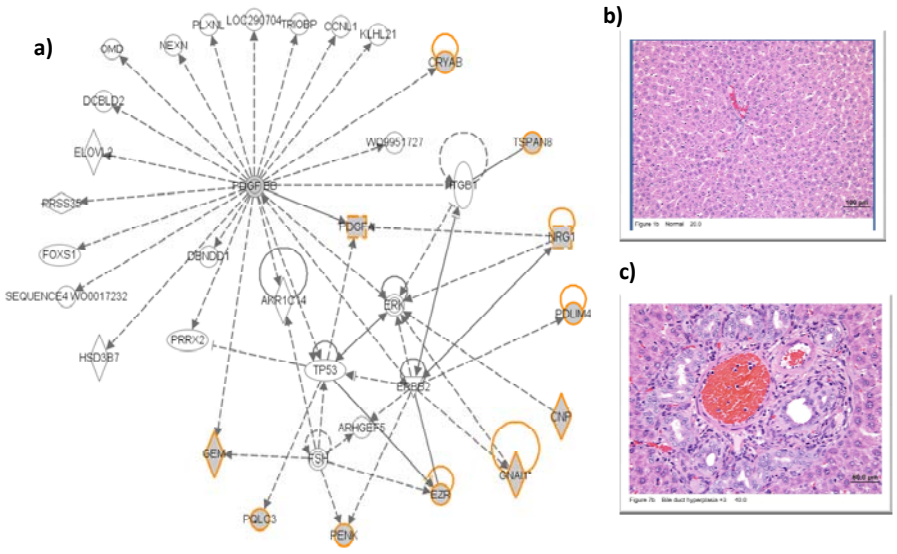


Figure 5: Pathway analysis for the enriched biological functions of the top 20 BDH signature genes. 11 out of top 20 genes are involved with p53 and ERBB2 pathways and highlighted with brown color in (a). A typical histopathological observation for BDH is shown in (c). The proliferation of bile duct consists of the cells that were often hypertrophic and hyperplastic, which often associated with inflammatory cell infiltrates, edema or even periportal fibrosis. (b) A section of normal liver.

4.4.6. Validation with complete independent datasets

To verify that the feature selection procedure with PPEA does not lead to overfitting and can achieve the objective of aggressively reducing the set of selected genes, a complete independent dataset is used for performance validation. The platform used for this validation set is RG230-2 (Affymetrix) and is different from the RU1 dataset used for signature generation. The summary result for inflammation signature is listed in Appendix C. As shown in Figure 6, top 6 genes for BDH have clear separation between two classes in the PCA analysis and achieve decent performance in the SVM classification. As mentioned in early section, one significant characteristic of overfitting is that easily train a classifier that correctly describes the training data but performs poorly on an independent set of data. Our result demonstrates that the BDH signature achieves a very good accuracy, sensitivity, and specificity in an independent dataset with a different platform, lending further credence to PPEA for avoiding overfitting.

4.4.7. Confirmation and assay development with qPCR

Microarray gene expression profiling is difficult to translate into a toxicity liability surrogate assay or a clinical prognostic tool given the large number of genes involved (Ramaswamy, 2004; Mendrick, 2008) and required time and expertise. qPCR is more clinically applicable, especially when working with a small group of highly selected genes. Motivated by the performance consistency of PPEA generated top 5 gene signatures for BDH, inflammation,

Figure 6: BDH signature validation with an independent dataset: (a) dataset information and test procedure; (b) A PCA analysis shows a clear separation between positive and negative compounds; and (c) a SVM classifier performance matrix shows a satisfactory sensitivity (88%) and specificity (78%) has been achieved with top 6 genes.

and cell death among two different microarray platforms (RU1 and RG230-2), we performed a small scale of RT-QPCR experiment to test the performance of top four genes for BDH signature along with two commonly used control gene, cyclophillin and ribosomal protein (RPLRO). Expression Fold Change (FC, the expression value of gene of interest vs. control genes) of each gene were plotted versus individual animal treated with BDH positive and negative

compounds (Figure 7). Apparently, there is a clear separation between positive and negative class based on the expression difference of top 4 genes in BDH signature, and achieved a very good performance by SVM and PAM (see Appendix D for detail). This result indicates that our methodology become applicable across many other biological phenotypes: compiling microarray data for bioinformatics analysis, generating a small list of robust genes involved in a given phenotype, and deriving a smaller QPCR gene signature to predict an outcome of interest or screen preclinical compounds for toxicity liability studies.

Figure 7: A scatter plot of Fold Change (FC, gene of interest vs. house-keeping gene) versus animals for RT-qPCR of top 4 genes in BDH signature. The red and blue bar indicates animals treated with BDH positive (animal 1-

18) and negative compounds (animal 19-48), respectively. Apparently, there is a clear separation between positive and negative class based on the expression level of top 4 genes in BDH signature.

4.5. Discussion

The recurring question when working with microarray data is how to handle the ubiquitous “overfitting” in gene expression profiling. Because of the uniqueness of the resulting microarray data whereas the sample size is typically far smaller than the feature size, this situation necessitates dimensionality reduction through gene selection to avoid data overfitting and improve generalization of discriminant analysis.

In this paper, we propose a novel feature selection algorithm termed as PPEA to alternatively tackle this fundamental issue. PPEA first applies two-way bootstrapping to manage the number of genes inversely equal or less than the number of samples in each splitting subset using for machine learning, and then assess the merit of each individual feature by evaluating its strength of class predictability under this new low dimensional sample-feature space. This approach is different from the other feature selection algorithms in that it assesses gene importance within the context of a multivariate model. That enables PPEA to access the gene information contained in complex biological interrelationships, rather than relying on the summation of univariate relationships within a set. For example, if two genes in a category were related to the samples' biological process or state by an “exclusive OR” association,

then PPEA could capture that relationship, whereas filter-based summations of univariate associations would be likely to overlook it.

The task of conventional feature selection in microarray analysis is considered as a search problem where each state in the search specifies a distinct subset of the possible relevant features. If the search space is too large, it is possible that the algorithm cannot discover the most selective genes within the search space. Moreover, having too many redundant or irrelevant genes increases the risk of overfitting, computational complexity, and cost and degrades estimation in classification error. The PPEA algorithm described here, in concept, approaches the search space like “divide and conquer”, breaking down the search space into certain number of sub-spaces of the same (or related) type. These sub-spaces become small enough with a new dimensionality (the sample size is reversely larger than the feature size in this case) to be solved directly. The solutions to the sub-space are then combined to give a solution to the original space. In practical, we realize that the random data split in each iteration may creates a characteristic of our method, which is that different runs of the algorithm may select different features. An unfortunate split of the data set may also remove an important feature, affecting thus negatively the classifier’s performance. This situation would be avoided if the number of iteration is large enough. In our algorithm, the iterative randomly splitting and classification process is terminated when the occurrence for each individual feature being randomly selected become roughly same (Figure 2), and the stability of ordered features according to their

predictive power within each predictive power enrichment matrix is reached (Figure 3).

Figure 2 illustrates the index of the 10 top-ranked genes (i.e., features) become stabilized when the iteration of splitting reached 280,000. As we can see, the rank for 8 out of 10 genes becomes consistent as early as the iteration reach to 80,000. We observe that when the number of iteration further increases, the stabilization of rank for each feature does not improve, due to the maximization of informative feature utilization. Empirically we prove the method's robustness regarding feature selection by verifying that most of the time the same features are selected in different runs providing high classifier performance.

We do not claim that our PPEA methods will find all interesting genes, because the schema for feature search in this algorithm is heuristics and suboptimal as it does not exhaustively search in the space of all possible combinations. However, we demonstrate that the rank transition becomes a plateau and the majority of features ranked at the top positions were very stable after certain number of iteratively search.

Our approach may inspire some insight into the biological mechanisms behind a biological phenotype as we consider protein-protein interaction (i.e. feature dependence) and largely avoiding overfitting. The functional analysis demonstrates that the signature genes were mechanistically related to the phenotype the signature intended to predict. For example, usually as a result of exposure to carcinogenic compounds such as Phomopsin (Peterson, 1990),

BDH manifests a purely cholangiolar proliferation and considered as pre-neoplastic lesions. Our result shows that 12 of top 20 genes for BDH signature are partially associated to the oncogenic p53 and ERBB2 pathways (Figure 5). We also observed that 17 out of top 20 genes for “inflammation” signature identified by PPEA were members of NF-kB pathway, which was a key pre-inflammatory pathway for a xenobiotic response (see Supplementary Figure 1). We believe our approach is a step in the right direction to find the genes that are truly reflect the alternation to a therapeutic intervention or disease, and may contribute to new approaches to dissect the heterogeneity of phenotypes and understand disease. Additionally, other methods of measuring gene expression such as qPCR can be used, as the number of genes to be monitored is rather small. For the liver injury datasets we demonstrated that quite accurate diagnoses could be achieved using only the gene-expression levels of 5-20 genes.

4.6. Conclusion

Our method demonstrated its efficiency in finding optimal feature subsets with small size and high classification performance. Results show that the top-5 classifier performs superior to other 5 comparing classifiers. Our approach also shows better performance, not only a providing a very good average accuracy, but also with respect to the performance sensitivity and specificity in complete independent dataset. Furthermore, our method is capable of identifying the gene signature truly functional relevant to what they

intend to predict. Thus, we believe that the PPEA model may largely circumvents the overfitting problem, and can be used to facilitate genomic biomarker discovery and development for predictive toxicology and to elucidate mechanism(s) of drug action and/or of toxicity.

4.7. Refrences

Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet. 5:101-13.

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000) Tissue classification with gene expression profiles. J Comput Biol. 7(3-4):559-83.

Bø TH & I Jonassen (2002) New feature subset selection procedures for classification of expression profiles. Genome Biology, 3(4):research0017.1-0017.11.

Bushel PR, Heinloth AN, Li J, Huang L, Chou JW, Boorman GA, Malarkey DE, Houle CD, Ward SM, Wilson RE, Fannin RD, Russo MW, Watkins PB, Tennant RW, Paules RS (2007) Proc Natl Acad Sci U S A. 104:18211-6

Das S (2001). Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the Eighteenth International Conference on Machine Learning (pp.74-81).

Dessi N & Pes B (2009) An Evolutionary Method for Combining Different Feature Selection Criteria in Microarray Data Classification, Journal of Artificial Evolution and Applications. pp.1-10.

Dougherty ER, Hua J, Sima C (2009) Performance of Feature Selection Methods. Curr Genomics. 10:365-374.

Eun JW, Ryu SY, Noh JH,Lee,MJ,Jang JJ, Ryu JC, Jung KH, Kim JK,Bae HJ, Xie H, Kim SY,Lee SH, Park WS, Yoo NJ, Lee JY, Nam SW (2008) Discriminating the molecular basis of hepatotoxicity using the large-scale characteristic molecular signatures of toxicants by expression profiling analysis. Toxicology 249:176-183

Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM (2006) Concordance among gene-expression based predictors for breast cancer. N Engl J Med 355(6):560-569.

Fielden MR, Eynon BP, Natsoulis G, Jarnagin K, Banas D, Kolaja KL (2005) A gene expression signature that predicts the future onset of drug-induced renal tubular toxicity. Toxicol Pathol. 33:675-83.

Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, Bostian KA, Brady L, Browne LJ, Calvin JT, Day GJ, Breckenridge N, Dunlea S, Eynon BP, Furness LM, Ferng J, Fielden MR, Fujimoto SY, Gong L, Hu C, Idury R, Judo MS, Kolaja KL, Lee MD, McSorley C, Minor JM, Nair RV, Natsoulis G, et al (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol. 119(3):219-44.

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, et al (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631-6.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek, M, et al (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286:531-537.

Guyon I,Weston J, Barnhill S, V Vapnik (2002) Gene selection for cancer classification using support vector machines. Mach. Learn., 46 389-422.

Guyon I. and A. Elisseeff (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3:1157-1182

Kohavi R & GH John “Wrappers for feature subset selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.

Liu J,Campen A, Huang S, Peng SB, Ye X, Palakal M, Dunker AK, Xia U, Shuyu Li (2008) Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. BMC Medical Genomics 1:39

Luo W, Fan W, Xie H, Jing L, Ricicki E, Vouros P, Zhao LP, Zarbl H (2005) Phenotypic Anchoring of Global Gene Expression Profiles Induced by N-Hydroxy-4-acetylaminobiphenyl and Benzo(a)pyrene Diol Epoxide Reveals Correlations between Expression Profiles and Mechanism of Toxicity. Chem. Res. Toxicol. 18:619-629

Mendrick DL (2008) Genomic and genetic biomarkers of toxicity, Toxicology 245:175-181

Natsoulis G, Pearson CI, Gollub J, P Eynon B, Ferng J, Nair R, Idury R, Lee MD, Fielden MR, Brennan RJ, Roter AH, Jarnagin K (2008) The liver pharmacological and xenobiotic gene response repertoire. Mol Syst Biol. 4:175

Ozer J, Ratner M, Shaw M, Bailey W, Schomaker S (2008) The current state of serum biomarkers of hepatotoxicity. 1: Toxicology. 245(3):194-205.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49:1373-9.

Ramaswamy S (2004) Translating cancer genomics into clinical oncology (comment). N Engl J Med 350:1814-6.

Ryan TP, Stevens, JL, Thomas CE (2008) Strategic applications of toxicogenomics in early drug discovery. Current Opinion in Pharmacology 8:1-7

Sima C & Dougherty ER (2008) The peaking phenomenon in the presence of feature selection. Pattern Recognit. Lett., 29:1667-1674.

Saeys Y, Inza I, Larra˜naga P (2007) A review of feature selection techniques in bioinformatics, Bioinformatics, 23:2507-2517.

Sistare FD & DeGeorge JJ (2007) Preclinical predictors of clinical safety: opportunities for improvement. Clin. Pharmacol. Ther. 82:210-214.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci USA, 99(10):6567-6572.

Peterson JE (1990) Biliary Hyperplasia and Carcinogenesis in Chronic Liver Damage Induced in Rats by Phomopsin. Pathology, 22: 213-222

Vittinghoff E & McCulloch CE (2007) Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 165:710-8

Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinformatics. 7:197.

Zidek N, Hellmann J, Kramer PJ, Hewitt PG (2007) Acute hepatotoxicity: a predictive model based on focused illumina microarrays. Toxicol Sci. 99:289-302.

CHAPTER FIVE: CONCLUSION

5.1. Summary

Our goal of this study is to develop new algorithm and methods for gene expression profiling in breast cancer and toxicogenomics. More specifically, this study seeks to develop and refine gene signatures or biomarkers for disease classifications, development, progression, outcome, and mechanism discovery. The proposed methods are built around these aims through following two different aspects.

The first one as described in Chapter Three is to build the predictive models with the genomic-scale molecular information from gene expression profiling integrating with prior knowledge of well-defined pathways. Using five independent data sets, we show that several molecular pathways involved in cancer development by directly regulating angiogenesis or metastasis processes, by regulating cell cycle, apoptosis, DNA repair, or by mediating cell signaling rely upon a single group of correlated genes to predict breast cancer outcome. We also applied the Amsterdam 70-gene signature and the breast cancer gene set including 264 genes as known molecular markers in the prognosis and diagnosis of breast cancer. Our intention was to examine if patients with differential expression patterns of these markers exhibited distinct survival probabilities as one would expect. This is a proof-of-concept test and served as the positive control in our study. There is indeed a significant difference in clinical outcome between the two patient groups with

distinct expression patterns of genes in the 70-gene signature or in the 264 breast cancer gene set. This result is reproducible in all of the five datasets ($P < 0.05$). We would like to emphasize that the five array datasets we analyzed were generated from different patient cohorts that included a total of 1,162 breast tumor samples. An un-supervised hierarchical clustering revealed a cohort of 159 patients was clustered into two groups with opposite expression patterns. The two groups exhibited a markedly different survival as displayed by the Kaplan-Meier analysis. The result also indicates that the pattern of gene expression in the cell cycle pathway can indeed serve as a powerful biomarker for breast cancer prognosis. We further built a predictive model for prognosis based on the cell cycle gene signature and found our model to be more accurate than the Amsterdam 70-gene signature when tested with multiple gene expression datasets generated from several patient populations.

The second aspect is to develop and refine a novel computational algorithm named PPEA for feature selection in order to tackle the 'overfitting' problem in microarray data analysis. As described in Chapter Four, PPEA attempts to take advantage from the combination of filter and wrapper algorithms by exploiting their best performance in two steps. The algorithm first iteratively applies a two-way bootstrapping procedure to estimate predictive power of each individual gene by splitting the dataset into certain number of small subsets, wherein the feature size is smaller than the sample size, and then assessed and ranked the individual features based on its merits by evaluating their strength of class predictability. This gave us the ability to

find feature subsets with small size and high classification performance. Used top 5 genes to build a model, we find that our model not only achieves superior predictive accuracy over several other existed models, but also has the ability to capture the functional relevance and validated in a complete independent dataset. Furthermore, the result obtained from wet lab validation with QPCR shows that top 4 genes of the BDH signature have the capability to clear discriminate BDH positive and negative compounds.

5.2. Limitations

We recognized the limitation of the pathway-based classifier in this study. One of potential shortcomings is that the pre-defined set of genes making up a pathway may be derived from conditions irrelevant to the disease of. Second, the majority of human genes have not yet been assigned to a definitive pathway. Third, as one of my research committee members in my research proposal review points out that this approach ignores the pathway-pathway interactions that could potentially improve the model performance. In addition, it is assumed that each gene within a pathway is equally important to pathway activity, which is often not the case.

As a novel algorithm and has not been tested on more hetero-genetic diseases like cancer except DILI, PPEA has a plenty room to be improved. One of possible weaknesses concerned by Dr. Jake Chen at my research proposal defense meeting is that PPEA could not find all interesting genes, because the schema for feature search in this algorithm is heuristics and

suboptimal as it does not exhaustively search in the space of all possible combinations. We demonstrate that the transition of feature rank based on their predictive capabilities becomes very stable after certain number of iteratively search. It may indicate the maximization of feature informative utilization.

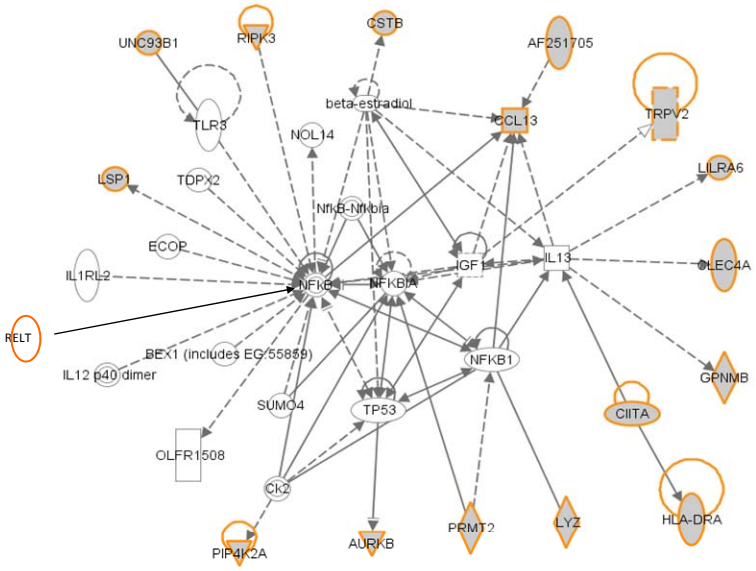
I realize that the structure-based approach will open up new avenues to fundamental understanding of the molecular mechanism associated with biological phenotype. However, our preliminary model incorporated the protein order/disorder information turns out to be only as efficient as the model based on gene expression alone. Thus, I dropped the proposed part of integration with protein structure-based modeling in this dissertation.

5.3. Future research

Futuristically, the strategy for mechanistic feature selection in the context of prior knowledge to integrate functional associations derived from various protein structural or unstructured properties, functional genomic and proteomic, or other 'omic' data sets obtained in both humans and model organisms. This integrated network modeling strategy will provide a ranking system to classify potential network components from low to high likelihood; the components will be evaluated genetically, structurally, and functionally.

APPENDICES

Appendix A: Pathway analysis for the inflammation signature genes



An analysis for the enriched biological functions of the top 20 inflammation signature genes shows that 16 out of top 20 genes are involved with inflammatory pathway NFkB and highlighted with brown color.

Appendix B: BDH positive and negative compounds used in qPCR assay development

Animal ID	Compound	BDH Class
1-1	1-napthyl isothiocyanate	positive
1-2	1-napthyl isothiocyanate	positive
1-3	1-napthyl isothiocyanate	positive
2-1	4,4'-methylenediamiline	positive
2-2	4,4'-methylenediamiline	positive
3-1	Carmustine	positive
3-2	Carmustine	positive
3-3	Carmustine	positive
4-1	Carvedilol	positive
4-2	Carvedilol	positive
4-3	Carvedilol	positive
5-1	Lomustine	positive
6-1	Methylpyrilene	positive
6-2	Methylpyrilene	positive
6-3	Methylpyrilene	positive
7-1	Vinblastine	positive
7-2	Vinblastine	positive
7-3	Vinblastine	positive
9-1	Ergocalciferol	negative
9-2	Ergocalciferol	negative
10-1	(+)-Pulegone	negative
10-2	(+)-Pulegone	negative
10-3	(+)-Pulegone	negative
11-1	Auranofin	negative
11-2	Auranofin	negative
11-3	Auranofin	negative
12-1	Diclofenac	negative
12-2	Diclofenac	negative
12-3	Diclofenac	negative
13-1	Ferrocene	negative
13-2	Ferrocene	negative
13-3	Ferrocene	negative
14-1	2-Acetylaminofluorene	negative
14-2	2-Acetylaminofluorene	negative
15-1	Clofibrate	negative
15-2	Clofibrate	negative
15-3	Clofibrate	negative
16-1	Indomethacin	negative
16-2	Indomethacin	negative
16-3	Indomethacin	negative
17-1	N-Nitrosodiethylamine	negative
17-2	N-Nitrosodiethylamine	negative
18-1	Rofecoxib	negative
18-2	Rofecoxib	negative
18-3	Rofecoxib	negative
19-1	Spironolactone	negative
19-2	Spironolactone	negative
19-3	Spironolactone	negative
20-1	CMC	vehicle

20-2	CMC	vehicle
20-3	CMC	vehicle
21-1	Corn oil	vehicle
21-2	Corn oil	vehicle
21-3	Corn oil	vehicle
22-1	water	vehicle
22-2	water	vehicle
22-3	water	vehicle

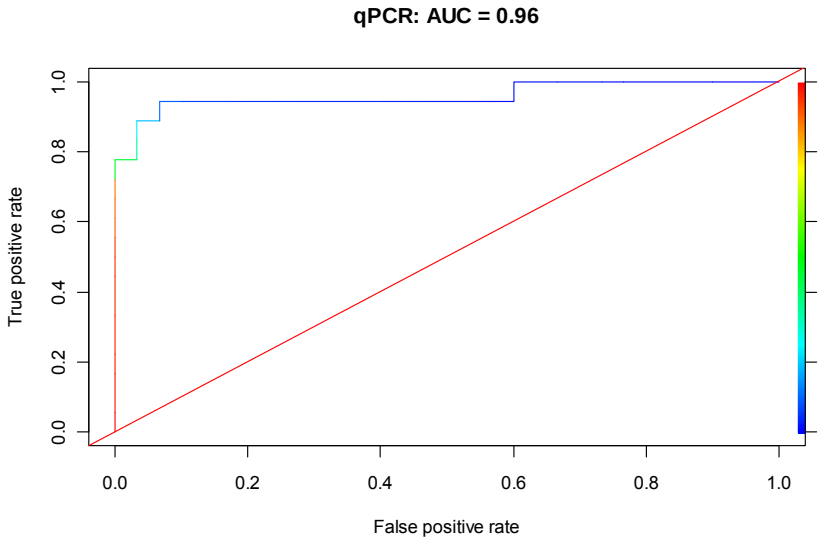
Appendix C: Summary result for the inflammation signature validated with a complete independent data set

TYPE	DATASET	GROUP	NUMP	ERROR	SENS	SPEC	PPV	NPV	AUC
TRAIN	ICONIX	TOP	9	23.3%	74.4%	78.9%	78.0%	75.6%	84.5%
TRAIN	ICONIX	RANDOM	9	32.2%	65.6%	70.0%	69.2%	67.1%	74.3%
TRAIN	ICONIX	BOTTOM	9	30.3%	67.8%	71.7%	70.7%	69.4%	76.4%
TRAIN	ICONIX	WHOLE	31042	11.7%	85.6%	91.1%	90.7%	86.4%	97.3%
TRAIN	ICONIX	SINGLE	9	18.3%	78.3%	85.0%	84.0%	79.8%	90.4%
TRAIN	ICONIX	ICONIX	222	15.3%	81.7%	87.8%	87.2%	83.0%	93.5%
TEST	ICONIX	TOP	9	18.8%	62.5%	82.3%	18.0%	97.3%	80.0%
TEST	ICONIX	RANDOM	9	40.0%	58.3%	60.1%	9.1%	95.8%	59.3%
TEST	ICONIX	BOTTOM	9	44.9%	49.2%	55.5%	6.4%	94.7%	53.7%
TEST	ICONIX	WHOLE	31042	17.6%	65.8%	83.4%	20.2%	97.6%	85.2%
TEST	ICONIX	SINGLE	9	13.1%	76.7%	87.6%	27.9%	98.4%	89.3%
TEST	ICONIX	ICONIX	222	17.9%	70.0%	82.9%	20.7%	97.9%	80.8%
IND	LRLcmpds	TOP	9	22.5%	54.0%	89.2%	72.2%	79.6%	78.6%
IND	LRLcmpds	RANDOM	9	45.3%	44.8%	59.6%	36.1%	68.4%	55.4%
IND	LRLcmpds	BOTTOM	9	49.5%	46.0%	52.8%	33.3%	65.6%	47.4%
IND	LRLcmpds	WHOLE	31042	23.6%	42.4%	93.4%	77.2%	76.5%	75.3%
IND	LRLcmpds	SINGLE	9	25.7%	38.4%	92.2%	71.2%	75.0%	70.2%
IND	LRLcmpds	ICONIX	222	24.9%	46.0%	89.6%	68.6%	77.0%	76.9%

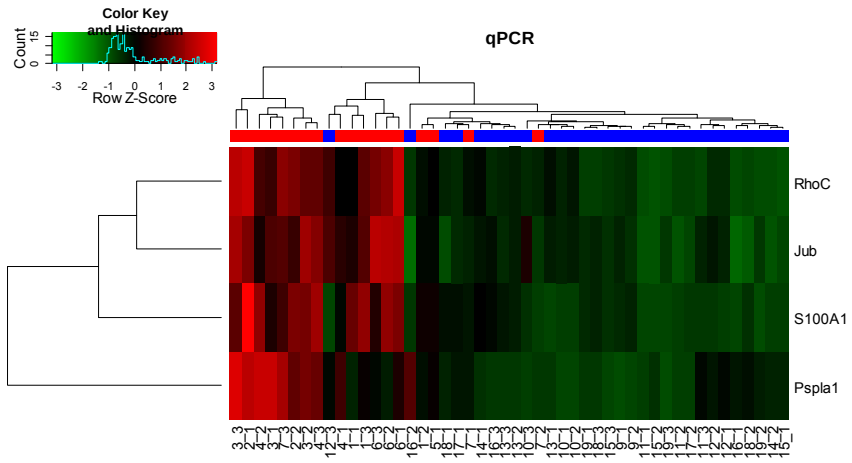
NUMP - number of probes; SENS - sensitivity; SPEC - specificity; PPV - positive predictive value; NPV - negative predictive value; AUC - area under curve

Appendix D: The ROC and heatmap for the BDH signature validated with qPCR

a)



b)



The RNA was extracted from livers of total 48 of animals (rats). 18 animals were observed Histopathologically BDH positive, and 30 animals BDH negative. Quantitative real-time PCR (qRT-PCR) was performed for measuring the expression of the first four genes of BDH signature. The performance assessment for these four gene signature is “Error = 0.104, sensitivity = 0.833, and specificity = 0.933”. a) ROC for First four genes of BDH signature, and the AUC = 0.96; b) a hierarchical clustering for the first four genes of the BDH signature.

Appendix E: The R code for PPEA

```
PPEA<-function (M, class.labels, epv, ratio.tt, threshold.error, num.iteration,
cor.r,num.eva, machine.eva, dir.s, nm.core) {
#The algorithm iteratively applies a two-way bootstrapping over p and n such
#that the sample size is larger than the gene number for each subset. The
#predictive power of individual gene is estimated from #using the prediction
#analysis of microarray (PAM).
#
#
# Inputs:
#       M: Let M_(p×n) be the expression data matrix for p genes and n
#       samples.
#       class.labels:  A vector of binary labels having the 1's and the 0's. The
#       positive class must be labeled as 1s, and the name of class.labels must
#       be matched with sample names of M
#       epv:            events per variable
#       ratio.tt        The ratio for size of training and testing dataset
#       threshold.error       The threshold for error cutoff
#       num.iteration The number of iteration
#       cor.r   The Spearman correlation threshold for the iteration stopping
#       num.eva       The number of iteration for ppea process evaluation
#       machine.eva PAM or SVM
#       dir.s   The directory used for saving the ppea matrix
#       nm.core       The stem of the file names
#
# Outputs:
#       ppea.list       The list contains a ppea matrix at different iteration
#       intervals and corresponding
#       Spearman correlation
#
        #ptm <-proc.time()
        cls.lev<-unique(class.labels)
        cls.p<-class.labels(class.labels==1) # positive class should be "1"
        cls.n<-class.labels(setdiff(names(class.labels), names(cls.p)))

        conf.mat<-NULL
        gene.lst<-NULL
        ppea.mat<-NULL
        cor.lst<-NULL
        gene.wset<-as.vector(row.names(M))
        num.print=1
        for (i in 1:num.iteration){
                ## Randomly select samples for training and testing
                clsp.tr<-sample(cls.p, round(length(cls.p)*ratio.tt))
```

```
clsp.tt<-cls.p((setdiff(names(cls.p), names(clsp.tr))))

clsn.tr<-sample(cls.n, round(length(cls.n)*ratio.tt))
clsn.tt<-cls.n((setdiff(names(cls.n), names(clsn.tr))))

class.tr<-c(clsp.tr,clsn.tr)
class.te<-c(clsp.tt,clsn.tt)

## randomly select the probes
prbs.r<-NULL
prbs.r<-as.vector(sample(row.names(M),
round(length(class.tr)*(1/epv))))
dta.tr<-M(prbs.r,names(class.tr))
dta.te<-M(prbs.r,names(class.te))

trainwts<-100/table(as.vector(class.tr))

# for PAM evaluation
if (machine.eva=="PAM"|machine.eva=="pam"){
        library(pamr)
        library(mda)
        train.dat <-NULL
train.dat <- list(x = as.matrix(dta.tr), y = as.factor(class.tr),
genenames = row.names(dta.tr), geneid = row.names(dta.tr), sampleid =
colnames(dta.tr))
        mod.pam <- pamr.train(train.dat, threshold.scale=trainwts)
        mod.cv <- pamr.cv(mod.pam, train.dat)

        # to find the optimized threshold
        #min0.pos<-which(mod.cv$error==min(mod.cv$error))
        #min.pos<-min(min0.pos)
        #if (mod.cv$size(min.pos)==1){
        #       min.pos=min(min0.pos)
        #       if (mod.cv$size(min.pos)==1){
        #               min.pos=1
        #       }
        #}
        #Delta=mod.cv$threshold(min.pos)
        Delta=0
        g.lst<-NULL
        g.lst<-pamr.listgenes(mod.pam, train.dat, Delta,
genenames = FALSE)
        g.lst<-list(as.vector(g.lst(,"id")))
        names(g.lst)<-"GENE"
        res.pam<-pamr.predict(mod.pam, dta.te, Delta)
```

```
                res.te<-cal_confusion(res.pam, class.te)
                conf.mat<-rbind(conf.mat, res.te)
                gene.lst(i)<-list(g.lst)
        }
        ###
        if (machine.eva=="SVM"|machine.eva=="svm"){
                ##SVM evaluation
                library(e1071)
                dta.tr<-t(dta.tr)
                dta.te<-t(dta.te)
                mod.cv <- svm(dta.tr, as.factor(class.tr), kernel="linear",
cross = 10, na.action = na.omit)
                res.svm <- predict(mod.cv, dta.te)
                res.te<-cal_confusion(res.svm, as.factor(class.te))
                g.lst<-NULL
                g.lst<-list(as.vector(prbs.r))
                conf.mat<-rbind(conf.mat, res.te)
                gene.lst(i)<-list(g.lst)
        }
        ## evaluate the PPEA matrix at different iteration interval ##
        if (i==num.print*num.eva){
                print (paste("i=", i, sep=""))
                tb.tot<-NULL
                tb.tot<-table(unlist(gene.lst))
                iter.pass<-NULL
                iter.pass<-conf.mat(,"Error")<=threshold.error
                tb.suc<-NULL
                        if (is.na(table(iter.pass)("TRUE"))){
                        tb.suc<-as.vector(rep(0, length(unlist(gene.lst))))
                        names(tb.suc)<-names(tb.tot)
                         }else{
                                tb.suc<-table(unlist(gene.lst(iter.pass)))
                        }
                match.tot<-as.vector(rep(0, length(gene.wset)))
                match.tot<-replace(match.tot,
        gene.wset%in%names(tb.tot), tb.tot)
                match.suc<-as.vector(rep(0, length(gene.wset)))
                match.suc<-replace(match.suc,
gene.wset%in%names(tb.suc), tb.suc)
                nm.tot<-paste("TOT.", i, sep="")
                nm.suc<-paste("SUC.", i, sep="")
                nm.ratio<-paste("RATIO.", i, sep="")
                nm.rank<-paste("RANK.", i, sep="")
                ratio.ts<-NULL
                ratio.ts<-match.suc/match.tot
```

```
                        rank.s<-as.vector(rep(0, length(gene.wset)))
                        names(rank.s)<-gene.wset
                        ratio.rank<-replace(ratio.ts, is.nan(ratio.ts),0)
                        rank.s<-replace(rank.s,rev(order(ratio.rank)),
1:length(gene.wset))
                        ppea.s<-NULL
                        ppea.s<-cbind(match.tot, match.suc, ratio.ts, rank.s)
                        colnames(ppea.s)<-as.vector(c(nm.tot, nm.suc, nm.ratio,
nm.rank))
                        row.names(ppea.s)<-gene.wset
                        ppea.mat<-cbind(ppea.mat, ppea.s)
                        if (!is.na(dir.s)){
                        file.o<-paste(dir.s, nm.core, "_ppea_matrix.csv", sep="")
                        write.table(ppea.mat, file=file.o, sep=",", row.names=T)
                        }
                        # Calculate the Spearman correlation
                        if (num.print>1){
                        nm.last<-paste("RANK.", num.print*num.eva, sep="")
                        nm.pre<-paste("RANK.", (num.print-1)*num.eva, sep="")
                        cor.s<-NULL
        cor.s<cor(as.vector(ppea.mat(,nm.last)),as.vector(ppea.mat(,nm.pre)),
use = "everything", method = c( "spearman"))
                        names(cor.s)<-nm.last
                        cor.lst<-c(cor.lst, cor.s)
                        print(paste("The Spearman correlation is ", cor.s, sep=""))
                        if(cor.s>cor.r){
                        stop("The number of iterations has reached the stable
stage")
                        }
                        print (paste("i=", i, sep=""))
                        }
                        num.print<-num.print+1
                }
        }
        ppea.list<-list(PPEA=ppea.mat, Correlation=cor.lst)
        #proc.time() - ptm
        return(ppea.list)
}

## construct the confusion matrix
cal_confusion<-function(c.pred, cls){
        # positve has to be "1"
        # to force the class be '1' and '0'
        conf.mat<-NULL
        cls<-replace(cls, as.vector(cls)!=1, 0)
```

```
        grp.s<-NULL
        grp.s<-split(c.pred, cls)
        TP<-NULL
        FN<-NULL
        tb.p<-NULL
        tb.p<-table(grp.s("1"))
        TP<-as.vector(tb.p("1"))
        FN<-as.vector(tb.p("0"))
        TN<-NULL
        FP<-NULL
        tb.n<-NULL
        tb.n<-table(grp.s("0"))
        TN<-as.vector(tb.n("0"))
        FP<-as.vector(tb.n("1"))
        sens<-TP/(TP+FN)
        spec<-TN/(TN+FP)
        err<-1-(TP+TN)/(TP+TN+FP+FN)
        ppv<-TP/(TP+FP)
        npv<-TN/(TN+FN)
        conf.mat<-NULL
        conf.mat<-as.vector(c(TP, FN, FP, TN, sens, spec, err, ppv, npv))
        names(conf.mat)<-as.vector(c("TP", "FN", "FP", "TN", "Sensitivity",
"Specificity", "Error", "PPV",
"NPV"))
        return(conf.mat)
}
```

CURRICULUM VITAE

Jiangang Liu

Education

Aug 2005 - Dec 2010, Ph. D. in Bioinformatics at Indiana University, Indianapolis, IN

Sep 2003 - Jun 2005, M. S. in Bioinformatics at Indiana University, Indianapolis, IN

Sep 1987 - Jun 1990, M. D. in Sport Medicine and Molecular Biology at Beijing Medical University, Beijing, China

Sep 1979 - Jun 1982, B. S. in Medicine at Shaoyang Medical College, Hunan, China

Professional Experience

Jan 1999 - Present, Computational Biologist in Bioinformatics Group, Eli Lilly and Company, Indianapolis, IN

Jan 1995 - Dec 1999, Research Associate in Department of Cell Biology, University of Alabama at Birmingham

Jul 1982 - Aug 1987, Surgeon, Department of General Surgery, Shanmin Hospital, Hunan, China

Selected Publications

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK (2006) Intrinsic disorder in transcription factors. Biochemistry 45:6873-6888.

Liu J,Campen A, Huang S, Peng SB, Ye X, Palakal M, Dunker AK, Xia U, Shuyu Li (2008) Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. BMC Medical Genomics 1:39

Liu J, Jolly RA, Thomas CE, Stevens, JL, Ryan TP, Watson DE, Searfoss GH, Goldstein KM, Dunker AK, Li D, Wei T (2010) Identification of Toxicogenomic Biomarkers by Development and Application of a New Mining Algorithm, In preparation

Liu J, Jolly RA, Thomas CE, Dunker AK, Li D, Wei T (2010) Relating ALT to necrosis in rat liver, in preparation