# BIOINFORMATICS ANALYSIS OF OMICS DATA TOWARDS CANCER DIAGNOSIS AND PROGNOSIS

by

Jianjun Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2007

Doctoral Committee:

Professor Arul M. Chinnaiyan, Chair
Professor David G Beer
Associate Professor Jill A. Macoska
Associate Professor Kerby A. Shedden
Associate Professor Debashis Ghosh, Pennsylvania State University

# ACKNOWLEDGEMENTS

I would first like to thank my mentor, Dr. Arul M. Chinnaiyan, for his expert guidance, support, encouragement, and inspiration. I would also like to thank Dr. Debashis Ghosh for his continuous statistical support and great advice, Dr. David G. Beer, Dr. Jill A. Macoska, and Dr. Kerby A. Shedden for serving on my Doctoral committee and giving me valuable suggestions on this thesis work.

I would like to thank Jindan Yu, Xiaoju Wang, Guoan Chen, Saravana Dhanasekaran, Daniel Rhodes, Scott A. Tomlins, and Sooryanarayana Varambally, who have contributed to most of the work described here. I would like to express my gratitude to all the members in the Chinnaiyan lab for their support. Without them, none of the work described here could have been completed. I would also like to thank William P. Worzel and Arpit A. Almal for their support on genetic programming project.

I would like to express my deepest gratitude to my wife and my love, Yipin, without whom I would be nowhere. Thanks for putting up with my late nights and giving me unconditional love and encouragement through my Doctoral study and the writing of this work. I would also like to thank my parents, my sister, and my grandparents for giving constant support and love. And last but not least, I would like to thank my friends and all whose support helped me completing this thesis in time.

## TABLE OF CONTENTS

**PART 1: INTRODUCTION**

**PART 2: GENE EXPRESSION-BASED CANCER DIAGNOSIS AND PROGNOSIS**

**PART 3: NON-INVASIVE TESTS FOR PROSTATE CANCER DIAGNOSIS**

**PART 4: AN INTEGRATIVE APPRAOCH TO MODEL PROSTATE CANCER PROGRESSION**

**PART 5: CONCLUSION**

## LIST OF FIGURES

Figure

# LIST OF TABLES

Table

# PART 1: INTRODUCTION

## CHAPTER 1

### DNA Microarray in Cancer Research

Having the essentially complete sequence of the human genome is similar to having all the pages of a manual needed to make the human body. The challenge to researchers and scientists in current post-genomic era is to find those missing definitions, to use genomic structural information to display and analyze biological processes on a genome-wide scale, and to explore gene functions. Although traditional molecular biology often takes a reductionist approach to biological questions, it has long been recognized that genes act in concert with other partners, often in separate dimensions of time and space. To fully understand the underlying biology, these molecular interactions may have to be studied from the conceptual framework of the entire genome. With the advent of high throughput technologies, the logistical challenge of which approach is being overcome. DNA microarray, as one of high-throughput technology, offers the ability to measure the expression of thousands of genes simultaneously, providing genome-wide views of gene expression from the yeast cells to human cancer cells. By supplying quantitative information on cell transcriptomes, this technology has become a powerful tool in biomedical research, especially in cancer research and has led to an explosion in global gene expression profiling studies. In theory, for human cancer, this knowledge has the potential to lead to optimized, individualized management of patients. The goal of this survey is to summarize the basic principles, the different steps involved in DNA microarray technology and, more importantly, the proper data analysis methodologies and applications in cancer biology.

In general, DNA microarrays are of two kinds, depending on the materials arrayed: cDNA microarrays (1) and oligonucleotide arrays (2) and. cDNA arrays are generated with a robotic arrayer, printing a double-stranded cDNA onto a solid surface such as glass, charged nylon membranes, or nitrocellulose filters. For oligonucleotide arrays, in situ synthesis usually produces short 20-25 mers by photolithography (Affymetrix) or lengths of up to 60 nt by inkjet technology (Agilent Technologies). Another type of oligonucleotide array is made by spotting longer presynthesized oligonucleotides (~70 nt) on glass slides (3). While cDNA is easily customizable, oligonucleotides generally offer greater specificity as they can be specifically tailored to minimize chances of cross-hybridization. Other advantages of oligonucleotide arrays include uniformity of probe length and the ability to distinguish specific variants (4).

The hybridization of a test sample to an array can be detected in one of two ways (**Figure 1.1**). cDNA arrays are commonly designed with 2-dye (also known as 2-channel) representing cDNAs from experimental and reference RNA samples experiments. Each cDNA sample is labeled with its own specific fluorophore. Expression values are reported as ratios between two fluorescent values, representing the quantitative difference between two cDNA sources. Alternatively, the Affymetrix-like oligonucleotide arrays use a single channel system to detect absolute level of gene expression. In addition, Affymetrix designs “perfect match” (PM) and “mismatch” (MM) probes for each probe set, where the “mismatch” probe is one nucleotide different from the “match” probe and is intended to indicate the level of non-specific binding. However, whether or not to use “mismatch” probes is still an open question. For example, MAS5, the default Affymetrix probe set algorithm, utilizes both PM and MM probe information, while RMA, an algorithm developed by UC Berkeley ignores MM probes.

Pat Brown and his colleagues at Stanford University published the first microarray paper using cDNA arrays in 1995 (1). In general, a typical cDNA microarray experiment proceeds as follows: (1) sample preparation and RNA isolation, (2) preparation of fluorescently labeled cDNA, (3) hybridization, (4) slide scanning, image and data analysis (**Figure 1.1**). During the experiment, multiple sources of variation may be introduced including: mixed cellular composition in tissue, genetic heterogeneity within tissue cells, difference in sample preparation, non-specific cross-hybridization of spots, different detection efficiencies for the fluorescent labels as well as differences between individual slides (4). Thus care must be taken in each step. Once slides are scanned, image analysis is first carried out to determine image quality, identify spots and output background adjusted fluorescent intensities. A good image shall have a low level, uniform background and high signal-to-noise ratio. For spot identification, most commercial scanners provide software to transform the colored spots into numerical intensities. During spot identification, background signal is also estimated. The most common method is to calculate the background signal locally in the vicinity of each spot and then subtract it. The ratios of measured background-subtracted Cy5 to Cy3 intensities are further subjected to normalization that is intended to remove systematic errors arising from the above variations.

Two major normalization steps are commonly used in the literature: (1) within-array normalization, (2) between-array normalization. For within-array normalization, a simple approach is global mean-centered or median-centered based on the assumption that the total integrated intensity across all spots in the microarray should be equal for both channels because of the equal amount of RNA used for labeling from each sample. Other often used methods include normalization against a subset of housekeeping genes, global loess, print-tip loess, robust spline or even 2D-loess to address more profound spatial biases (5,6). To adjust differences between arrays, further between-array normalization may be required. Common methods include global scale normalization, quantile normalization or variance stabilizing normalization (vsn) (5,7,8). As Affymetrix oligonucleotide arrays use a single-channel system, the normalization steps are slightly different. The first step for affymetrix oligonucleotide arrays is to estimate the absolute intensity of individual probe sets which represent primarily annotated transcripts. Each probe set is typically represented by a set of 11–20 PM and MM probe pairs. This multiple probe feature allows for more robust background assessments and gene expression measures, and has facilitated the development of computational or statistical methods to translate image data into a single normalized "signal" for mRNA transcript abundance. To date, there are many probe set algorithms that have been developed, with a gradual movement away from chip-by-chip methods (MAS5), to project-based model-fitting methods (dCHIP, RMA, GC-RMA etc.). However, it is debating that which one is the rational best method (9,10). The best probe set algorithm may vary from project to project (9). After calculating summary probe set intensity, further between-array normalization may be needed for some probe set algorithms, for example, MAS5.

The normalized microarray expression values are typically log2-transformed, and stored as a two-dimensional table, with genes in the rows and profiles in the columns. As thousands of data points may be observed per array, microarray data is often grouped and visualized as a heatmap. Each data point can be presented as a color that quantitatively and qualitatively reflects its relative expression within the data. For example, high expression is presented as red while low expression is presented by green; and the intensity of the color represents the degree to which the gene is expressed. To better present such heatmaps, an unsupervised hierarchical clustering is typically adopted along with visualization (11). It orders genes or samples based on their similarity of expression. For example, one could cluster the samples in a collection of cancer patient cohort into subgroups based on the similarity of their aggregate expression profiles. On the other hand, genes that share similar patterns of expression in a biological context could be also clustered together. As such a method does not require any priori assumption, it has the advantage of being unbiased. More importantly, it allows one to detect the inherent patterns hidden in a complex dataset. A success example is given in a breast cancer profiling reported by David Botstein and Patrick Brown at Stanford University (12) . By employing a two-way hierarchical clustering, they grouped both genes and samples based on the similar patterns of gene expression profile, leading to the discovery of molecular subtypes of breast cancer. Clustering of the tumors based on overall expression profiles firstly divided the samples into two distinct clusters. One cluster of tumors shared relatively high expression of a set of genes expressed in ER+ tumors or breast luminal cells, thus being defined as ER+/luminal subtype. The other group of tumor samples having relatively low expression of these genes, were further sub-categorized into basal cell-like, Erb-B2+, and normal-like subtype, each subtype with a characteristic gene expression signature. A follow-up study showed that the ER+/luminal type could be further divided into at least two subgroups with different clinical outcome (13).

In addition to unsupervised learning, one common analysis is to perform supervised learning analysis, incorporating the prior knowledge of sample information. A typical schema for microarray data analysis is to select a subset of genes that can best distinguish two classes of training samples such as disease vs. healthy controls and build a computation or statistical model that is able to classify training samples as well as predict independent, blinded test samples into these classes. Such supervised analyses are particularly useful for cancer diagnosis and prognosis. However, they are relying on accurate sample information, which may be an issue in cancer given the limitations of current histopathologic accuracy.

A significant effort has been put forth to apply microarray technique to the study of cancer (12,14-17), both due to the complexity and heterogeneity of the disease and the lack of efficient clinical diagnostic tools and treatments. Cancer can be considered a genetic disease, occurring as a result of the progressive accumulation of genetic alterations in somatic cells. Because hundreds of genes may be simultaneously involved in the mechanisms of tumor formation, high-throughput DNA microarray is in particular useful to screen a large number of genes and thus identify potential interesting marker genes. Presently, microarrays have been extensively used in cancer research for several applications, including (but not limited to) the following: (1) discovery of novel cancer diagnostic and prognostic biomarkers; (2) identification of novel target genes for oncogenes or tumor suppresser genes; (3) molecular class discovery, classification and prediction; (4) identification of genes associated with drug resistance and prediction of clinical response to drug. In the next section, we will review the main strategies thus far employed in microarray gene expression profiling studies, as well as the significant results obtained from them.

A common goal for cancer microarray profiling is to identify genes differentially expressed between two groups of samples, e.g. benign and tumor tissue. Many statistical tests can be used to determine the significance of difference of gene expression between two groups. Some common tests include student's t-test, signal to noise ratio, permutation test and significance of microarray analysis (SAM). However, as thousands of genes are being tested simultaneously, the chance of false positive rate is increased. Thus, there is a need to adjust for multiple hypothesis testing. The most often used procedures to control false positive rate are estimation of family-wise error rate developed by Westfall and Young (18), and false discovery rate introduced by Benjamini and Hochberg (19,20). More details of feature gene selection are discussed in **Chapter 2**.

There have been many successes in using gene expression profiling to identify markers of diagnostic and prognostic value. Our previous study in prostate cancer is a good example (15). Prostate cancer is the most frequently diagnosed cancer in American men. Screening for prostate-specific antigen (PSA) has led to earlier detection of prostate cancer, but elevated serum PSA levels are present in non-malignant conditions such as benign prostatic hyperlasia (BPH). cDNA microarrays have been used to examine gene-expression profiles of more than 50 normal and neoplastic prostate specimens and three common prostate-cancer cell lines. Statistical testing was used to sort differentially expressed genes between the sample groups. The highest scoring genes were then subjected to independent tissue microarrays for validation at the protein level. Hepsin, a transmembrane serine protease, was found to be highly expressed at the mRNA level and protein level in nearly all of the cancer samples, but not in the benign samples, suggesting its role as a novel biomarker for prostate cancer. Another paradigm in biomarker discovery is alpha-methylacyl CoA racemase (AMACR) (21-24). Luo et al. (2002) and our group reported simultaneously that AMACR is a novel tissue biomarker for prostate cancer by cDNA microarrays and independent tissue microarrays. Our group further demonstrated that the humoral immune response against AMACR was more sensitive and specific than PSA (a clinical prostate cancer marker) in distinguishing sera from prostate cancer patients to control subjects (24).

Several groups have used DNA microarrays for classifying tumors from benign tissues or distinguishing tumor subtypes on the basis of certain discriminant function. **Chapter 2** interrogated a wide range of molecular classification methods in detail. The first proof-of-principle for microarray-based histological classification was reported by Golub et al (16) in 1999. This study demonstrated the feasibility of using expression profiling for cancer diagnosis. Using unsupervised learning on oligonucleotide microarrays, leukemia samples are neatly clustered into known acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) solely based on gene expression. Supervised learning demonstrated that a set of 40 genes that are differentially expressed in AML and ALL could accurately predict a group of unknown samples into correct categories, again solely based on gene expression profile. Although this distinction can be detected using modern histopathology and cell surface phenotypes, this study has established a paradigm that tumor expression profiling can be used for cancer classification. More recently, Armstrong et al. (25) identified mixed-lineage leukemia (MLL), a new molecular subtype of leukemia with a decidedly unfavorable prognosis. MLL arises from a chromosomal translocation involving the mixed-lineage leukemia gene and has typically been classified with ALL. This study showed that MLL has a unique gene expression profile distinct from AML and ALL, demonstrating that the differences in gene expression are robust enough to classify disease subtypes.

One of major obstacles to cancer therapy is the development of drug resistance. Cancers may be either primarily resistant to the treatment or develop resistance during the process of treatment. Multiple mechanisms of drug resistance have been reported and drug resistance is likely to involve a diversity group of genetic factors such as tumor suppresser genes, growth factor receptors, DNA repair factors and cell death regulators. Presently, it is difficult to predict whether chemotherapy will be effective for individual patients. By using cDNA microarrays, Kudoh et al. (26) monitored the expression profiles of Doxorubicin-induced and Doxorubicin resistant cancer cells. A subset of the

Doxorubicin-induced genes was found to be constitutively over-expressed in cells selected for resistance to doxorubicin and may represent the signature profile of doxorubicin resistance phenotype. This study demonstrated the feasibility of obtaining potential molecular profile or fingerprint of anticancer drugs in cancer cells by DNA microarray, which might yield further insights into the mechanisms of drug resistance and suggest alternative methods of treatment.

Gene expression profiling of tumors has been also used for outcome prediction. Investigators have demonstrated the utility of using pretreatment gene expression profiling to determine prognosis. In a retrospective study of 38 patients with diffuse large B-cell lymphoma (DLBCL), Alizadeh et al. (14) firstly demonstrated expression-based correlates of outcome. They clustered cDNA microarray data and defined two molecularly distinct forms of DLBCL. By examining patient survival, they found that the defined germinal center B-like (GCB) DLBCL had a significantly better overall survival than those with activated B-like (AB) DLBCL. Van't Veer et al. (27) have also reported the use of gene expression profiling to develop an outcome predictor for breast cancer metastasis. Primary breast tumors from patients who developed distant metastases within 5 years were compared with tumors from patients who continued to be disease-free after a period of at least 5 years.  Supervised classification was used to identify a set of 70 genes strongly predictive of a short interval to distant metastases. In a follow-up study (28), by using the previously established 70-gene prognosis profile, they classified a series of 295 primary breast carcinomas as having a gene-expression signature associated with either a poor prognosis or a good prognosis. Among the 295 patients, 180 had a poor-prognosis signature and 115 had a good-prognosis signature. These two groups showed markedly different outcome (10-year distant metastasis-free survival, 50.6% vs. 85.2%). They also demonstrated that the prognosis profile could add value to existing clinical and histological criteria.

To date, hundreds of mRNA expression profile studies of various cancers have been reported in the literature and a large number of datasets have been made available (**Figure 1.2**). This tremendous resource would speed up the identification of robust cancer markers as well as facilitate the development of improved molecular signatures if it could be properly and fully utilized. However, due to the lack of a unifying bioinformatic resource, the majority of these data sit stagnant and disjointed following publication, massively underutilized by the cancer research community. While standards and repositories have begun to be established, the full potential of cancer microarray data will only be reached when it is unified, logically analyzed, and easily accessible. To this end, our lab initiated an effort to systematically curate, analyze and make available all public cancer microarray data via a web-based database and data-mining platform, designated '*ONCOMINE*' (http://www.oncomine.org) (29). Besides data collection, our effort also includes centralizing gene annotation data from various genome resources to facilitate rapid interpretation of a gene's potential role in cancer. Furthermore, we have integrated microarray data analysis with other resources including gene ontology annotations and a therapeutic target database so that clinically interesting subsets of genes can be focused on. Currently the ONCOMINE database houses 310 independent datasets comprising over 500 million gene expression measurements from nearly 22,000 microarray experiments. By making these resources easily accessible to public, we hope that this work could benefit the identification of potential cancer markers, maximize the utility of data, promote an increase in validation performance, and ultimately lead to the improved understanding of cancer and the development of novel diagnostic and therapeutic strategies.

As noise is known significant in DNA microarrays due to genomic variations, experimental artifacts, sampling bias, and cross-hybridization so on, there is high demand to validate potential cancer markers or gene signatures in independent datasets or through independent experimental techniques. While it is common to use the microarray as a screening tool and then to validate a few promising candidates using such as reverse transcriptase polymerase chain reaction (RT-PCR), or tissue microarrays, it may under-utilize the microarray dataset and overlook other potential markers. With the increasing number of publicly available gene expression datasets, meta-analysis in combining multiple studies to determine the repeatability of one microarray result becomes a promising method for in silico validation. For example, our previous study (22) demonstrates a statistical model for performing meta-analysis of independent microarray datasets. Instead of using the actual expression measurements which may be complicated due to distinct microarray technologies, the model utilizes statistic *p*-values derived from individual studies. Differential expression was first assessed independently for each gene in each dataset based on a *p*-value. Then individual study *p*-values were combined using a result that -2log(*p*-value) has a chi-squared distribution under the null hypothesis of no differential expression. The model was first implemented on four publicly available prostate cancer gene expression data sets that compared the gene expression profiles of clinically localized prostate cancer to that of benign prostate tissue with the goal of identifying genes differentially expressed between the two groups. The analysis revealed that four prostate cancer gene expression datasets shared significantly similar results, independent of hybridization platforms, demonstrating that combining *p*-values is useful to obtain more precise estimates of significances. Based on this statistical framework for inter-study validation, our lab has extended the approach to a large compendium of public cancer microarray datasets in a follow-up study (30). We characterized a common transcriptional profile that is universally activated in most cancer types relative to the normal tissues from which they arose, likely reflecting essential transcriptional features of neoplastic transformation (**Figure 1.3**). In addition, a meta-signature of undifferentiated cancer was also uncovered, consisting of 69 genes that were over-expressed in undifferentiated cancer relative to well differentiated cancer, suggesting common molecular mechanisms by which cancer cells progress and dedifferentiate.

While the above studies highlight the use of expression profiling for addressing important questions in clinical oncology and demonstrate the potential of DNA microarrays in clinic, many challenges remain. The first challenge lies in microarray assay development and standardization. Microarray technology is known to be susceptible to measurement error due to a long and convoluted chain of decisions on sampling, preprocessing, hybridization, calibration, and analysis. Errors and biases may involve the sampling of the specimens, their quality, the amount of tissue obtained, storage, fixation, plating, and readout of microchips (31). The analytical calibration and informatics analysis plan can also be very convoluted. Major decisions need to be made for transformation, normalization, data filtering, removal of technical artifacts, and background correction. For each decision node, there are numerous possibilities, and so far there is no standard informatics platform available. Another challenge lies in gene annotation. For a given probe, there is some uncertainty to map to the correct target gene due to non-specific probe design, cross-hybridization or transcript splice variants of same gene; some probes may actually represent a different gene than advertised. In addition, DNA microarrays measure gene expression at the mRNA level, while gene products function at the protein level.  Some inconsistence may exist between mRNA and protein level expression. An mRNA can be alternatively spliced prior to translation and eventually yield different proteins. Additionally, various post-translational modifications may occur in proteins. Another important challenge lies in the availability of sufficient numbers of samples. Until now, gene expression profiling has depended mostly on small numbers of clinical specimens. Validation of claims has been uncommon, fragmented, and incomplete (31). A final challenge relates to the integration of data sets from different laboratories using different profiling technologies. While it is surely best to use these multiple datasets to validate one another so that the most promising candidate biomarkers can be identified, this task is challenging because microarray data exists on a variety of scales depending on the specific technological platform utilized as well as the experimental procedure. Although there are some successes to integrate different datasets to date, more sophisticated methods are required for efficient data comparison and integration.

In conclusion, DNA microarray is an invaluable and promising technology. Developing molecular diagnostic tools by tumor gene expression profiling is conceivable. Although many challenges remain ahead, identifying novel molecular targets and classifying novel molecular subtypes of cancer on the basis of DNA microarray data may facilitate the development of new cancer drug, the design of clinical trials, and the planning of cancer therapy.

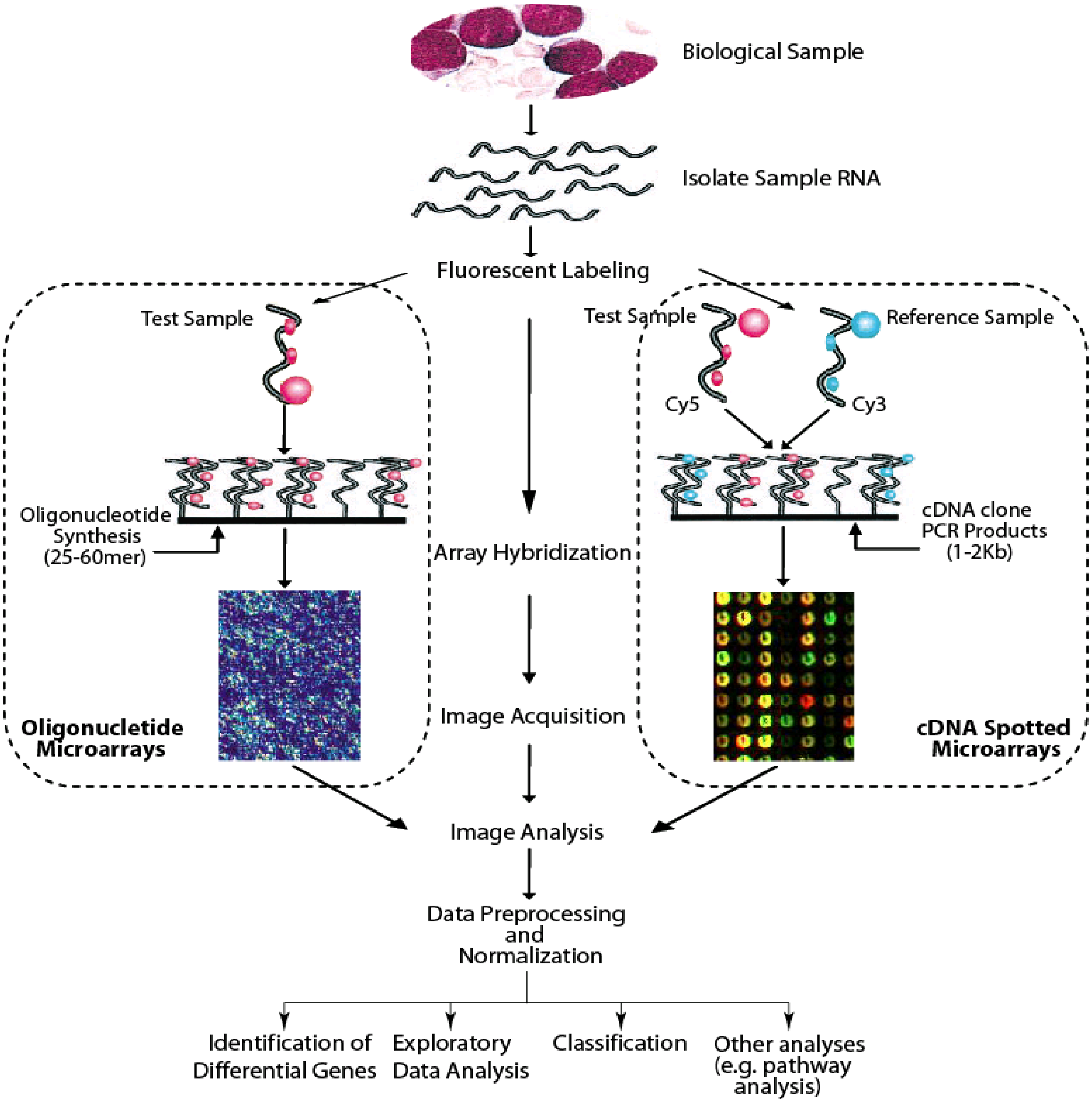

Biological Sample
Isolate Sample RNA
Fluorescent Labeling
Test Sample
Oligonucleotide Synthesis (25-60mer)
Oligonucletide Microarrays
Array Hybridization
Image Acquisition
Test Sample
Reference Sample
Cy5
Cy3
cDNA clone PCR Products (1-2Kb)
cDNA Spotted Microarrays
Image Analysis
Data Preprocessing and Normalization
Identification of Differential Genes
Exploratory Data Analysis
Classification
Other analyses (e.g. pathway analysis)

**Figure 1.1**. Experimental workflow of a typical microarray using either oligonucletide or cDNA spotted array technique. [Adapted from Ramaswamy et al., Journal of Clinical Oncology. 2002. 20(7):1932-1941]

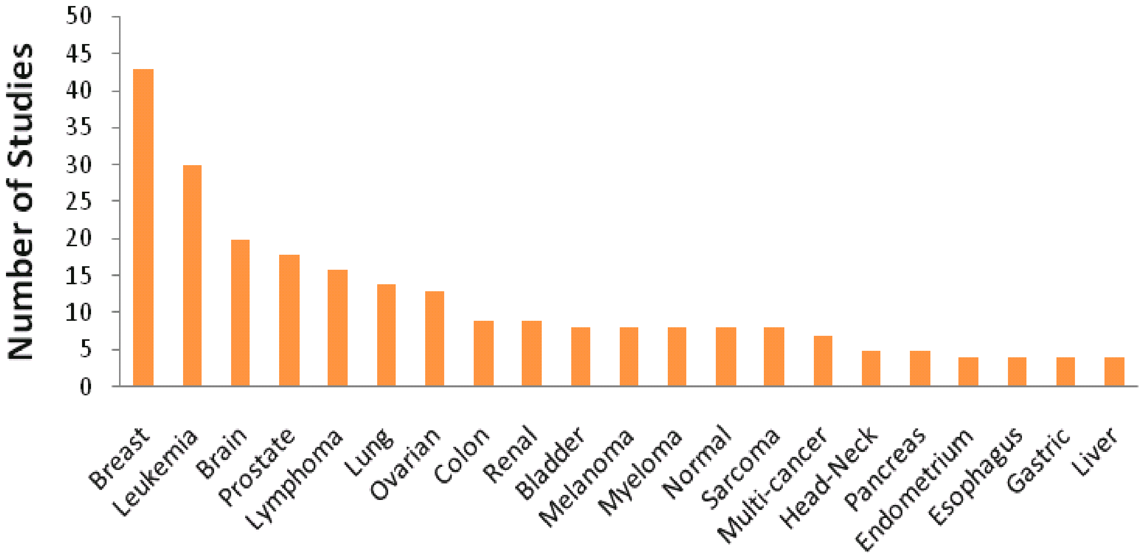


**Figure 1.2**. Number of expression profiling studies carried out for individual cancer type as of early 2007 in ONCOMINE database (http://www.oncomine.org).

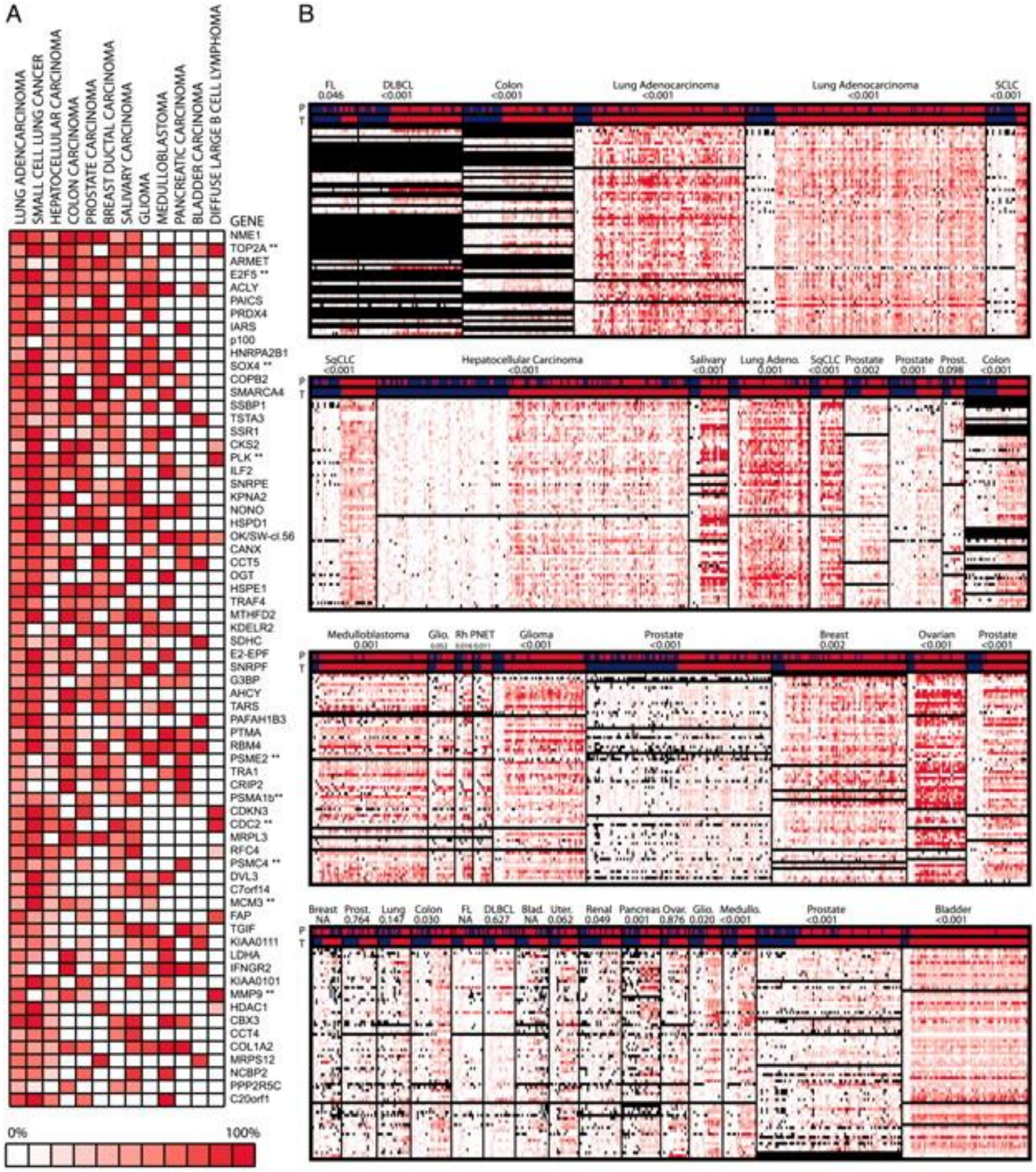


**Figure 1.3.** Meta-signature of neoplastic transformation. (A) Sixty-seven genes overexpressed in cancer relative to normal tissue counterpart in at least a dozen ‘‘cancer vs. normal’’ signatures from independent microarray studies. White boxes signify either not present or not significant. Red boxes signify significant overexpression in cancer relative to normal tissue, the shade of red indicating the percentage of cancer samples that had an expression value greater than the 90th percentile of normal samples. (B) The signature was able to significantly predict ‘‘cancer vs. normal’’ status in 32 of 39 analyses. The two bars above each heat map represent the predicted class (P) and the true class (T): red signifies cancer and blue signifies normal tissue. In the color maps, black signifies data not available, white signifies less than or equal to the normal class mean expression level, and red signifies the degree of over-expression relative to the mean normal class expression level.

# REFERENCES

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

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

3. Arcellana-Panlilio, M, and Robbins, SM (2002). Cutting-edge technology. I. Global gene expression profiling using DNA microarrays. *Am J Physiol Gastrointest Liver Physiol* **282**, G397-402.

4. Ramaswamy, S, and Golub, TR (2002). DNA microarrays in clinical oncology. *J Clin Oncol* **20**, 1932-1941.

5. Smyth, GK, and Speed, T (2003). Normalization of cDNA microarray data. *Methods* **31**, 265-273.

6. Xiao, Y, Hunt, CA, Segal, MR, and Yang, YH (2004). Novel stepwise normalization method for two-channel cDNA microarrays. *Conf Proc IEEE Eng Med Biol Soc* **4**, 2921-2924.

7. Huber, W, von Heydebreck, A, Sultmann, H, Poustka, A, and Vingron, M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104.

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

9. Seo, J, and Hoffman, EP (2006). Probe set algorithms: is there a rational best bet? *BMC Bioinformatics* **7**, 395.

10. Shedden, K, Chen, W, Kuick, R, et al. (2005). Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics* **6**, 26.

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

12. Perou, CM, Sorlie, T, Eisen, MB, et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747-752.

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

14. Alizadeh, AA, Eisen, MB, Davis, RE, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

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

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

17. Hedenfalk, I, Duggan, D, Chen, Y, et al. (2001). Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344**, 539-548.

18. Westfall, PH, and Young, SS. (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment: Wiley).

19. Benjamini, Y, and Hochberg, Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B* **57**, 289-300.

20. Storey, JD, and Tibshirani, R (2003). Statistical methods for identifying differentially expressed genes in DNA microarrays. *Methods Mol Biol* **224**, 149-157.

21. Luo, J, Zha, S, Gage, WR, et al. (2002). Alpha-methylacyl-CoA racemase: a new molecular marker for prostate cancer. *Cancer Res* **62**, 2220-2226.

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

23. Rubin, MA, Zhou, M, Dhanasekaran, SM, et al. (2002). alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *Jama* **287**, 1662-1670.

24. Sreekumar, A, Laxman, B, Rhodes, DR, et al. (2004). Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer. *J Natl Cancer Inst* **96**, 834-843.

25. Armstrong, SA, Staunton, JE, Silverman, LB, et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30**, 41-47.

26. Kudoh, K, Ramanna, M, Ravatn, R, et al. (2000). Monitoring the expression profiles of doxorubicin-induced and doxorubicin-resistant cancer cells by cDNA microarray. *Cancer Res* **60**, 4161-4166.

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

28. van de Vijver, MJ, He, YD, van't Veer, LJ, et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009.

29. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

30. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* **101**, 9309-9314.

31. Ioannidis, JP (2007). Is molecular profiling ready for use in clinical decision making? *Oncologist* **12**, 301-311.

# CHAPTER 2

## Towards Cancer Classification Using Gene Expression Data

One challenge of cancer treatment is to develop specific therapies for pathogenetically distinct tumor types, to maximize efficacy and minimize toxicity. Cancer classification and class discovery have thus been central to advances in cancer treatment (1). Previously, cancer classification has primarily been on the basis of morphological and clinical characteristics of the tumor. However, these traditional methods have been reported to have serious limitations (1). Tumors with similar histopathological appearance can be molecularly heterogeneous, differently responsive to particular therapy, and thus may require different clinical courses (1). To gain a better insight into this issue, demand on developing more systematic approaches to examine global gene expression has been on the rise. The recent advent of microarray technology has made it straightforward to simultaneously monitor the expression patterns of thousands of genes. Although still in its early stage of development, current successes have indicated its promising future.

To date, various statistical or machine learning techniques have been proposed for molecular cancer classification. In this survey report, a comprehensive overview of current cancer classification methods will be presented. Due to the high-dimensional nature of gene expression data, we will also summary the prevailing feature gene selection methods as it is an integrated part for molecular classification. Finally, we will discuss several challenges related to cancer classification and present solutions. A typical workflow of molecular cancer classification can be seen in **Figure 2.1**.

Gene Selection

Different from traditional data used for classification, gene expression data has several unique characteristics as follows: high dimensionality, small sample size and a large number of redundant and irrelevant genes. Gene expression data sets usually contain thousands to tens of thousands of genes. However, a majority of genes do not have expression change between cancer classes. In addition, many genes are redundant and highly correlated. Further, current gene expression data sets in the literature have relative small set of samples (often less than 100). With such a huge dimension space, it appears easy for classic statistical or computational methods to over-fit the data. Moreover, inclusion of a large number of irrelevant genes not only increases the computation time, but also introduces noise and confuses the classifiers. A common way to deal with this issue is to perform gene selection prior to classification in the literature in order to improve the performance of classifiers, reduce computational running time, and facilitate post-classification analysis for biological insights of genes involved in the classification.

The most commonly used gene selection approach is individual gene ranking based on some correlation measuring criteria. Each gene is ranked by its correlation with the class labels and the top ones are selected. Conventional statistical methods for individual gene selection include student's t-test, Wilcox rank sum test, logistic regression, and Pearson correlation so on. Golub et al. (1) proposed a correlation metric measuring the relative correlation between the expression values of a gene and the class labels, termed signal-to-noise (S2N). For a gene with two classes (e.g., Class 1 vs. Class 2), the signal-to-noise statistic is $(\mu_{Class1}-\mu_{Class2})/(\sigma_{Class1}+\sigma_{Class2})$ where $\mu$ and $\sigma$ are the mean and standard deviation of the expression for the gene. This method favors genes that have large between-class mean difference and small within-class variation. Comparing to t statistic, this method penalizes genes that have higher variance in each class more than those genes that have a high variance in one class and a low variance in another. Similarly, in order to penalize genes with small standard errors, several attempts have been made based on an *ad hoc* fix by simply adding a constant to the observed standard error:

$$\tilde{t}_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{se_i + s_0}$$ , where $s_0$ is a fudge factor.

Efron et al. (2) simply computed $s_0$ as the 90th percentile of the $se_i$'s. In the SAM method developed by Tibshirani group at Stanford University (3), $s_0$ is determined based on minimizing the coefficient of variation of $\tilde{t}_i$ as a function $se_i$.

Alternatively, Smyth (4) developed a hierarchical model and derived an empirical Bayes estimate for the gene-wise variance. This empirical Bayes approach is equivalent to shrinkage of the estimated sample variances towards a pooled estimate, resulting in far more stable inference when the sample size is small. Similarly, Baldi et al. (5) developed a regularized t-test that uses a Bayesian estimate of the variance among gene measurements within an experiment for the identification of statistically significantly differential genes.

While genes selected by the above methods are highly correlated with the class individually, combining them together may not give the best discrimination power as these methods lack the capability of exploiting correlation and interaction among genes. In addition, these methods may include redundant genes. For example, genes regulated in the same pathway may be included as they contain similar high correlation information with the class labels.  Moreover, these methods may not be able to detect genes that are complement to each together and contribute to the classification while individual of them does not exhibit high correlation with class labels. This is common as tumor heterogeneity has been observed in many cancers and some genes may be dys-regulated only in a subtype of cancer. One approach to overcome these barriers is to find a group of genes that serve together to maximize the classification accuracy. One can monitor the change on the expected value of error when one gene is removed. The expected value of error is the error rate computed on an infinite number of examples, which can be approximated by a cost function computed on the training samples given a training set. Guyon et al. (6) proposed a recursive feature elimination (RFE) approach to perform gene selection using support vector machine. The basic idea is to apply the SVM classification algorithm on the training data, compute the change in cost function for the removal of each gene, find the gene that minimizes the cost function change after its removal, and then remove that gene and repeat the entire procedure. Finally, a ranked feature list is generated. The subset of genes that are top ranked (eliminated last) together yields the optimal classification performance, but those genes are not necessarily the ones that are individually most relevant (6) as the method has no effect on correlation metrics. The authors found that the SVM-RFE method worked better for cancer classification than the individual gene ranking approach and was able to select genes that are directly related with cancer, whereas the other method tends to pick up genes that are differential because of the different cell compositions in two classes of tissues.

## Cancer Classification Method

While we believe that a fair amount of attention should be paid to gene selection as an integral preprocessing step, the central role of cancer classification is to develop classification methods to accurately classify cancer classes. One promising use of DNA microarray data in cancer classification is to accurately determine individual patient's diagnostic and prognostic status based on his/her individual genomic profile, eventually leading to personalized cancer therapy for individual patients. Typically, a classifier, which consists of a set of discriminant functions, will be built on a "training" set, and then evaluated on an independent "test" dataset that does not participate in the development of the classifier. To date, a wide range of supervised classification methods have been developed for gene expression data sets. Golub and his co-workers have pioneered a molecular classification approach for gene expression-based histological classification (1). They selected 50 "informative" genes based on signal-to-noise ratio and proposed a weighted voting method to classify acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL). They have demonstrated that AML and ALL can be accurately distinguished solely based gene expression values without previous knowledge of these classes.

Khan et al. (2001) (7) first attempted to classify four types of the small, round blue-cell tumors (SRBCTs) that share similar histopathological characteristics to specific diagnostic categories based on their gene expression signatures. They performed dimension reduction on the full set of gene expression data by using Principle Component Analysis. The first ten components were then trained in artificial neural networks (ANNs). The ANNs correctly classified all samples and identified the genes most relevant to the classification based on measuring the sensitivity of the classification to a change in the expression level of each gene. To test the ability of the trained ANN models to recognize SRBCTs, the authors analyzed additional blinded samples, and correctly classified them in all cases. The study have successfully demonstrated the potential applications of gene expression-based classification methods to classify histopathologically similar cancers.

Decision tree, also known as classification trees, is a widely used classification method. The construction of the decision tree involves two phases: the growing phase and the pruning phase. In the growing phase, a decision tree is built from the training data. The purity-based entropy function selects the best gene at each internal node to split the data set into subsets that minimizes the misclassification error. In the pruning phase, the tree is pruned using some heuristics to avoid overfitting of data and increase the generality of the classifier. Using a public colon cancer data set, Zhang et al. (2001) (8)

introduced a recursive partitioning classification method based on classification tree and demonstrated its high accuracy for discriminating among distinct colon cancer tissues with a cross validation misclassification rate of 6-8%. In an extended study (9), in order to improve classification and prediction accuracy, the authors proposed a deterministic procedure to form forests of classification trees. When two published and commonly used data sets are used, they found that the deterministic forests performed far better than the single trees.

Some similarity-based classification methods have been also applied for molecular cancer classification. One simple yet common method is Nearest Neighbor (NN) or its variant, *K*-Nearest Neighbor (*K*NN). Briefly, for each testing sample *s*, its class label is determined by the training sample whose expression profiling is most similar to *s*, according to certain distance measure. The distance measure can be any similarity/dissimilarity matrix such as Pearson correlation, Euclidean distance, Manhattan distance etc. If using *K*NN ($K$>1), the class label of *s* is assigned using majority vote from *K* training samples with highest similarity to *s*. Utilizing three public cancer gene expression data sets, Dudoit et al. (10) compared the performance of different classification methods including *K*NN, linear discriminant analysis, classification trees and more recent aggregating classifiers. They found that the nearest-neighbor, diagonal linear discriminant analysis (DLDA) in general had the smallest misclassification rates, whereas fisher linear discriminant analysis (FLDA) had the highest.

From a different point of view, one may consider the training process of a classifier as a process to find a hyperplane that best separates the training samples into different groups according to their classes. The best hyperplane could be the one with maximum margin, where margin is defined as the distance from a hyperplane to the sets of data points that are closest to it. Such a hyperplane is more robust and may less prone to change when given a slightly different training set. One of max-margin classification algorithms is Support Vector Machine (SVM), which has been widely used in data mining applications including molecular classification based on gene expression data (11-14). Mukherjee et al. (11) first demonstrated that SVM yielded superior performance for gene expression-based classification tasks. Ramaswamy et al. (12) extended SVM method to solve multiclass problems by employing a simple one-versus-all technique. Guyon et al. (6) proposed a SVM-RFE method to perform gene-selection, and recent study (13) extended it to be MSVM-RFE for multiclass gene selection. The ability of SVM for producing hyperplane with maximized margin and for tuning the amount of training errors has made SVM especially suitable for the gene expression data classification (15).

Another popular similarity-based classification method in molecular cancer classification is nearest shrunken centroid method (PAM) developed by Tibshirani et al. (16). One major modification to standard nearest centroid classification is that it "shrinks" each of the class centroids toward the overall centroid for all classes by an amount termed as the threshold. This shrinkage has led to two advantages: 1) more accuracy on classification by reducing the effect of noisy genes, 2) automatic gene selection. In a comparison of PAM to six other classification algorithms including SVM, *K*NN, DLDA, and RandomForest, the authors have observed that PAM in overall has the lowest average error rate and is just slightly behind SVM in average rank. In a recent extended study (17), the authors introduced a modified version of linear discriminant analysis, termed the "shrunken centroids regularized discriminant analysis" (SCRDA). They have claimed that this method often outperforms the PAM method and can be as competitive as the support vector machines classifiers.

Recently, one class of machine learning technique, Evolutionary Algorithm (EA) has also been introduced to cancer classification on gene expression data (18-23). For example, researchers including our lab have demonstrated one EA approach, genetic programming (GP) could be a promising approach for discovering comprehensible rule-based classifiers from gene expression profiling data (19,20,22,24). Our lab applied GP to cancer gene expression data to select feature genes and develop molecular classifiers (24). By examining GP on one Small Round Blue Cell Tumors (SRBCTs), one lung adenocarcinoma and five prostate cancer datasets, we have found that GP classifiers, which often comprise five or less genes, successfully predicted cancer classes. Further, we have demonstrated that GP classifiers remain predictive ability on independent datasets across microarray platforms.

Gene expression profiles have been also used to predict disease or treatment outcome of patients. Van 't Veer et al. (25) compared primary breast tumors from patients who developed distant metastases within 5 years to tumors from patients who continued to be disease-free after a period of at least 5 years.  Correlation-based supervised classification successfully identified a set of 70 genes with an expression signature strongly predictive of a short interval to distant metastases. Beer et al. (26) developed a compound covariate predictor and generated a risk index based on the top 50 genes which identified low-risk and high-risk stage I lung adenocarcinomas with significantly different outcome. Other common methods include semi-supervised principle components (27), penalized Cox regression (28), and threshold gradient descent method (29) etc.

## Challenges in Cancer Classification

Although results of molecular classification obtained thus far seem promising, there are still considerable challenges. In this section, we discuss some important issues in cancer classification and review current solutions thus far. However, these questions are still open. Further research is needed to fully address these issues.

The first challenge lies in the unique characteristic of high-throughput data: the huge dimensionality and high co-linearity. High-throughput data such as DNA microarray usually contain a large number of genes yet relatively small sample size. Such data disable application of standard discrimination methods. For example multivariate logistic regression, can not be directly applied to obtain the parameter estimates on gene expression data. Presently, the prevailing strategies include pre-filtering by gene selection as described in the previous section, performing dimension deduction, or using regularized statistical models. One way to achieve dimension reduction is to transform the large number of genes to a new set of variables which are uncorrelated and ordered such that the first few account for most of the variation in the data. Principle component analysis (PCA) is one of well known methods. It transforms the original variables (genes) to a new set of predictor variables, which are linear combinations of the original variables. In mathematical terms, PCA sequentially maximizes the variance of the original data. Khan et al (7) applied PCA to SRBCT gene expression data and used the first 10 principle components to train a neural network. Other dimension deduction methods include singular value decomposition (SVD), the partial least squares (PLS) and sliced inverse regression (SIR) so on. One major disadvantage of these dimension deduction methods is the loss of gene information as the followed classification algorithm is developed solely upon the new variables. Interestingly, some researchers utilized these dimension-reduction methods to remove highly correlated genes in a gene predictor where individual gene selection may be carried out first to form the predictor (27,30). A large portion of genes selected by individual gene ranking are often redundant or highly correlated. In term of classification accuracy, it is thus necessary to remove such genes as they do not contribute much towards the performance of a classifier although they may be important in biological relevance. For example, such groups of genes may reflect an essential de-regulated pathway for the cancer progression.

In addition to dimension deduction, the other approach is to use the regularized estimation methods. A common regularization is to add a penalty function to a multivariate partial likelihood in order to stabilize the parameter estimates. Commonly employed penalty functions include $L_2$ and $L_1$ penalizations. For example, classical linear regression with $L_2$ penalty is known as “ridge regression”(31). Li et al. (32) was the first to investigate $L_2$ penalized estimation of the Cox model in the high-dimension and low-sample size settings and applied their method to gene expression profile for censored patient outcome. One limitation of $L_2$ penalization is that it uses all the genes in the prediction and does not provide a way of select relevant genes. An alternative is to use $L_1$ penalized estimation, which was proposed by Tibshirani et al. (33) and was called the least absolute shrinkage and selection operator (Lasso). Using newly developed least angle regression (LARS) by Efron et al. (34), Gui et al. (28) proposed an efficient way to estimate $L_1$ penalized Cox regression model , termed LARS-Lasso. Friedman et al. (35) have recently proposed a step-wise optimization method termed threshold gradient descent (TGD) and demonstrated its application in classification problems. Interestingly, they showed that with different threshold value, TGD can approximate the estimates of partial least square, ridge regression, Lasso and LARS. Gui et al. (29) further extended the TGD method to the Cox regression model for selecting genes that are associated with patient survival and building a predictive model to predict the risk of a future patient.

It has to be anticipated that in situations where the number of genes exceeds by far the number of samples in the data, the overfitting of naively applied statistical strategies and resulting over-optimism of the prediction error may be overwhelming. This leads to another challenge in molecular classification about how to estimate unbiased prediction error rate. The standard practice for performance validation is to use a set of samples as a training set for the development of the prediction model and use a completely independent set of samples for estimating the prediction error. However, it is rare to obtain readily available sufficiently large numbers of specimens that are amenable to microarray analysis and accompanied by the necessary clinical information. Thus the sample size of a typical microarray study is usually less than 100. To maximize the utility of samples, gene expression profile studies generally estimate prediction accuracy using the same data by proper application of resampling methods such as cross-validation or bootstrapping. These methods use the data efficiently and are almost unbiased when used correctly. However, it does have significant variance when used with small sample size and can be subject to bias if used naively. For example, an inappropriate usage of cross-validation may lead to two types of biases: selection bias and optimization bias (parameter selection bias). A proper cross-validation leaves out a single 'test fold' of the data, selects the model, variables and parameters solely based on the remaining 'training folds' and then evaluates the misclassification rate on the test fold. When averaged over folds, this should provide a nearly unbiased estimate of the true misclassification rate of the classifier. A 'selection' bias' can occur when a subset of variables are selected based on all the available data and then the error rate is estimated by cross-validation using this fixed set of variables. On the other hand, an 'optimization bias' may occur if cross-validation is used to estimate the error rate for multiple sets of free parameters, and then the set of parameter values with the lowest estimated error rate is chosen for the final classifier (36). This happens because the same data is used to both select a set of parameters and to estimate the error rate. To deal with this issue, a separate two-level cross-validation is needed to estimate its error rate. As suggested by Wood et al (36), two-level cross-validation should be used as follows. Assumed $K$ fold cross validation is used, at the top level, one of $K_1$ folds of data is left out for the purpose of assessing the error rate of the finished classifier. At the lower level, $K_2$-fold cross-validation is then performed on the remaining data to select the optimal value of any free parameters. When all parameters are selected, the classifier can be tested on the left-out fold at the top level. By repeating this for all $K_1$ folds at the top level, one can generate a cross-validated assessment of the cross-validated choice. To simplify the procedure, one may select $K_2 = K_1$-1, so that the same fold structure can be used for both levels.

Even perfect and complete cross-validation may suffer from unknown external validity, making molecular classifiers difficult to move towards clinical practice. One challenge lies in the limited sample size of the data. The profiled set of samples may not represent the general populations in clinic. Thus, in order to truly assess the classifier performance, one may perform a completely independent validation which may include but not limited to different data samples from independent disease centers, same protocols and same definition of analytical end point, and independent testing by independent research investigators so on. Another challenge is the fact that the genes selected for each proposed profile may be not stable. Different splits of the training and validation data may result in very different sets of genes being selected. Some genes may be valid only in the reported dataset. Ein-Dor et al. (37) have reported that thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. This is about 100-fold larger than the sample sizes currently being used to date. Thus new methods may be needed in selecting the robust genes relevant to cancer classification.

One may question why not to estimate prediction accuracy using existing external datasets as there are numerous gene expression profiles reported in the literature. In theory, this is a most rigorous validation. However, one challenge lies in the lack of a unified microarray platform and standardized procedure to integrate public datasets. The challenge is two-folds. First, there is no unified microarray platform thus far. For an instance, Affymetrix platforms measure the absolute mRNA expression levels of individual genes of individual samples while cDNA microarrays measure the relative mRNA expression change between two samples. Different reference samples may be used for cDNA arrays on different studies. In addition, datasets from different platforms may contain different sets of genes. To address these issues, efficient methods for data transformation and gene annotation need to be developed. Second, even when a unified platform is used for gene expression profiling, systematic biases may still exist between datasets from different laboratories. Non-standardization of data may introduce noise and error into the classification accuracy. Special care must be taken to inter-study classification. A simple approach to normalize inter-study datasets is to standardize individual genes within each dataset with zero mean and unit variance after between-array normalization. This procedure transforms and makes same genes of different datasets at the same location and scale. In practice, this is similar to calculate relative gene expression levels of individual sample to a reference sample where each gene's expression level equals mean expression value of the gene across all samples in the dataset. This method was used in **Chapter 3** to validate a breast cancer outcome signature developed from a training set on multiple independent datasets. The signature successfully dichotomized the patients in individual datasets into high-risk and low-risk groups with strongly different outcome. However, this method may not work well on datasets with small sample sizes. Warnat et al. (38) used median rank scores and quantile discretization to derive numerically comparable measures of gene expression from different platforms. The basic idea of this method is to transform gene expression values of different microarray platforms to a common numerical range by replacing numerical values of one study by numerical values from the other study, with respect to the relative ranks of expression values within each study. Our lab also developed a data integration method based on *poe* (probability of expression) transformation (39). The *poe* model transformed the raw gene expression data into signed probability of differential expression for each gene in each sample, thus providing a unified measure across studies. The platform-free scaleless property of this model is particularly useful for data integration in the domain of gene expression profiling. Further, the transformation improved contrast in each data set by removing the influence of extreme expression values. Following this *poe* model, we combined multiple breast cancer studies (n = 305 samples) and developed a 90-gene meta-signature, which demonstrated strong association with survival in breast cancer patients.  A more advanced method was proposed by Benito et al. (40) for the identification and adjustment of systematic biases present within microarray data sets. They presented a new approach, called 'Distance Weighted Discrimination (DWD)', to adjust system biases in microarray datasets. The new method was shown to be very effective in removing systematic biases present in published breast tumor cDNA microarray data sets and could be used to merge multiple breast tumor data sets completed on different microarray platforms.

Taken together, in this chapter, we provided a comprehensive survey on the existing cancer classification methods. As an important step of classification, feature gene selection was presented in detail as well. Molecular classification based on gene expression profiling has been rapidly evolving from an interesting scientific concept to a clinical tool in the last decade. It provides a more systematical and unbiased way for cancer diagnosis and prognosis. Through this survey, we conclude that, although the progress of molecular classification varies for different cancers and there are still a great amount of work that needs to be further addressed, the results obtained so far is promising and the future is possibly fascinating.

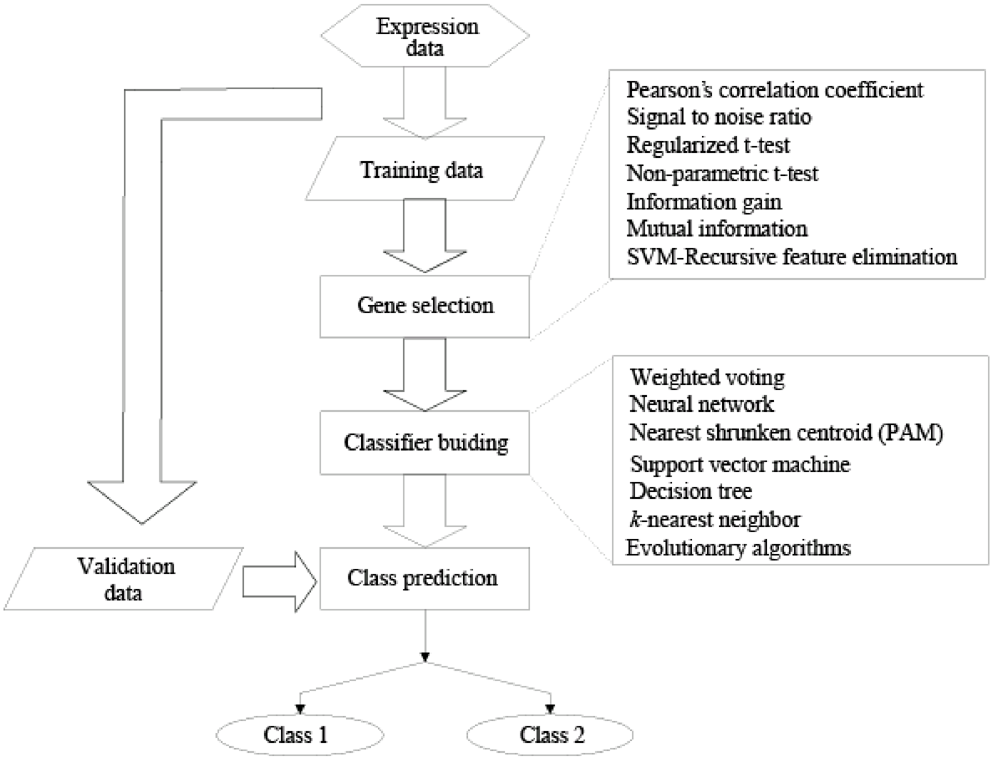


**Figure 2.1**.A typical workflow of molecular cancer classification and prediction model.

## REFERENCES

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

2. Efron, B, Tibshirani, R, Storey, JD, and Tusher, V (2001). Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Soc* **96**, 1151-1160.

3. Tusher, VG, Tibshirani, R, and Chu, G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121.

4. Smyth, GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3.

5. Baldi, P, and Long, AD (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-519.

6. Guyon, I, Weston, J, Barnhill, S, and Vapnik, V (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389-422.

7. Khan, J, Wei, JS, Ringner, M, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7**, 673-679.

8. Zhang, H, Yu, CY, Singer, B, and Xiong, M (2001). Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci U S A* **98**, 6730-6735.

9. Zhang, H, Yu, CY, and Singer, B (2003). Cell and tumor classification using gene expression data: construction of forests. *Proc Natl Acad Sci U S A* **100**, 4168-4172.

10. Dudoit, S, Fridlyand, J, and Speed, TP (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association* **97**, 77--87.

11. Mukherjee, S, Tamayo, P, Mesirov, JP, et al. (1999). Support vector machine classification of microarray data: MIT, CBCL).

12. Ramaswamy, S, Tamayo, P, Rifkin, R, et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* **98**, 15149-15154.

13. Zhou, X, and Tuck, DP (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics* **23**, 1106-1114.

14. Brown, MP, Grundy, WN, Lin, D, et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A* **97**, 262-267.

15. Lu, Y, and Han, J (2003). Cancer classification using gene expression data. *Information Systems* **28**, 243-268.

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

17. Guo, Y, Hastie, T, and Tibshirani, R (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86-100.

18. Ho, SY, Hsieh, CH, Chen, HM, and Huang, HL (2006). Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems* **85**, 165-176.

19. Hong, JH, and Cho, SB (2006). The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. *Artif Intell Med* **36**, 43-58.

20. Langdon, WB, Buxton, B.F. (2004). Genetic Programming for Mining DNA Chip Data from Cancer Patients. *Genetic Programming and Evolvable Machines* **5**, 251-257.

21. Li, L, Weinberg, CR, Darden, TA, and Pedersen, LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **17**, 1131-1142.

22. Mitra, AP, Almal, AA, George, B, et al. (2006). The use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. *BMC Cancer* **6**, 159.

23. Ooi, CH, and Tan, P (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* **19**, 37-44.

24. Yu, J, Yu, J, Almal, AA, et al. (2007). Feature selection and molecular classification of cancer using genetic programming. *Neoplasia* **9**, 292-303.

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

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

27. Bair, E, and Tibshirani, R (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, E108.

28. Gui, J, and Li, H (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001-3008.

29. Gui, J, and Li, H (2005). Threshold Gradient Descent Method for Censored Data Regression with Applications in Pharmacogenomics. *Pro. Pac. Symp. Biocomput.* **10**, 272-283.

30. Shen, R, Ghosh, D, Chinnaiyan, A, and Meng, Z (2006). Eigengene-based linear discriminant model for tumor classification using gene expression microarray data. *Bioinformatics* **22**, 2635-2642.

31. Hoerl, AE, and Kennard, RW (1970). Ridge regression:Biased estimation for Anonorthogonal problems. *Technornetrics* **12**, 55-67.

32. Li, H, and Luan, Y (2003). Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data. *Pro. Pac. Symp. Biocomput.* **8**, 65-76.

33. Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.* **58**, 267-288.

34. Efron, B, Hastie, T, Johnstone, I, and Tibshirani, R (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

35. Friedman, JH, and Popescu, BE (2004). Gradient Directed Regularization for Linear Regression and Classification. Technical report, Stanford University.

36. Wood, IA, Visscher, PM, and Mengersen, KL (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics* **23**, 1363-1370.

37. Ein-Dor, L, Zuk, O, and Domany, E (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* **103**, 5923-5928.

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

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

40. Benito, M, Parker, J, Du, Q, et al. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105-114.

# PART 2: GENE EXPRESSION-BASED CANCER DIAGNOSIS AND PROGNOSIS

## CHAPTER 3

### A Transcriptional Fingerprint of Estrogen in Human Breast Cancer Predicts Patient Survival

Estrogen signaling plays an essential role in breast cancer progression, and Estrogen Receptor (ER) status has long been a marker of hormone responsiveness. However, ER status alone has been an incomplete predictor of endocrine therapy, as some ER+ tumors, nevertheless, have poor prognosis. Here we sought to use expression profiling of ER+ breast cancer cells to screen for a robust estrogen-regulated gene signature that may serve as a better indicator of cancer outcome. We identified 532 estrogen-induced genes and further developed a 73-gene signature that best separated a training set of 286 primary breast carcinomas into prognostic subtypes by step-wise cross-validation. Notably, this signature predicts clinical outcome in over ten patient cohorts as well as their respective ER+ sub-cohorts. Further, this signature separates patients who have received endocrine therapy into 2 prognostic subgroups, suggesting its specificity as a measure of estrogen signaling, and thus hormone sensitivity. The 73-gene signature also provides additional predictive value for patient survival, independent of other clinical parameters, and outperforms other previously reported molecular outcome signatures. Taken together, these data demonstrate the power of using cell culture systems to screen for robust gene signatures of clinical relevance.

Breast cancer is the most common type of cancer among women in the industrialized world, accounting for nearly 1 of every 3 cancers diagnosed. Estrogen is essential for the normal growth and differentiation of the mammary gland, and plays a critical role in the pathogenesis and progression of breast cancer (1). Increased lifetime exposure to estrogen is a well-known factor for increased breast cancer risk (1), and drugs that block the effects of estrogen has been used to inhibit the growth of hormone-dependent breast cancers (2). In the last few decades, systemic adjuvant therapy to patients with predicted poor prognosis has significantly increased breast cancer survival (3). Current prognostic markers for breast cancer include tumor stage, size, histological grade, and estrogen receptor status. However, approximately 1 out of 4 patients diagnosed with breast cancer nevertheless die from the disease (4), indicating the insufficiency of current prognostic biomarkers. In addition, a large number of patients with ER-positive tumors failed on endocrine therapy, suggesting the need of more precise biomarkers of therapy prediction.

Taking advantage of global expression profiling, molecular predictors have been developed to classify and predict patient prognosis (5-10). This prognostication of breast cancer outcome may be used for the selection of high-risk patient for adjuvant therapy. Transcriptional changes of these predictor genes are presumed to reflect the activity of essential signaling pathways in tumors and thus greatly increase the prediction power.

For example, the expression of prostate-specific-antigen (PSA) indicates the activation of androgen receptor and serves as a much better diagnostic/prognostic biomarker of prostate cancer than androgen receptor itself. Similarly, for several decades ER status has been used as a marker of hormone responsiveness to guide adjuvant therapy, with ER+ tumors having significantly better clinical outcome (11). Some ER+ tumors, nevertheless, incur disease recurrence, indicating that ER status alone is an incomplete assessor and additional biomarkers are required. A transcriptional fingerprint of estrogen may better reflect the activity of estrogen signaling, thus being a more definitive predictor of breast cancer recurrence and patients' response to hormonal therapy.

In this study, we attempted to delineate downstream effector genes of estrogen signaling. We hypothesized that these genes may indicate an activated state of estrogen receptor, and thus predict cancer outcome and hormone responsiveness. To identify robust estrogen-regulated genes, we employed three ER+, estrogen-responsive breast cancer cell lines, MCF-7, T47D and BT-474. We stimulated these cells with 17β-estradiol to emulate the transcriptional events induced by estrogen signaling *in vivo*. To ensure that we capture the transcriptional changes due to direct regulation by estrogen, rather than downstream effects, we focused primarily on early time-points (0, 1, 2, 4, 8, 12 and 24hrs) following estrogen stimulation (12). By a time-course analysis on expression profiling of these cell lines, we identified 532 estrogen-induced probe sets, representing 446 unique genes (FDR<0.01, see Methods and **Figure 3.1a**).

Several lines of evidence support that the genes we selected represent a true downstream transcriptional network of estrogen signaling. Firstly, a subset of these genes, including PGR, PDZK1, CTSD, MYC, MYB, MYBL1, MYBL2, STK6, Ki-67 and GREB1, have been previously confirmed to be induced by estrogen (13-15). Secondly, Molecular Concept Map (MCM) analysis (16), which allows for the identification of molecular correlates of our gene set, revealed significant enrichment of 'up-regulated by estrogen treatment' signatures (*P*-values<=0.001, Odds ratios >=4.35) previously identified by several independent groups (17-19) (**Figure 3.1b**). To evaluate the biological relevance of our gene set *in vivo*¸ MCM analysis of cancer profiling concepts found strong enrichment of 'over-expressed in ER+ breast cancer' concepts derived from a number of human breast cancer profiling studies executed by independent investigators (5,8,10,20). Therefore, our estrogen-regulated gene set is relevant to previously identified gene sets of estrogen regulation reported from both *in vitro* cell line experiments and i*n vivo* tumor profiling. Interestingly, integrative analysis with a public genome-wide location data of ER occupancy (21) showed that a highly significant portion (*P* <0.00001) of our estrogen-induced genes are direct targets of ER, suggesting that our gene set may represent the direct transcriptional network evoked by activated ER.

To obtain an overall annotation of our estrogen-regulated genes, we performed MCM analysis on Gene Ontology (GO) concepts. Significantly enriched gene ontology (GO) categories include "DNA replication", "regulation of cell cycle", "protein folding", "tRNA processing", "cytokinesis", "DNA replication", and "DNA repair" (**Figure 3.1c**).

This result is consistent with previously reported functions of estrogen-regulated genes (13,22).

Intriguingly, another distinct interaction network revealed by MCM analysis enriched in the 'over-expressed in high grade breast cancer' signatures from various datasets such as the Miller et al. (5), Sotirious et al (20), and van de Veer et al. (9) datasets (**Figure 3.1d**). Notably, this enrichment network also includes several concepts of 'over-expressed in metastasis, dead or recurrent breast cancers', suggesting a link between our gene signature and breast cancer outcome. Thus, we next attempted to confirm this survival association using breast cancer expression profiling datasets. We performed k-mean clustering (k=2) with Pearson correlation distance of 286 node-negative primary breast carcinomas (10). Kaplan-Meier (KM) survival analysis revealed that the resulted two clusters differed significantly in patient outcome ($P$ = 0.002). The "high-risk" group with poorer outcome has higher expression of several known ER targets (13,15), such as MYBL1, MYBL2, MKI67, and MCM2. By contrast, "good-outcome" genes that are over-expressed in the "low-risk" group include PGR, CD44, ADD1, and PTGER3.

To develop an optimal outcome predictor using top survival-related genes, we ranked the 532 estrogen-regulated genes by their corresponding survival significance and performed step-wise cross-validation. Our results demonstrated a set of top-ranked 73 genes (**Table 3.1**) that yielded optimal survival association with the least cross-validation error (**Figure 3.2a**). This 73-gene signature successfully dichotomized the 286 training samples into high-risk and low-risk groups with significantly different outcome (*P* < 0.00001, **Figure 3.2b**). Importantly, by performing 1000 Monte Carlo simulations we found that the probability for a randomly selected subset of 73 genes to cluster the same samples with equivalent or better significance was less than 0.001, re-affirming that the performance of our 73-gene signature could not be achieved by chance.

To validate the prediction power of our 73-gene signature, we collected all public breast carcinoma datasets (n=11) with available patient survival information from ONCOMINE (23) database. The 73-gene signature was then applied to predict individual samples within each dataset into either "high-risk" or "low-risk" group using nearest centroid classification. Strikingly, in 10 out of these 11 datasets KM survival analysis revealed a remarkable outcome difference between the predicted "high-risk" and "low-risk" groups (**Figure 3.3a-j**). For the only dataset wherein our outcome signature failed to predict, it revealed a marginally significant (log-rank *P* = 0.15, **Figure 3.3k**) association with distance metastasis within 5 years. To the best of our knowledge, this is the first study thus far that reports a breast cancer outcome predictor which is validated extensively in such many independent patient cohorts.

We observed that our gene signature correctly predicted most ER- breast tumors within individual datasets as "high-risk". As a subset of ER+ tumors relapses regardless of standard anti-hormone therapy, they may as well have poor prognosis. It is therefore important to identify these patients for more effective adjuvant therapies. We thus examined the ability of our predictor in stratifying the ER+ tumors into prognostic subgroups. We have taken the ER+ samples from each dataset and carried out KM survival analysis for the predicted “high-risk” and “low-risk” groups by the 73-gene signature. Notably, KM survival analysis demonstrated a strong discriminative power of our 73-gene signature in distinguishing ER+ patients with different prognoses (**Figure 3.4**).

Prognostication of breast cancer outcome may guide the respective selection of patients at high risk for systemic adjuvant therapy. However, there is no guarantee that these selected patients will actually benefit from the therapy. It is therefore of important clinical value to predict therapy responsiveness and to spare some patients from unnecessary adjuvant therapies which have side effects that may cause more harm than good. For example, endocrine therapy may be sufficient for some node-positive and ER-positive patients, and more aggressive adjuvant therapy may not additionally help these patients. Out of the 11 datasets we analyzed above, four contained patient treatment information. We extracted hormone-treated samples from each dataset and assessed whether our gene predictor was able to predict patient response to hormonal therapies. Again, we predicted the hormone-treated samples into “high-risk” and “low-risk” groups. Importantly, in each cohort we observed significantly different outcome for the two predicted groups, suggesting an ability of our signature in therapy prediction (**Figure 3.3j, Figure 3.5a-c**).

To further confirm the association of our gene signature with estrogen sensitivity, we determined whether the 73-gene signature is able to classify ER+/ER- cell lines *in* *vitro*. We performed hierarchical clustering based on the expression pattern of the 73 genes in 5 ER- and 3 ER+ cell lines. Interestingly, we found that the 73 genes perfectly separated the 8 cell lines into their respective ER+ and ER-clusters, demonstrating that our signature genes are specific to estrogen signaling. Furthermore, as we selected our estrogen-induced genes based on expression induction at relative early time points (no later than 24hrs) following 17β-estradiol treatment, we hypothesized that this subset of 73 genes is also enriched for direct targets of ER. Concordantly, comparative analysis with ER-occupied genes described in a previous study (21) identified a significant overlap (*P*=0.0001), re-confirming the specificity of our signature to estrogen activity.

As estrogen may also play an important role in the development of glioma (24) and lung cancer, especially lung adenocarcinoma (25,26), we examined our outcome signature in 3 glioma and 1 lung adenocarcinoma datasets. Notably, our gene signature successfully predicted patient outcome, with *P*=0.0006 for the Freije et al. Glioma (27), *P*=0.008 for the Phillips et al. Glioma (28), *P*=0.11 for the Nutt et al. Glioma (29) and *P*=0.006 for the Bhattacharjee et al. lung adenocarcinoms (30) dataset (**Figure 3.5d-j**).

Global gene expression profiling of breast cancer has yielded a number of prognostic signatures in the last decade. To properly evaluate the predictive power of our signature, we compared it with established clinical parameters as well as previously reported gene predictors. We first compared our signature with an 822-gene estrogen-regulated signature (termed as "estrogen-SAM") developed by Oh and the co-workers (14) based on SAM analysis that classified the ER+ cases of the Rosetta data set (n=225) into prognostic subtypes (8). We selected the Rosetta data as the test dataset since it has been routinely used as a validation dataset for breast cancer outcome signatures. Multivariate Cox proportional-hazards regression analysis of these patients showed that both our signature and the estrogen-SAM signature were significant predictors for relapse-free survival (RFS), independent of standard clinical factors (RFS *P*=0.002 and *P*=0.004 respectively, **Table 3.2**). Importantly, our outcome signature was by far the strongest predictor for both relapse-free and overall survival (OS) (RFS *P*=0.002, Hazard ratio [HR]: 2.24, 95% Confidence Interval [CI]: 1.35-3.70; and OS *P*=0.001, HR: 3.27, 95% CI: 1.62-6.62). Thus, our outcome signature achieved better predictive power while using substantially fewer genes. In addition, our signature comprised solely of estrogen-regulated genes, thus representing the biological significance of estrogen activity. By contrast, the estrogen-SAM signature genes were selected based on their differential expression between two tumor subtypes predefined by estrogen-regulated genes, and hence may or may not themselves be regulated by estrogen.

We next extended the comparison of our signature and the estrogen-SAM signature to the Rosetta 70-gene signature as well using the Rosetta data set. As the Rosetta signature utilized a subset of 44 samples during its development, to avoid potential bias these samples were excluded from our analysis. Importantly, our signature and the Rosetta 70-gene signature were both significant predictors of relapse-free survival (*P*=0.026 and *P*=0.021 respectively, **Table 3.3**) in this dataset. Surprisingly, our signature was the only significant predictor of overall survival (*P*=0.008), independent of other clinical parameters and signatures. To further compare the performance of our signature to previously reported breast cancer gene signatures, we examined their respective predictive abilities on multiple datasets. As shown in the **Table 3.4**, the Rosetta 70-gene signature, Oncotype DX gene predictor, and our gene signature demonstrated superior performance over other signatures while our gene signature showed overall best performance.

To investigate the molecular difference between our signature and other breast cancer gene predictors of similar size, we examined the number of overlapping genes. Interestingly, only two (PRC1 and CENPA), one (CD44) and three (BRRN1, CDCA8, and MYBL2) genes overlapped between our 73-gene signature and the Rosetta 70-gene signature (9), the Wang et al (10) 76-gene signature, and the Miller et al (5) 32-gene signature, respectively. This lack of overlap suggests that our signature is comprised of genes distinct from previously reported gene predictors. Nevertheless, two-way contingency table analysis revealed strong associations between prediction results of individual samples made by our outcome signature and the Rosetta 70-gene signature, the wound-response signature and the intrinsic-subtype model (7) (**Table 3.5**). These findings are consistent with previously reported study that distinct gene predictors, although with little overlap in terms of gene identity, may have high rates of concordance in prediction results for individual samples (31). Taken together, our distinct gene signature outperformed other known predictors while being concordant in outcome prediction of individual samples.

There is established precedence for clinical use of molecular markers to help decide customized therapy for individuals with breast cancer. For example, ER and PR, and ERBB2 have been used to assess potential response to hormonal therapy and Herceptin, respectively. However, a single marker such as ER has been found insufficient to fully stratify patient into different diagnostic/prognostic subtypes. In this study, we aimed to identify a transcriptional fingerprint of estrogen, which reflects the downstream activity of estrogen signaling pathway, and thus may be a more efficient predictor of breast cancer recurrence.

Unlike most previously reported breast cancer signatures that were developed using supervised analysis based on patient diagnosis/prognosis status (5-10), our signature was discovered by specifically selecting estrogen-regulated genes, thus representing the activities of estrogen signaling, a key biological characteristic of breast cancer tumors. We profiled gene expression of three breast cancer cell lines during early time-points following estrogen treatment. We observed that over 80% of our estrogen-regulated genes were already activated within 1-2 hr following estrogen treatment in MCF7 breast cancer cell line. Genome-wide location analysis confirmed that a significant portion of these genes are directly occupied by ER, suggesting an enrichment of direct ER target genes in our signature. In addition, our gene signature distinguishes ER+ and ER- patients, as well as separates patients who did well with hormonal therapy from those who did not, indicating its specificity in monitoring estrogen activity.

In developing the 73-gene outcome signature we focused on *in vitro* estrogen-regulated genes and further selected a subset that is associated with patient outcome *in vivo* in human breast tumors. These genes are unique as they represent a subset of downstream targets of estrogen signaling that are predictive of breast cancer outcome. The 73-gene signature predicts breast cancer outcome in 10 out of 11 datasets we analyzed. Besides correctly assigning most ER- tumors in each dataset into high-risk group, this signature is able to stratify the ER+ samples into prognostic subtypes, suggesting that it may better reflect tumor aggressiveness than ER status alone. Most importantly, our signature provides additional prognostic information beyond standard clinical factors and yields overall best performance against previously reported breast cancer outcome predictors.

Further validation and refinement of our signature using additional datasets with larger cohorts of breast cancer patients will help to strengthen its clinical value. This study lays the ground for future characterization of individual signature genes to facilitate in the understanding of breast cancer progression as well as help select genes with critical roles in estrogen response for breast cancer therapy. Furthermore, as RT-PCR assays of paraffin-embedded tissues have recently been developed (6), it is technically feasible to develop an RT-PCR assay of our 73-gene signature for future validation and, potentially later on, for clinical usage. Our signature may be useful in selection of high-risk patients for adjuvant therapy as well as in sparing some hormone-sensitive patients from aggressive therapy.

## Methods and Materials

Cell culture: Breast cancer cell lines (MCF-7, T47-D, BT-474) were maintained as previously described (32). For defined estrogen culture experiments, cells were rinsed in PBS, grown in steroid-depleted media (phenol red-free IMEM (Improved Minimal Essential Media) supplemented with 10% charcoal stripped calf bovine serum for 2 days, and treated with $10^{-9}$ M 17β-estradiol for 1, 2, 4, 8, 12 or 24 hours as described previously (13).

RNA extraction and microarray experiments: RNA was isolated, labeled and hybridized according to the Affymetrix protocol (Affymetrix GeneChip Expression Analysis Technical Manual, Rev. 3) by the University of Michigan Comprehensive Cancer Center Affymetrix and cDNA Microarray Core Facility as described previously (13). All primary array data have been deposited in the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) with series number GSE3834.

Affymetrix microarray data analysis: Data from microarray experiments were calculated, normalized and log2-transformed using RMAExpress (33). As described previously (13), the MCF-7 profiles were generated on the Affymetrix U133A platform, and the other profiles were generated on the Affymetrix U133 Plus 2.0 platform; we thus only considered 22,283 probe sets that were common in both platforms for subsequent analysis. Expression values within each cell line were first *z*-transformed to zero mean and unit variance. Time-course experiments were analyzed using EDGE (12) to identify genes differentially expressed in estrogen-treated relative to estrogen-starved cells.

Multiple hypothesis testing was adjusted by False Discovery Rate (FDR). 1,314 probe sets were identified differentially expressed over time with FDR less than 0.01. These genes were then subjected to hierarchical clustering, which resulted in one estrogen-induced gene cluster containing 532 probe sets and the other estrogen-inhibited gene cluster containing 782 probe sets. For subsequent analyses, only genes in the estrogen-induced cluster were used as we are more interested in estrogen-activated events during tumor progression.

Analysis of primary breast tumor data using the estrogen-regulated gene set: All primary breast tumor sets used in this study were collected by ONCOMINE (23) from previous publications or from the NCBI GEO database. Genes within each dataset were normalized to zero mean and unit variance. The largest Affymetrix U133A breast cancer dataset (10), containing 286 primary breast carcinomas, was used as the training set to conduct cross-validation and to develop an optimal gene set as previously described (34). All other datasets were used as independent test sets for validation purpose. The basic cross-validation procedure is as follows: (1) fit Cox regression model and calculate the Cox score for each gene in the Wang et al training set; (2) choose a set J of possible values of Cox scores S from step (1), and let $p_{min}$=1, $e_{min}$=1. (3) For each S in J, do the following: (4) perform k-means clustering (k=2) using only genes with absolute Cox scores greater than S. (5) perform a log-rank test to test whether the two clusters have different survival rates. Name the p-value of this test as p. (6) If p > $p_{min}$, then return to step 3. (7) perform 10-fold cross-validation by nearest centroid classification based on the class memberships defined by the clusters obtained in step 3. Name the misclassification error as e. (8) If $e \leq e_{min}$, then let $S_{opt} = S$, $p_{min} = p$, and $e_{min} = e$, and return to step 3. Otherwise return to step 3 without changing the value of $S_{opt}$. The optimal value of S is the value of $S_{opt}$ when cycle of this procedure terminates, and the optimal gene signature is designated as genes with absolute Cox scores greater than $S_{opt}$. The two clusters from k-means clustering based on these optimal genes are designated accordingly as either "high-risk" or "low-risk" by Kaplan-Meier (KM) survival analysis. Individual samples in the test data sets are then predicted as "high-risk" or "low-risk" by nearest centroid classification. When both the training and the test datasets used the same affymetrix platform, probe set IDs were used to cross-refer the two datasets. Otherwise, gene symbols were used to map genes from the training set to the test sets. When multiple report identifiers were found for one gene on a given platform, expressions of such reporter IDs were averaged per gene.

Survival Analysis: KM survival plots were compared by log-rank test in R (the R Foundation, http://www.r-project.org) for individual datasets. The end point of interest for survival analysis is recurrence-free survival unless the dataset only provides overall survival information. Multivariate Cox proportional-hazards regression analysis was conducted on van de Vijver et al. dataset in R. Concordance of sample prediction memberships by different signatures was tested in SPSS 11.5 for windows (SPSS Inc., Chicago, IL, USA).

**Table 3.1.** Description of the 73 Genes in the outcome signature

| Affy_U133A Probe Set | Gene Symbol | Affy_U133A Probe Set | Gene Symbol |
|---|---|---|---|
| 202148_s_at | PYCR1 | 220038_at | SGK3 |
| 203564_at | FANCG | 204498_s_at | ADCY9 |
| 209773_s_at | RRM2 | 208922_s_at | STX5 |
| 202954_at | UBE2C | 218620_s_at | HEMK1 |
| 202095_s_at | BIRC5 | 208688_x_at | EIF3S9 |
| 202870_s_at | CDC20 | 212022_s_at | MKI67 |
| 221436_s_at | CDCA3 | 206364_at | KIF14 |
| 214096_s_at | SHMT2 | 218663_at | HCAP-G |
| 218336_at | PFDN2 | 206976_s_at | HSPH1 |
| 221520_s_at | CDCA8 | 218270_at | MRPL24 |
| 214095_at | SHMT2 | 218009_s_at | PRC1 |
| 203145_at | SPAG5 | 209408_at | KIF2C |
| 204092_s_at | AURKA | 204817_at | ESPL1 |
| 218726_at | DKFZp762E1312 | 38158_at | ESPL1 |
| 211881_x_at | IGL@ | 204962_s_at | CENPA |
| 206472_s_at | TLE3 | 203755_at | BUB1B |
| 202107_s_at | MCM2 | 222039_at | LOC146909 |
| 216913_s_at | KIAA0690 | 204441_s_at | POLA2 |
| 219215_s_at | SLC39A4 | 212949_at | BRRN1 |
| 201710_at | MYBL2 | 219502_at | NEIL3 |
| 201584_s_at | DDX39 | 210466_s_at | SERBP1 |
| 204252_at | CDK2 | 204633_s_at | RPS6KA5 |
| 219910_at | HYPE | 203710_at | ITPR1 |
| 201421_s_at | WDR77 | 215193_x_at | HLA-DRB1 |
| 213906_at | MYBL1 | 212473_s_at | MICAL2 |
| 211576_s_at | COL18A1 | 213933_at | PTGER3 |
| 218984_at | PUS7 | 202464_s_at | PFKFB3 |
| 205284_at | KIAA0133 | 220266_s_at | KLF4 |
| 220177_s_at | TMPRSS3 | 212848_s_at | C9orf3 |
| 204489_s_at | CD44 | 202417_at | KEAP1 |
| 204490_s_at | CD44 | 204792_s_at | IFT140 |
| 209835_x_at | CD44 | 200706_s_at | LITAF |
| 205322_s_at | MTF1 | 215273_s_at | TADA3L |
| 218481_at | EXOSC5 | 221261_x_at | MAGED4 |
| 220029_at | ELOVL2 | 214736_s_at | ADD1 |
| 208305_at | PGR | 220935_s_at | CDK5RAP2 |
| 209273_s_at | HBLD2 | | |

**Table 3.2.** Multivariate Cox proportional hazards analysis of the 73-gene outcome signature in the Van de vijver et al. ER+ data set. The total number of samples is 225.

| | Relapse-Free Survival | | Overall Survival | |
|---|---|---|---|---|
| **Variable** | **Hazard Ratio (95% CI)** | **p-value** | **Hazard Ratio (95% CI)** | **p-value** |
| Our estrogen-regulated signature | 2.24 (1.35-3.70) | **0.002** | 3.27 (1.62-6.62) | **0.001** |
| The Oh et al. Estrogen-SAM genesignature (IIE vs. IE) | 2.32 (1.31-4.11) | **0.004** | 2.24 (0.95-5.28) | 0.066 |
| Age | 0.94 (0.89-0.98) | **0.004** | 0.94 (0.89-1.00) | 0.069 |
| Size (diameter >2cm vs. <2cm) | 1.49 (0.93-2.37) | 0.095 | 1.41 (0.76-2.61) | 0.280 |
| Tumor Grade | | | | |
| (intermediate vs. well diff.) | 1.40 (0.72-2.72) | 0.320 | 2.02 (0.65-6.28) | 0.230 |
| (poorly vs. well diff.) | 1.30 (0.64-2.63) | 0.460 | 2.86 (0.91-9.02) | 0.070 |
| Node status | | | | |
| (1-3 vs. 0 positive nodes) | 1.82 (0.93-3.57) | 0.082 | 1.65 (0.66-4.18) | 0.290 |
| (>3 vs. 0 positive nodes) | 2.87 (1.23-6.74) | **0.015** | 2.22 (0.69-7.11) | 0.180 |
| Hormonal or chemotherapy vs. no adjuvant therapy | 0.33 (0.16-0.66) | **0.002** | 0.43 (0.17-1.13) | 0.086 |

**Table 3.3.** Multivariate Cox proportional hazards analysis of the 73-gene outcome signature with two known predictors in the Van de vijver et al ER+ data set. Samples used for Van't veer et al. training model were excluded, leading to 181 samples in total.

| | Relapse-Free Survival | | Overall Survival | |
|---|---|---|---|---|
| **Variable** | **Hazard Ratio (95% CI)** | **p-value** | **Hazard Ratio (95% CI)** | **p-value** |
| Our estrogen-regulated signature | 2.01 (1.09-3.72) | **0.026** | 3.63 (1.40-9.42) | **0.008** |
| 70-gene Signature (poor vs. good) | 2.42 (1.14-5.14) | **0.021** | 2.37 (0.72-7.85) | 0.160 |
| The Oh et al. Estrogen-SAM gene signature (IIE vs. IE) | 1.83 (0.95-3.52) | **0.070** | 1.71 (0.62-4.75) | 0.300 |
| Age | 0.97 (0.92-1.03) | 0.340 | 1.00 (0.92-1.07) | 0.900 |
| Size (diameter >2cm vs. <2cm) | 1.18 (0.68-2.04) | 0.560 | 1.29 (0.60-2.77) | 0.510 |
| Tumor Grade | | | | |
| (intermediate vs. well diff.) | 0.97 (0.47-2.00) | 0.930 | 1.33 (0.39-4.49) | 0.650 |
| (poorly vs. well diff.) | 0.67 (0.29-1.54) | 0.350 | 1.63 (0.46-5.75) | 0.450 |
| Node status | | | | |
| (1-3 vs. 0 positive nodes) | 1.86 (0.90-3.88) | 0.096 | 1.76 (0.63-4.92) | 0.280 |
| (>3 vs. 0 positive nodes) | 3.56 (1.39-9.14) | **0.008** | 2.84 (0.77-10.5) | 0.120 |
| Hormonal or chemotherapy vs. no adjuvant therapy | 0.33 (0.16-0.68) | **0.003** | 0.38 (0.14-1.03) | 0.056 |

**Table 3.4**. Performance comparisons of the estrogen-regulated signature with previously reported breast cancer signatures. Individual gene signatures were extracted from the original literature. Except for Oncotype DX and estrogen-regulated predictors were trained in Wang et al. breast dataset, the other signatures were trained in each respective dataset. Each signature was used to perform k-mean clustering and patients were separated into high-risk and low-risk groups, which were used as a training model to predict samples in other datasets by nearest centroid classification. The best signature for each dataset was highlighted in bold.

| **Signature Source** | Our Signature | | | Wang et al. | | | Pawitan et al. | | |
|---|---|---|---|---|---|---|---|---|---|
| | Log-rank P-value* | Hazard Ratio (95% CI) | C-Index (95% CI)** | Log-rank P-value | Hazard Ratio (95% CI) | C-Index (95% CI) | Log-rank P-value | Hazard Ratio (95% CI) | C-Index (95% CI) |
| Number of Genes | | 14 | | | 76 | | | 64 | |
| Wang_Breast | **1.03E-06** | **2.60 (1.75, 3.86)** | **0.63 (0.58, 0.67)** | 0.0007 | 1.90 (1.30, 2.78) | 0.58 (0.54, 0.63) | 0.09 | 1.39 (0.95, 2.04) | 0.55 (0.50, 0.59) |
| Pawitan_Breast | **1.20E-07** | **8.60 (3.3, 22.41)** | **0.73 (0.66, 0.80)** | 0.0002 | 3.65 (1.77, 7.53) | 0.65 (0.57, 0.74) | ######## | 10.3 (3.13, 34.0) | 0.71 (0.65, 0.78) |
| Miller_Breast | 0.0004 | 2.59 (1.50, 4.46) | 0.62 (0.56, 0.69) | **0.0001** | **2.71 (1.60, 4.61)** | **0.63 (0.56, 0.69)** | 0.001 | 2.51 (1.43, 4.41) | 0.61 (0.55, 0.68) |
| Vantveer_Breast | 0.0003 | 2.90 (1.58, 5.33) | 0.64 (0.57, 0.71) | 0.03 | 1.87 (1.04, 3.36) | 0.58 (0.51, 0.65) | 0.04 | 1.91 (1.03, 3.56) | 0.58 (0.51, 0.65) |
| Sotirious_Breast | **9.09E-06** | **2.98 (1.80, 4.93)** | **0.66 (0.60, 0.72)** | 0.0005 | 2.30 (1.42, 3.74) | 0.64 (0.58, 0.70) | 0.0007 | 2.41 (1.42, 4.08) | 0.64 (0.58, 0.69) |
| Bild_Breast | **0.0001** | **3.10 (1.69, 5.68)** | **0.64 (0.58, 0.71)** | 0.005 | 2.17 (1.24, 3.81) | 0.60 (0.52, 0.67) | 0.33 | 1.34 (0.75, 2.40) | 0.55 (0.47, 0.62) |
| Oh_Breast | 0.01 | 2.71 (1.23, 5.95) | 0.62 (0.53, 0.71) | 0.009 | 2.59 (1.23, 5.42) | 0.61 (0.52, 0.71) | 0.04 | 2.21 (1.01, 4.86) | 0.60 (0.51, 0.69) |
| Sorlie_Breast | 0.003 | 2.44 (1.32, 4.48) | 0.62 (0.54, 0.69) | 0.12 | 1.58 (0.88, 2.84) | 0.57 (0.49, 0.64) | 0.19 | 1.49 (0.82, 2.72) | 0.57 (0.49, 0.64) |
| Vandevijver_Breast | 1.76E-06 | 2.76 (1.79, 4.25) | 0.64 (0.59, 0.69) | 0.001 | 1.99 (1.31, 3.03) | 0.59 (0.54, 0.65) | ######## | 2.73 (1.73, 4.31) | 0.63 (0.58, 0.68) |

| **Signature Source** | Miller et al. | | | Van't veer et al. | | | Oncotype Dx | | |
|---|---|---|---|---|---|---|---|---|---|
| | Log-rank P-value | Hazard Ratio (95% CI) | C-Index (95% CI) | Log-rank P-value | Hazard Ratio (95% CI) | C-Index (95% CI) | Log-rank P-value | Hazard Ratio (95% CI) | C-Index (95% CI) |
| Number of Genes | | 32 | | | 70 | | | 16 | |
| Wang_Breast | 0.62 | 1.12 (0.73, 1.72) | 0.52 (0.48, 0.57) | 0.0002 | 2.05 (1.39, 3.01) | 0.60 (0.55, 0.65) | 0.001 | 1.86 (1.26, 2.74) | 0.59 (0.54, 0.63) |
| Pawitan_Breast | 0.07 | 1.94 (0.93, 4.06) | 0.57 (0.49, 0.65) | ######## | 8.67 (3.03, 24.8) | 0.71 (0.65, 0.78) | ######## | 4.73 (2.11, 10.6) | 0.68 (0.60, 0.76) |
| Miller_Breast | 0.008 | 2.04 (1.19, 3.48) | 0.59 (0.53, 0.66) | 0.0002 | 2.66 (1.54, 4.59) | 0.64 (0.57, 0.70) | 0.008 | 2.03 (1.19, 3.46) | 0.59 (0.53, 0.66) |
| Vantveer_Breast | 0.002 | 2.53 (1.37, 4.67) | 0.59 (0.53, 0.66) | **########** | **6.81 (3.35, 13.9)** | **0.71 (0.65, 0.78)** | 0.0003 | 2.98 (1.61, 5.50) | 0.64 (0.57, 0.71) |
| Sotirious_Breast | 0.33 | 1.30 (0.77, 2.19) | 0.55 (0.49, 0.61) | ######## | 2.93 (1.77, 4.87) | 0.66 (0.60, 0.72) | 0.0009 | 2.23 (1.37, 3.64) | 0.65 (0.59, 0.70) |
| Bild_Breast | 0.0006 | 2.67 (1.49, 4.79) | 0.62 (0.54, 0.69) | 0.04 | 1.78 (1.01, 3.13) | 0.60 (0.54, 0.67) | 0.04 | 1.80 (1.03, 3.16) | 0.59 (0.52, 0.66) |
| Oh_Breast | 0.12 | 1.80 (0.86, 3.77) | 0.57 (0.47, 0.67) | 0.001 | 3.59 (1.59, 8.12) | 0.66 (0.57, 0.75) | **0.0003** | **4.91 (1.87, 12.9)** | **0.66 (0.58, 0.74)** |
| Sorlie_Breast | 0.001 | 2.61 (1.41, 4.81) | 0.61 (0.54, 0.67) | 0.01 | 2.17 (1.18, 3.99) | 0.60 (0.53, 0.68) | **0.0002** | **3.10 (1.64, 5.83)** | **0.64 (0.57, 0.72)** |
| Vandevijver_Breast | 0.0006 | 2.15 (1.37, 3.36) | 0.58 (0.53, 0.63) | ######## | 3.31 (2.10, 5.22) | 0.66 (0.61, 0.70) | **########** | **3.40 (2.18, 5.32)** | **0.67 (0.62, 0.71)** |

*This result is not adjusted for established clinical parameters such as stage, grade, and receptor status

**C-index: concordance index (area under the curve) for censored data calculated using Hmisc package in R.

**Table 3.5.** Two-way contigency table analysis measuring the association association among different breast cancer outcome signatures in the van de vijver et al. data set

**A. Two-way contigency table on Van de vijver et al. Data Set (n=295)**

| Our Estrogen-regulated Signature | 70-gene signature (# of patients) | |
|---|---|---|
| | Good | Poor |
| Low-Risk | 106 | 53 |
| High-Risk | 9 | 127 |

| Statistics for two-way table analysis | |
|---|---|
| p-value | <0.0001 |
| Cramer's V | 0.617 |

**B. Two-way contigency table on Van de vijver et al. Data Set (n=295)**

| Our Estrogen-regulated Signature | Wound-response signature (# of patients) | |
|---|---|---|
| | Activated | Quiescent |
| Low-Risk | 98 | 61 |
| High-Risk | 130 | 6 |

| Statistics for two-way table analysis | |
|---|---|
| p-value | <0.0001 |
| Cramer's V | 0.404 |

**C. Two-way contigency table on Van de vijver et al. Data Set (n=295)**

| Intrinsic Subtype | Our Estrogen-regulated Signature (# of patients) | |
|---|---|---|
| | Low-Risk | High-Risk |
| Basal-like | 3 | 50 |
| Luminal A | 108 | 15 |
| Luminal B | 13 | 42 |
| HER2+ and ER- | 9 | 26 |
| Normal-like | 26 | 3 |

| Statistics for two-way table analysis | |
|---|---|
| p-value | <0.0001 |
| Cramer's V | 0.720 |

**D. Two-way contigency table on Van de vijver et al. ER+ tumors (n=225)**

| Our Estrogen-regulated Signature | Estrogen-SAM signature (# of patients) | |
|---|---|---|
| | Group IE | Group IIE |
| Low-Risk | 90 | 58 |
| High-Risk | 12 | 65 |

| Statistics for two-way table analysis | |
|---|---|
| p-value | <0.0001 |
| Cramer's V | 0.431 |

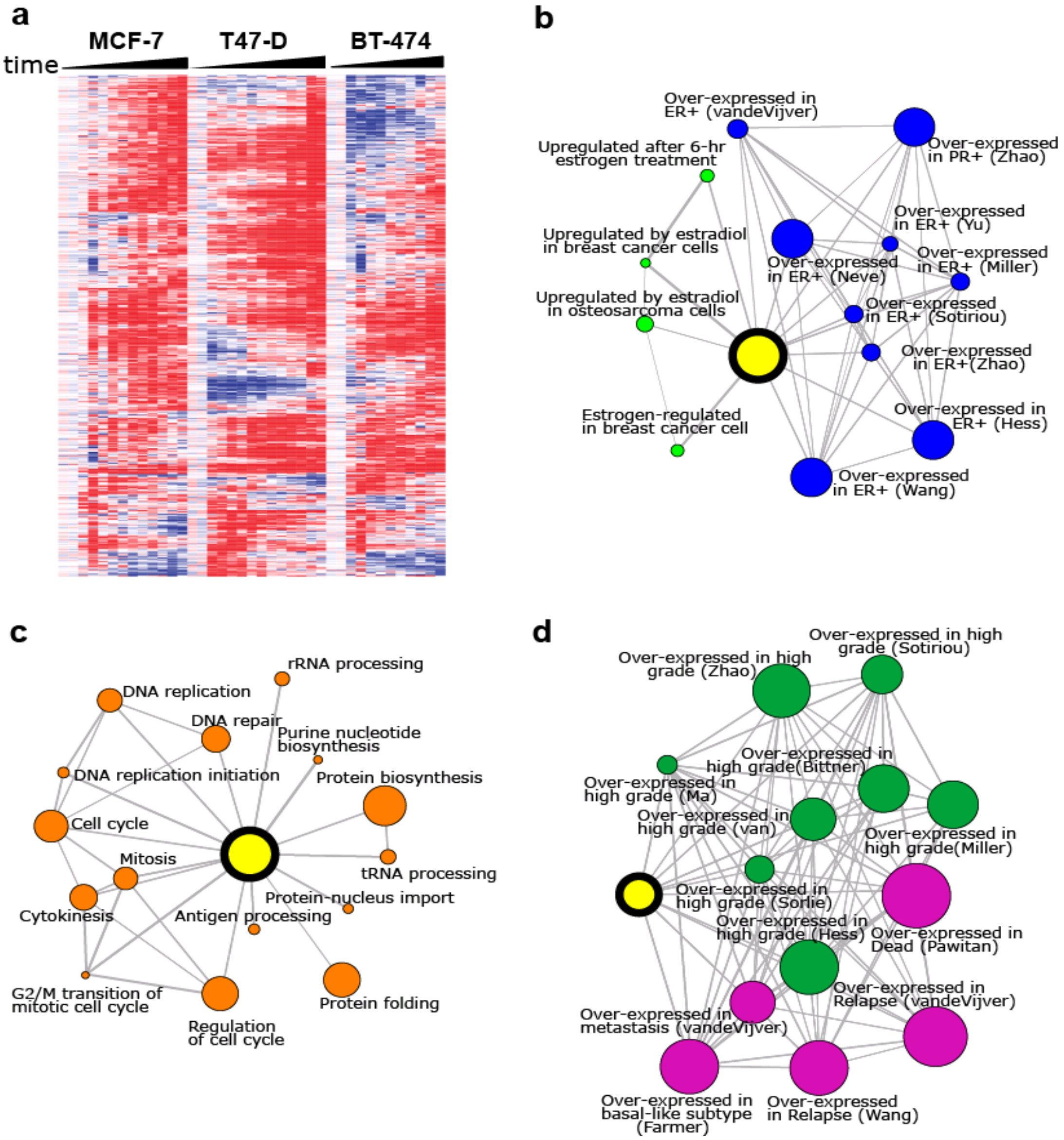


**Figure 3.1**. Identification and molecular concept map analysis of estrogen-induced genes. **a**. Heatmap representation of 532 *in vitro* estrogen-induced genes across three ER+, estrogen sensitive breast cancer cell lines (MCF-7, T47-D, and BT-474) following 17β-estradiol treatment. Each row represents a gene, and each column represents a sample treated with estrogen for different time periods (0, 1, 2, 4, 8, 12 or 24 hours with replicates). **b-d.** Molecular concept map analysis (MCM) of the estrogen-induced genes (yellow node with black frame) showing enrichment networks of (**b**) previously reported estrogen-regulated molecular concepts both *in vitro* and *in vivo*, (**c**) gene ontology concepts, and (**d**) breast cancer progression and prognosis concepts. Each node represents a molecular concept. The node size is proportional to the number of genes in the concept. Each edge represents a statistically significant enrichment. Concepts of "up-regulated genes by estrogen treatment" are indicated by light green nodes. Blue, holly green and purple nodes represent genes up-regulated in ER+ cancer, high-grade breast cancer, and patients with poor outcome, respectively. Enriched gene ontology terms are represented by orange nodes.

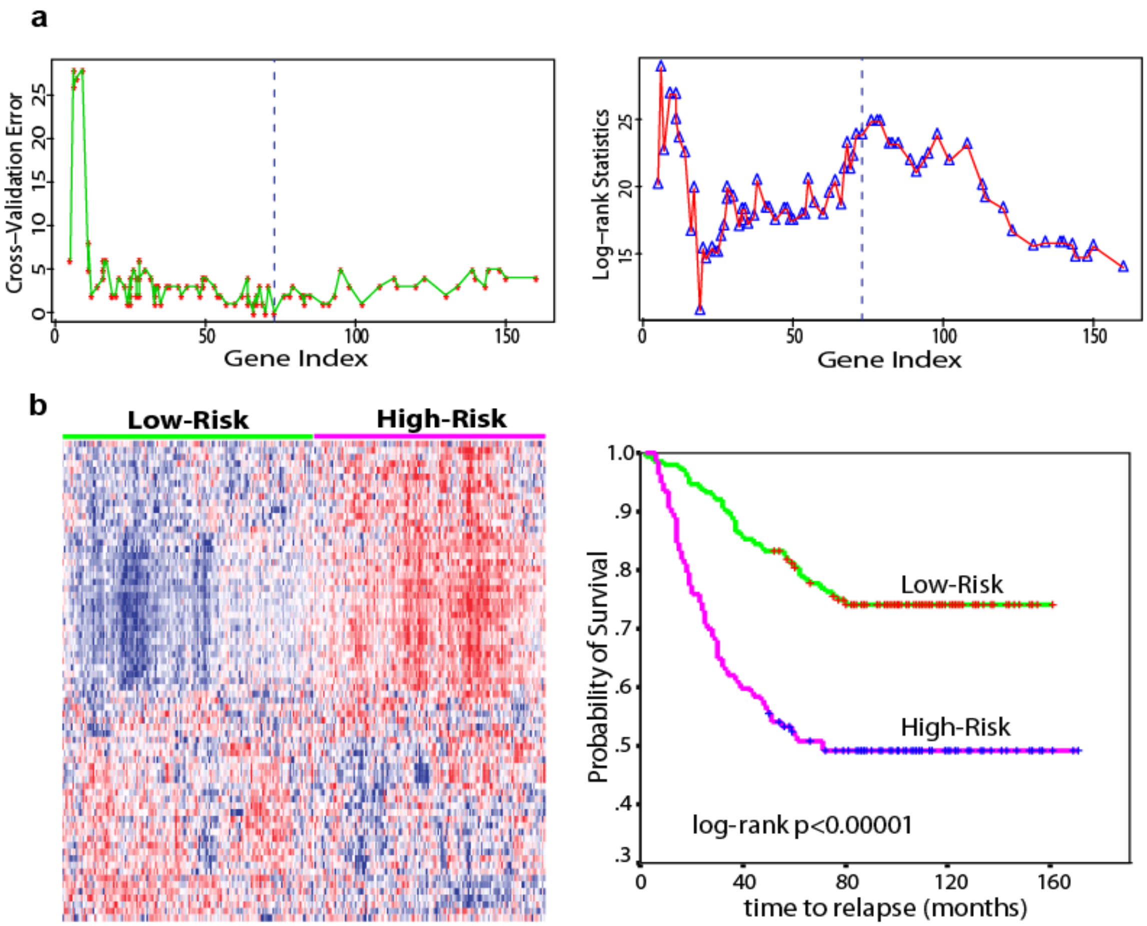


**Figure 3.2**. Estrogen-regulated genes stratified breast cancer samples into two groups with significantly different prognoses. **a**. Representation of step-wise cross-validation on the Wang et al. training set. The left panel presents the number of misclassified samples by cross-validation, and the right panel presents survival difference of the resulted two clusters when a particular set of genes were used. The X-axis represents the number of top genes, ordered by their corresponding survival significance. The dashed line indicates the threshold used to select the optimal gene signature. **b**. K-mean clustering representation of the 73 estrogen-regulated genes in the training cohort (left) and its Kaplan-Meier survival plot (right). The 73 genes were selected based on minimal misclassification error by 10-fold cross validation in the space of the initial identified 532 genes (Panel **a**.)

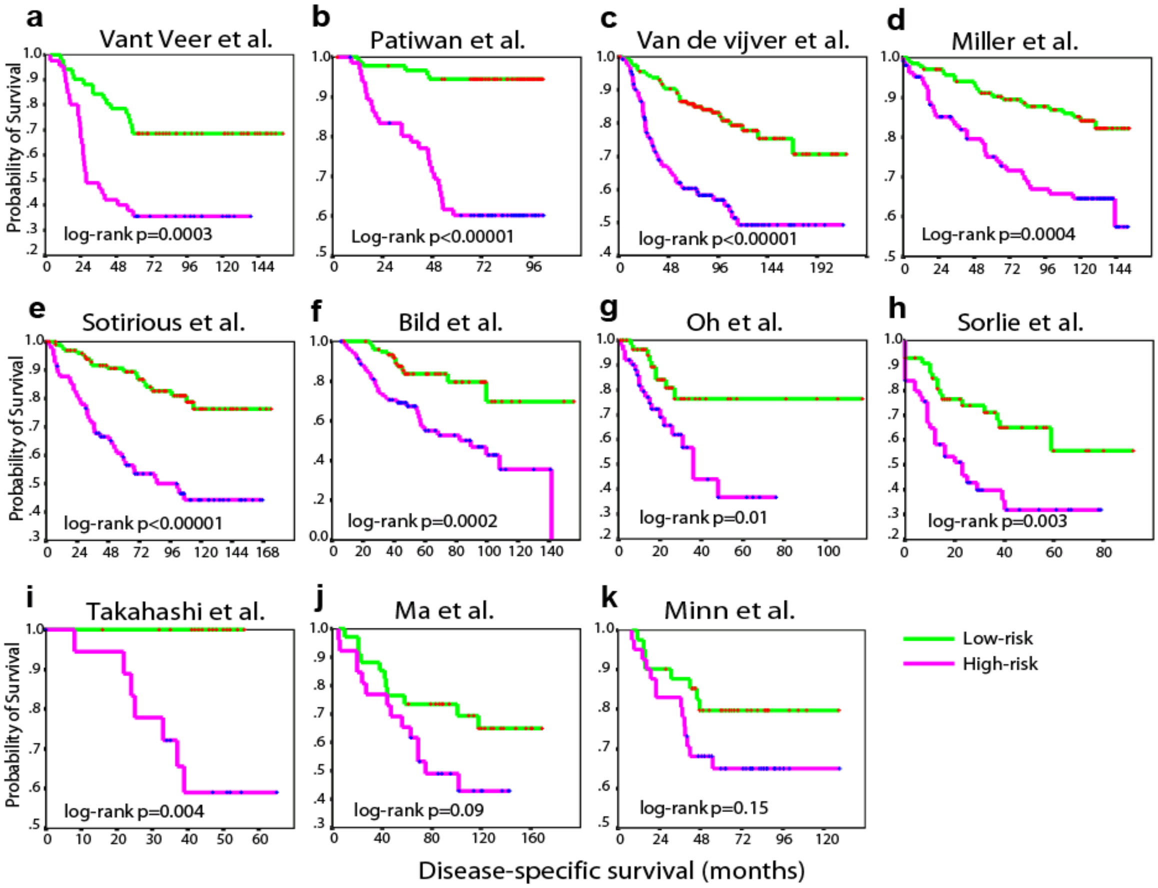


**Figure 3.3**. The 73-gene outcome signature predicts clinical outcome of breast cancer. The low-risk and high-risk groups in each study were predicted on the basis of the expression patterns of the 73 signature genes as described in the Method. KM analysis was used to evaluate the significance of outcome difference between the two groups. *P* values were calculated by the log-rank test.

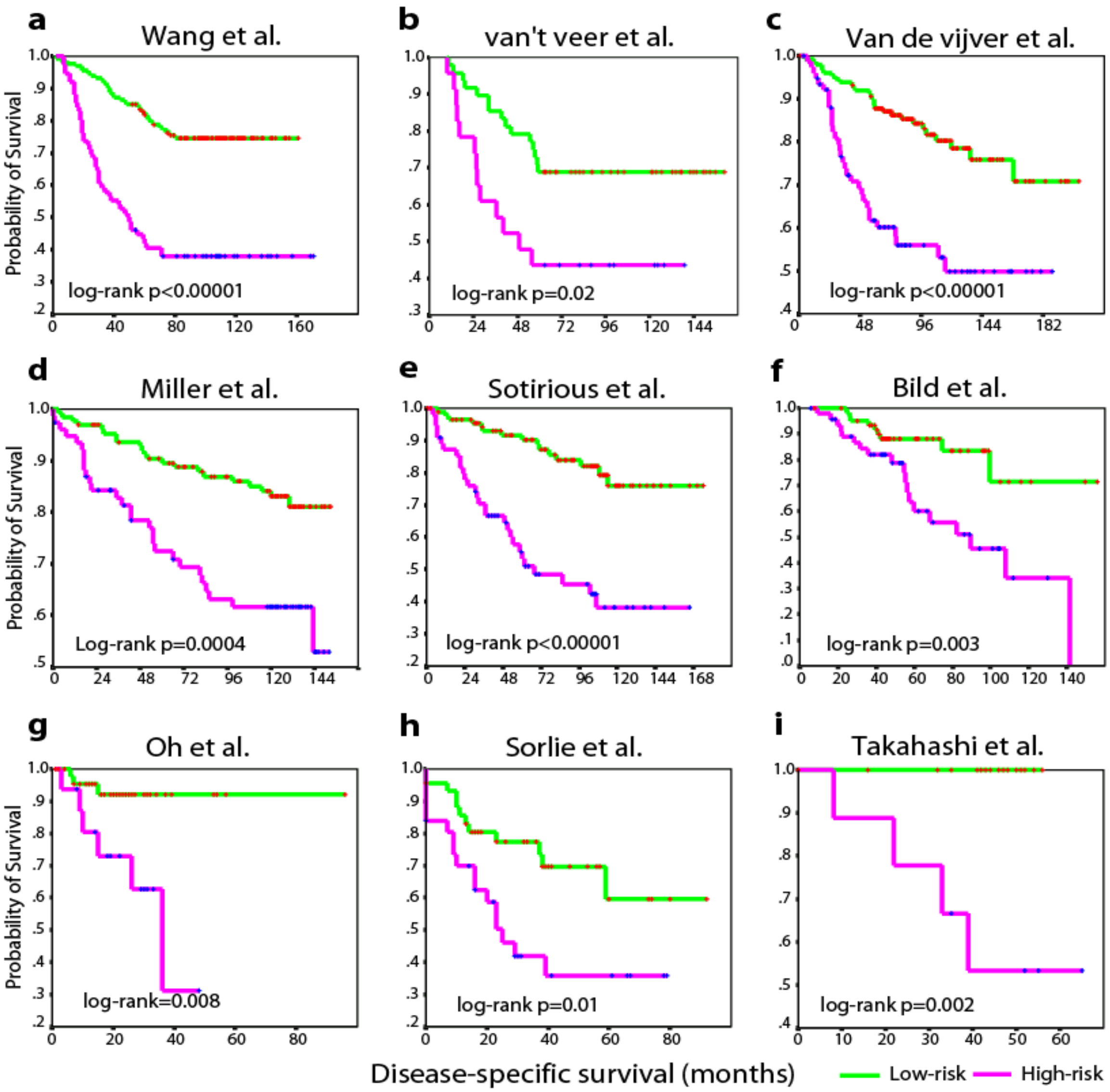


**Figure 3.4**. The 73-gene outcome signature predicts clinical outcome of ER+ breast cancer. The ER+ breast cancer samples were extracted from their respective datasets and the significance of outcome difference between the low-risk and high-risk groups were estimated by KM survival analysis. *P* values were calculated by the log-rank test. The Ma et al. data set is not included in this analysis since nearly all of its samples are ER+ and thus have been presented in **Figure 3.3**.

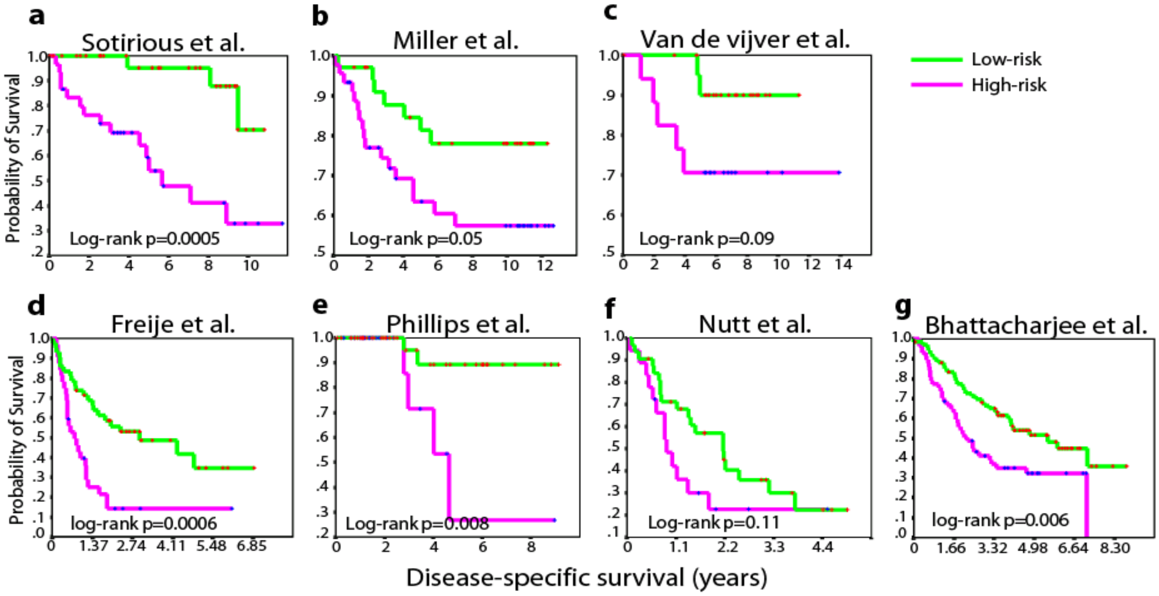


**Figure 3.5**. The 73-gene outcome signature predicts clinical outcome in tamoxifen-treated breast cancer subcohorts (a-c), gliomas (d-f), and lung adenocarcinoma (g). The low-risk and high-risk groups were predicted by the 73-gene signature with nearest centroid classification. KM analysis was used to evaluate the significance of outcome difference between the two groups. *P* values were calculated by the log-rank test.

# REFERENCES

1. Clemons, M, and Goss, P (2001). Estrogen and the risk of breast cancer. *N Engl J Med* **344**, 276-285.

2. Jordan, VC (2003). Tamoxifen: a most unlikely pioneering medicine. *Nat Rev Drug Discov* **2**, 205-213.

3. (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**, 1687-1717.

4. Brenner, H (2002). Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. *Lancet* **360**, 1131-1135.

5. Miller, LD, Smeds, J, George, J, et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* **102**, 13550-13555.

6. Paik, S, Shak, S, Tang, G, et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-2826.

7. Perou, CM, Sorlie, T, Eisen, MB, et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747-752.

8. van de Vijver, MJ, He, YD, van't Veer, LJ, et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009.

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

10. Wang, Y, Klijn, JG, Zhang, Y, et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-679.

11. Pujol, P, Daures, JP, Thezenas, S, et al. (1998). Changing estrogen and progesterone receptor patterns in breast carcinoma during the menstrual cycle and menopause. *Cancer* **83**, 698-705.

12. Leek, JT, Monsen, E, Dabney, AR, and Storey, JD (2006). EDGE: extraction and analysis of differential gene expression. *Bioinformatics* **22**, 507-508.

13. Creighton, CJ, Cordero, KE, Larios, JM, et al. (2006). Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. *Genome Biol* **7**, R28.

14. Oh, DS, Troester, MA, Usary, J, et al. (2006). Estrogen-regulated genes predict survival in hormone receptor-positive breast cancers. *J Clin Oncol* **24**, 1656-1664.

15. Frasor, J, Danes, JM, Komm, B, et al. (2003). Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. *Endocrinology* **144**, 4562-4574.

16. Tomlins, SA, Mehra, R, Rhodes, DR, et al. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* **39**, 41-51.

17. Buterin, T, Koch, C, and Naegeli, H (2006). Convergent transcriptional profiles induced by endogenous estrogen and distinct xenoestrogens in breast cancer cells. *Carcinogenesis* **27**, 1567-1578.

18. Frasor, J, Stossi, F, Danes, JM, et al. (2004). Selective estrogen receptor modulators: discrimination of agonistic versus antagonistic activities by gene expression profiling in breast cancer cells. *Cancer Res* **64**, 1522-1533.

19. Stossi, F, Barnett, DH, Frasor, J, et al. (2004). Transcriptional profiling of estrogen-regulated gene expression via estrogen receptor (ER) alpha or ERbeta in human osteosarcoma cells: distinct and common target genes for these receptors. *Endocrinology* **145**, 3473-3486.

20. Sotiriou, C, Wirapati, P, Loi, S, et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262-272.

21. Carroll, JS, Meyer, CA, Song, J, et al. (2006). Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38**, 1289-1297.

22. Lin, CY, Strom, A, Vega, VB, et al. (2004). Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biol* **5**, R66.

23. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

24. Sribnick, EA, Ray, SK, and Banik, NL (2006). Estrogen prevents glutamate-induced apoptosis in C6 glioma cells by a receptor-mediated mechanism. *Neuroscience* **137**, 197-209.

25. Marquez-Garban, DC, Chen, HW, Fishbein, MC, Goodglick, L, and Pietras, RJ (2007). Estrogen receptor signaling pathways in human non-small cell lung cancer. *Steroids* **72**, 135-143.

26. Stabile, LP, and Siegfried, JM (2004). Estrogen receptor pathways in lung cancer. *Curr Oncol Rep* **6**, 259-267.

27. Freije, WA, Castro-Vargas, FE, Fang, Z, et al. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* **64**, 6503-6510.

28. Phillips, HS, Kharbanda, S, Chen, R, et al. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157-173.

29. Nutt, CL, Mani, DR, Betensky, RA, et al. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* **63**, 1602-1607.

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

31. Fan, C, Oh, DS, Wessels, L, et al. (2006). Concordance among gene-expression-based predictors for breast cancer. *N Engl J Med* **355**, 560-569.

32. Rae, JM, Johnson, MD, Scheys, JO, et al. (2005). GREB 1 is a critical regulator of hormone dependent breast cancer growth. *Breast Cancer Res Treat* **92**, 141-149.

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

34. Bair, E, and Tibshirani, R (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* **2**, E108.

# CHAPTER 4

## Molecular Classification of Cancer using Genetic Programming

Despite important advances in microarray-based molecular classification of tumors, its application in clinical settings remains formidable. This is in part due to the limitation of current analysis programs in discovering robust biomarkers and developing classifiers with a practical set of genes. Genetic Programming (GP) is a type of machine learning technique that uses evolutionary algorithm to simulate natural selection as well as population dynamics, hence leading to simple and comprehensible classifiers. Here we applied GP to cancer gene expression profiling data to select feature genes and build molecular classifiers by mathematical integration of these genes. Analysis of thousands of GP classifiers generated for a prostate cancer dataset revealed repetitive use of a small set of highly discriminative feature genes, many of which are known disease-associated. GP classifiers often comprise five or less genes and successfully predict cancer types and subtypes. Importantly, GP classifiers generated in one study is able to predict samples from an independent study, which may have used different microarray platforms. In addition, GP yielded better or similar classification accuracy as conventional classification methods. Further, the mathematical-expression of GP classifiers provides insights into relationships between classifier genes. Taken together, this study has demonstrated that GP may be valuable for generating effective classifiers containing a practical set of genes for diagnostic/prognostic cancer classification.

The development of high-throughput microarray-based technology will potentially revolutionize cancer research in a number of areas including cancer classification, diagnosis and treatment. Expression profiling at the mRNA level can be used in the molecular characterization of cancer by simultaneous assessment of a large number of genes (1-5). This approach can be used to determine gene expression alterations between different tissue types such as those obtained from healthy controls and cancer patients. Analysis of such large-scale gene expression profiles of cancer will facilitate the identification of a subset of genes that could function as diagnostic or prognostic biomarkers. The development of molecular classifiers that allow segregation of tumors into clinically relevant molecular subtypes beyond those possible by pathological classification may subsequently serve to classify tumors with unknown origin into different cancer types or subtypes. However, due to the large number of genes and the relatively small number of patient cases available from such studies, it remains a challenge to find a robust gene signature for reliable prediction.

As discussed in **Chapter 2**, a number of computational and statistical models have been developed for molecular classification of tumors. However, many of methods are often developed using parametric statistical techniques and thus have difficulty in finding non-linear relationships between genes. Alternatively, complex models such as neural networks often deliver “black box” solutions for classification and do not give insight into relationships between genes. In this study, we present a machine learning approach called Genetic Programming (GP) for molecular classification of cancer. GP belongs to a class of evolutionary algorithms and was first introduced by Koza (6) in 1992. Recently, GP has been shown to be a promising approach for discovering comprehensible rule-based classifiers from medical data (7,8) as well as gene expression profiling data (9-14). However, the potential of GP in cancer classification has not been fully explored. For example, GP classifiers identified from one dataset have not been validated in independent datasets. Here, we applied GP algorithm to cancer expression profiling data to identify potentially informative feature genes, build molecular classifiers and classify tumor samples. A basic flowchart of GP has been described in **Figure 4.1** and typical parameters used in GP have been detailed in **Table 4.1**. We tested GP in one Small Round Blue Cell Tumors (SRBCTs), one lung adenocarcinoma and five prostate cancer datasets (**Table 4.2**), and evaluated the generality of GP classifiers within and across datasets. In addition, we compared the performance of GP with that of other common classification techniques, such as linear discriminant analysis and support vector machines, for prediction accuracy.

To investigate the ability of GP to robustly select feature genes, we examined gene occurrences across classifiers generated from our GP system. Our results revealed that a small set of genes was frequently selected. For example, an analysis of feature genes in a set of 1000 best classifiers from GP to distinguish primary prostate cancer from metastatic samples on LaTulippe et al. (Memorial Sloan-Kettering Cancer Center, MSKCC) prostate dataset (15) indicated a high tendency of GP in selecting certain genes across classifiers (**Figure 4.2)**. **Figure 4.2A** presents the normalized $z$ score (16) of the frequency of each gene in the 1000 classifiers that contains a total of 2000 gene occurrence, with the X-axis representing Gene Index. As shown in the figure, only 261 out of the total 3547 genes used for this study occurred at least once. Interestingly, 46 of them occurred at least twelve times ($z$ score >=15, p<0.0001, **Table 4.3**). The fact that this small set of genes has dominated the generated classifiers implies that such genes may be truly important for prostate cancer metastasis, and may serve as discriminative biomarkers for cancer progression. As GP is stochastic and may give different solutions in each run, it is interesting to examine the reproducibility of gene selection across independent runs. Thus, we created another independent set of 1000 classifiers using identical GP parameters on the same training set. A total of 264 genes occurred at least once in this set of GP classifiers. Notably, 206 of them were common in both sets and a highly positive correlation of $z$ scores of these gene between the two sets was observed ($R^2 = 0.94$, $P < 1x10^{-5}$, **Figure 4.2B**).

Next we examined the 46 most frequently occurring feature genes in the above analysis (**Table 4.3**). Strikingly, the top 3 probes represented the same gene, MYH11, which has been reported to be down-regulated in multiple metastatic cancers (17). Another top-listed gene was EZH2, encoding a polycomb group protein that we and others have previously characterized as over-expressed in aggressive epithelial tumors (18,19). We therefore hypothesized that the top frequently occurring genes might serve as a multiplex signature to distinguish metastatic prostate cancer from primary prostate cancer. To test this, hierarchical clustering was performed to group cancer samples based on the expression patterns of these genes. As shown in **Figure 4.3A**, these top 46 genes clustered tumor samples into their corresponding diagnostic classes (metastatic or primary prostate cancer), each with a unique expression signature. Interestingly, the same set of genes also successfully classified the independent Yu et al. (Pittsburgh) prostate cancer dataset. Similar results were observed when samples of the SRBCT dataset were clustered based upon the top 54 frequent feature genes ($z$ score >=14) derived from the training samples of this dataset (**Figure 4.3B**). In addition, we also selected the top 26 feature genes ($z$ score >=40) from the 2000 classifiers developed from the Lapointe et al. (Stanford) prostate cancer training dataset. Hierarchical clustering based on the expression pattern of these genes grouped tumors of four independent prostate cancer datasets with high classification accuracy (**Figure 4.3C-F).**

To further investigate whether such feature genes can be used to predict class memberships of validation samples, we carried out class prediction of the SRBCT dataset by diagonal linear discriminant analysis (DLDA) and *k*-nearest neighbor analysis (kNN, *k* =3). The top 54 frequent genes selected from the 2000 classifiers generated from the training samples of SRBCT data were used as a gene signature to predict the validation samples. Both DLDA and kNN analysis predicted all of the 20 validation samples with 100% accuracy (data not shown), confirming that the frequent genes derived from GP are truly discriminative genes and capable of predicting unknown samples.

Next we sought to examine the performance of GP classifiers comprising only a handful feature genes. We first evaluated the ability of GP classifiers to accurately classify four diagnostic classes of cancers (NB, RMS, EWS and BL) within the SRBCT dataset (20). A set of 63 training samples was used by GP to generate distinguishing classifiers through cross-validation. Classification was performed in a binary mode (target versus non-target class). For each target class, the top 10 best classifiers were selected and employed to predict a validation set of 20 samples. Most of the classifiers achieved 100% sensitivity and specificity on the training set. Similar prediction accuracy was observed when these classifiers were applied to the 20 blinded validation samples. The best classifiers (**Table 4.4**) perfectly predicted all of the validation samples. The average prediction accuracy of the top 10 classifiers for each target class was 98.5% for BL (95% confidence interval [CI] = 0.97-1.00), 92.5% for EWS (95% CI = 0.89-0.96), 95.5% for NB (95% CI = 0.91-1.00), 95.5% for RMS (95% CI = 0.92-0.99). Overall, GP classifiers achieved comparable classification and prediction performance as the method described in the original study, while using much less genes. This high prediction accuracy, however, might be partially due to the fact that the 4 cancer types here are much more heterogeneous than the subtypes of any single cancer.

Thus, we next examined GP in classifying subtypes of lung adenocarcinoma, wherein samples were designated as “high-risk” or “low-risk” based on the original publication information (21). One hundred classifiers were generated by GP from 66 training samples and the top five were found to have the highest training accuracy of 98.5%. When these 5 classifiers were applied to the 20 test samples, we found a maximal prediction accuracy of 98.5% and an average prediction rate of 84.0% (95% CI, 0.70-0.99), being comparable with that of other classification methods as described in the later session.

A more challenging work is to validate classifiers across independent datasets. We thus investigated whether GP could distinguish molecular subtypes of a single cancer class from independent datasets. Two prostate cancer datasets (Pittsburgh, and MSKCC sets) were used to evaluate GP in classifying primary or metastatic prostate cancer. Genes within each dataset were standardized to have zero mean and unit variance, given that similar proportion of metastatic samples was observed in both datasets. The MSKCC samples were used as a training set to generate GP classifiers. The 20 classifiers that perfectly classified primary from metastatic prostate cancer in the training set were selected for prediction. When these classifiers were applied to predict the independent Pittsburgh prostate cancer samples, the best classifiers (**Table 4.4**) correctly predicted all metastatic prostate cancers, and 58 out of 62 clinically localized prostate cancers. This led to 100% sensitivity and 93.5% specificity. The average prediction accuracy of all of the 20 classifiers was 95.2% sensitivity (95% CI = 0.87-1.00) and 82.1% specificity (95% CI = 0.65-0.99).

The above two prostate cancer datasets were hybridized using the same Affymetrix HG-U95Av2 platform and shared similar proportion ratios of target/non-target samples. Next, we examined whether classifiers generated by GP could predict samples from independent studies that have used different microarray platforms. Three prostate cancer datasets (2,22,23) (UM, Stanford, and Pittsburgh datasets) were used to test GP classifiers in predicting benign prostate and primary prostate cancer (PCA) samples. Among them, the Stanford and UM datasets used spot cDNA microarrays while the Pittsburgh data used affymetrix HG-U95Av2 oligonucleotide arrays. Two-thirds of the Stanford samples were used as a training set to generate GP classifiers, whereas the other one-third, the UM and Pittsburgh samples were all considered as validation samples. We used GP to generate 2000 classifiers and selected the top 26 frequently occurring genes (*z*-score >=40) as potential feature genes. To examine whether these genes are present in all three microarray platforms we cross-referenced them to the UM and Pittsburgh datasets using gene symbols. Out of these 26 genes, 12 are present in all three datasets. We thus entered these 12 genes into the GP system to start a new round of 5-fold cross-validation on the Stanford training set. Five perfect classifiers were achieved and applied to the validation set. Prediction accuracy in the Stanford validation samples ranged from 84.4% to 90%. However, the classifiers performed poorly on UM and Pittsburgh datasets. We suspected that this might be due to the discrepancy in the proportion ratio of PCA/benign samples and/or the probe intensity difference across array platforms, which led to divergence in the constant D of a classifier (e.g. GENE[A] / GENE[B] - GENE[C] > D). However, we believe that the relationships between the classifier genes, although with varying values of D, may still be predictive across studies, given that the classifier genes are putative discriminative genes. For instance, one of the classifier, formulated as ‘IF (MYO6 + AMACR) >= -2.6776 THEN PCA’ (see **Table 4.5**), did not predict well on the validation sets. However, the expression value of “MYO6 + AMACR” might still be predictive. To test this, we transformed the five classifiers individually as described in Methods and calculated a prediction score for each validation sample by computing the left-side of each classifier-inequality on a continuous scale. The predictive ability of each classifier on each validation set was then assessed using the Area under ROC Curve (AUC). Notably, all classifiers were strongly significant (p-values < 5x10$^{-4}$, **Figure 4.4, Table 4.5**) in both the Stanford and Pittsburgh validation sets. The lowest AUC was 0.91 (95% CI = 0.80-1.00) and 0.87 (95% CI =

0.79-0.95) respectively. For UM dataset, except for one classifier being marginally significant (AUC=0.64, p-value =0.09), all other classifiers were also strongly significant (p-values < $5x10^{-4}$).

An ensemble "meta-classifier" combining multiple classifiers in general yields better prediction performance, as it involves more genes and multiple predictive signatures. Thus, we composed a "meta-classifier" based on the above 5 classifiers. For each sample, the calculated prediction scores of the five classifiers were totaled to an overall prediction score which was then defined as the prediction score of the "meta-classifier" for that sample. As expected, this "meta-classifier" revealed higher AUCs in each dataset (0.96, 0.99, 0.99 for the Stanford, UM and Pittsburgh set respectively, p-values <$5x10^{-4}$, see **Figure 4.4**, and **Table 4.5**)

Examination of classifier genes have revealed that GP classifiers (**Table 4.4** and **4.5**) are much simpler than predictors reported by other approaches (1,3,5,20,21,24-27), where more than ten genes are often required to build an effective predictor. GP, by contrast, can utilize only 2-5 genes to produce effective classifiers and achieve high prediction power. This simplicity may owe to the relatively strict expression constraints (**Table 4.1**) and the use of a non-parametric method in selecting informative genes rather than usual parametric statistical techniques. Further, unlike some other non-parametric approaches such as neural networks and support vector machines, GP is transparent in that the entire procedure for classifier generation and evolution is readily available for inspection and adjustment. GP also revealed interesting quantitative relationship between within-classifier genes. Studying the specific genes used by a classifier and their relationships may provide valuable information about gene interactions, transcriptional regulatory pathways, and clinical diagnosis.

One important criterion to assess a classification approach is how it performs in comparison to other commonly used algorithms in the same research area. To evaluate the performance of GP, the Burkitt lymphomas ('BL') in the SRBCT dataset and the high-risk class in the lung adenocarcinoma data were chosen as the target classes and five classification methods including Compound Covariate Predictor (CCP), 3-Nearest Neighbors (3NN), Nearest Centroid (NC), Support Vector Machines (SVM), and Diagonal Linear Discriminant Analysis (DLDA) were selected as comparing counterparts of the GP method. To produce a fair comparison, we took into account the small number of genes used by GP classifiers and conducted the comparison tests based on either 5- or 10-gene classifiers.

The same training and validation sets as described previously were used to evaluate the performance of each classification method. The basic procedure was defined by two steps: (1) two-sample student t-test was conducted for each gene in the training set, and the 5 or 10 genes with the smallest p-values were selected as test classifiers, (2) expression data of the selected 5 or 10 genes across the training samples were used to build a training model, which was subsequently applied to the validation samples. Each individual validation sample was predicted as either 'target-class' or 'non-target class'. Misclassification rate was defined as the percentage of validation samples that were misclassified by a test classifier. Since GP generates multiple classifiers, the average of the misclassification rates of the top GP classifiers derived from the training set was used to represent the misclassification rate of a typical well-performing GP classifier. For the SRBCT data, we used the averaged misclassification rate of the top 10 classifiers because there were 10 perfect classifiers generated from the training samples to classify ‘BL’ and ‘Non-BL’. Similarly, the 5 classifiers having the least classification error in the training set were used for the lung adenocarcinoma dataset. As shown in **Table 4.6**, the error rates were comparable across different methods. The GP system ranked the 2$^{nd}$ and the 3$^{rd}$ in the SRBCT and the lung adenocarcinoma datasets respectively when 5-gene classifiers were evaluated. We believe that this may reflect the general prediction strength of GP system when only a small number of genes are chosen.

An intrinsic advantage of GP is that it automatically selects a small number of feature genes during “evolution” (12). The “evolution” of classifiers from the initial population seamlessly integrates the process of gene selection and classifier construction. By contrast, gene selection must be performed in a separate stage for many other classification algorithms such as kNN, weighted voting, and DLDA. Moreover, it is relatively easier for GP to keep the number of genes used in one classifier small. As GP searches a larger space than most traditional classification approaches, there is an increased chance of GP in finding a better performing classifier. By identifying and utilizing a small number of genes and developing transparent and human-comprehensible rule-based classifiers, GP stands as a good algorithm of choice.

The challenge of the field of molecular classification lies in the tradeoff of prediction power and the number of genes used. We have therefore stringently tested GP classifiers, which comprised of 5 or less genes, in achieving high prediction accuracy in datasets with varying levels of classification complexity. Unlike other studies (13,28) validating their classifiers using cross-validation within a dataset, our result not only demonstrated that the top GP classifiers easily classified and predicted the SRBCT dataset, which contained 4 classes of physiologically heterogeneous cancers, with 100% accuracy, but also showed optimal performance in classifying and predicting subtypes of prostate cancer, for samples either of the same study or of a different study that used the same microarray platform.  In addition, GP-selected feature genes stay discriminative even for cancer samples examined in different studies that used greatly different microarray platforms. Due to this robustness and stability of GP feature genes, we expect GP classifier to be highly applicable to clinical diagnosis.

A major issue in GP as well as other machine learning systems is data over-fitting due to a large number of variables and a small number of cases in microarray profiling. This occurs when the classifier is strongly biased towards the training set and generates poor prediction generality in validation samples. To address this, our study restricted the complexity of classifiers and adopted an n-fold cross-validation strategy. By limiting the size and complexity of classifiers using the minimum description length principle of risk minimization (29), the system was forced to generate the most salient features likely to be the most general solutions (30). By re-sampling using n-fold cross-validation, classifiers derived from the training samples were re-examined in the test-fold samples to test how well the learning algorithm could be generalized. If the fitness on the training data in one fold is significantly better than the fitness on the test data, it may indicate that there is an issue of over-fitting in the data. Therefore, a careful examination of the samples may be necessary.

Another issue for GP is that it is computationally intensive. The estimated running time increases along with the complexity of the problem, and the number of variables. This can be partially resolved by using parallel processing which segments the problem into parts running on different processors simultaneously and then synchronizes among them. In addition, variable pre-filtering may also reduce the running time. As described in the result section, GP typically selects those inherently discriminative genes and usually a small set of genes dominates the selection. Thus, a pre-filtering such as excluding genes with small variances may significantly reduce the running time yet not affect the performance of classifiers.

Taken together, in this study we systematically evaluated the feasibility of GP in feature selection and cancer classification. By examining the feature genes used by GP classifiers we have demonstrated that GP is able to robustly select a set of highly discriminative genes. In addition, the mathematical expression of GP classifiers reveals interesting quantitative relationships between genes. By testing GP classifiers generated from training sets in validation sets, we have shown that GP classifiers can successfully predict tumor classes and outperform most of other classification methods when only a limited number of genes are allowed to build a classifier. Our work suggests that GP may be useful for feature selection and molecular classification of cancer using a practical set of genes.

## MATERIALS AND METHODS

Datasets. All datasets were obtained from ONCOMINE (31) or requested from the original authors. The Small Round Blue Cell Tumor (SRBCT) data (20) contained 88 samples from four types of cancer cells: neuroblastoma (NB), rhabdomyosarcoma (RMS), the Ewing family of tumors (EWS), and Burkitt lymphomas (BL). The entire dataset, excluding five non-SRBCT samples, was divided into a training set (63 samples) and a validation set (20 samples) as described in the original study. The lung adenocarcinoma dataset (21) contains 86 lung cancer samples. The samples were then subdivided into a high- or low-risk group as requested from the authors. Twenty eight high-risk and 38 low-risk samples were included in the training set while the remaining 20 samples were considered as the validation set.

Three prostate cancer datasets (2,22,23) from the University of Michigan (Dhanasekaran et al, UM), Stanford University (Lapointe et al., Stanford) and the University of Pittsburgh (Yu et al, Pittsburgh) respectively were used to classify primary prostate cancer (PCA) from benign prostate samples (BENIGN). A total of 56 samples were randomly selected from Stanford dataset as the training set. The rest of the Stanford samples, UM and Pittsburgh sets were treated as validation sets. Two prostate cancer datasets including the Pittsburgh and MSKCC datasets (15,23) were also retrieved to distinguish metastatic prostate cancer (MET) and PCA. Detailed study information is shown in **Table 4.2**.

Genetic Programming for classification: In this study, we used genetic programming to discover classifiers that are capable of classifying samples into different cancer types based on gene expression patterns. Genetic programming (GP) (6) is an evolutionary algorithm that simulates natural selection and population dynamics to search for intelligible relationships amongst the constituents in a system (classifiers in this study). A generic GP classifier-based prediction is shown as: IF '(GENE[A] / GENE[B] - GENE[C]) > D' THEN 'TARGET CLASS', where the IF clause is generated by GP, "TARGET CLASS" is pre-defined in the initial configuration file, D is a constant, and GENE[A], GENE[B] and GENE[C] represent the expression levels of gene A, B, and C respectively.

A basic flowchart for the GP system is given in **Figure 4.1A**. Briefly, the system randomly selects inputs such as gene identifiers and constant values, which are used to represent the expression values of corresponding genes. Such selected inputs are then combined with the function operators such as arithmetic or Boolean operators to compose tree-based GP classifiers, an example of which is given in **Figure 4.1B**. Such classifiers are eventually accumulated to form an initial population, where a small subgroup of classifiers is then selected to create a 'mating group'. Each classifier in this 'mating group' is assessed by a fitness function defined as the area under the receiver-operator characteristic curve (ROC-AUC), which is used widely to assess the accuracy of a diagnostic test that yields continuous test results in clinical research areas. The two fittest classifiers are then selected as 'mating' parents by a tournament selection scheme and ‘mated’ to produce ‘offspring’ via selective genetic operators such as *crossover*, or *mutation*. The *crossover* operator exchanges a subtree of one parent with the other to generate offspring (**Figure 4.1C**), while the *mutation* operator probabilistically chooses a node in a subtree and replace it with a new created subtree randomly. The generated offspring then replaces the least-fit parent classifiers in the population. Once new offspring fully replaces parent classifiers in the entire population, a new generation that in general contains better classifiers is created. This process of mating pool selection, fitness assessment, mating and replacement is repeated over generations, progressively creating better classifiers until a termination criterion is met (e.g., a perfect classifier with a fitness score of 1 or the maximum number of ‘generations’ is reached).

**Table 4.1** shows an example of primary GP parameters used to analyze the prostate cancer dataset from LaTulippe et al. (MSKCC) study. Given the limited sample size of each dataset we employed n-fold cross-validation procedure to estimate the generalization of classifiers in predicting samples with unknown class membership. For example, when a dataset is selected as the training set, it is randomly subdivided into n parts (or folds), wherein classifiers are developed as described in the above GP process using samples in n-1 folds. These classifiers are then tested on samples in the left-out fold to assess their potential generalization capability since such samples are not involved in the development of the classifiers. A good classifier is expected to classify well in the training samples as well as the samples in the left-out fold. This process is repeated n times with each fold taking turns as the testing fold and the best classifiers are then selected based upon overall performance on the training folds and the test fold.

We implemented parallel genetic programming algorithm in C (patented by Genetics Squared, Inc., Ann Arbor, MI 48104; http://www.genetics2.com). The analyses were performed on a parallel computer cluster (7 Dell 1850 1U racks with 2x3.2GHz Xeon processor and 1 Dell 1750 1U rack with 1x3.06GHz Xeon processor) running the Debian Linux operating system. The running times for different datasets varied from a few minutes to a few days, depending on a large number of parameters like the complexity of the problem, size of population used in the evolution, number of generations, cost of fitness calculation, number of classifiers, and size of the data set, etc. For the LaTulippe et al. prostate cancer dataset with the parameters listed in **Table 4.1**, it took approximately three and a half hours to complete a set of 1000 classifiers.

**Table 4.1** Settings for primary GP parameters used to analyze the Latulippe et al.prostate cancer data

| **Parameter** | **Setting** | **Description*** |
|---|---|---|
| Terminal Set | All inputs including gene expression values, and constant values. | A set where all end (leaf) nodes in the parse trees representing the programs must be drawn. A terminal could be a variable, a constant or a function with no arguments |
| Function Set | Boolean and floating point operators: <, >, <=, =>, *, /, +, - | A set of operators, e.g. +, -, *, ÷. These act as the branch points in the parse tree, linking other functions or terminals |
| Selection | Generational, tournament size 5 | An evolution is called “generational” when the entire existing population of classifiers is replaced by a new created population at every generation. Tournament selection is a mechanism for choosing classifiers from a population. A group of classifiers are selected at random from the population and the best one(s) is chosen |
| Initial population | Each tree was created by ramped half-and-half | Ramped half and half operates by creating an equal number of trees with each depth between a pre-determined minimum and maximum. |
| Population size | 20000 | The number of candidate classifiers in a population |
| Number of demes | 12 | A deme is a separately evolving subset of the whole population. The subsets may be evolved on different computers. Emigration between subset may occur every generation. |
| Crossover probability | 0.2 | The probability of creating a new individual from parts of its parents |
| Mutation probability | 0.2 | The probability of a subtree replaced by another, some or all of which is created at random |
| Termination criteria | Fitness score reaches 1 or max generations (50) | A statement or condition to stop the genetic programming cycle. |
| Initial tree depth | 3 | The initial distance of any leaf from the root of a tree |
| Initial node count | 3 | The initial number of nodes in a tree. |
| Maximum tree depth | 7 | The maximum distance of any leaf from the root of a tree |
| Maximum node count | 8 | The maximum number of nodes in a tree. |
| Number of folds | 4 | The number of parts a training set will be subdivided into. |
| Deme migration frequency | Every generation | The frequency of moving classifiers between isolated demes |
| Deme migration percentage | 5% of individuals | The percentage of classifiers moving between two demes. |
| Fitness | the area under the receiver-operator characteristic curve | A process which evaluates a member of a population and gives it a score or fitness. |

*Source of some term descriptions: Langdom, WB. (1998). Genetic Programming and Data Structures: Genetic Programming + Data Structures = Automatic Programming! Amsterdam: Kluwer.

**Table 4.2** Gene expression datasets applied to GP system

| Class description | Authors | Journal | Array type | # of Genes |
|---|---|---|---|---|
| Four classes: neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), Ewing sarcoma (EWS) | Khan, J., et al. | Nature Medicine, 7:673 | cDNA | 2308 |
| Two classes: high-risk group and low-risk group. | Beer, DG et al. | Nature Medicine, 30:41 | Affymetrix Hu6800 | 7070 |
| Two classes: primary prostate cancer (PCA) and metastatic prostate cancer (MET) | Latulippe, E., et al. | Cancer Research. 62:4499 | Affymetrix HG_U95A | 3547 |
|  | Yu, YP., et al. | J Clin Oncol. 22:2790 | Affymetrix HG_U95A | 3547 |
| Two classes: Benign/normal prostate (BENIGN) and primary prostate cancer (PCA) | Lapointe, J. et al. | PNAS. 101:811 | cDNA | 4168 |
|  | Dhanasekaran, SM. et al. | Nature. 412(6849):822 | cDNA | 16965 |
|  | Yu, YP., et al. | J Clin Oncol. 22:2790 | Affymetrix HG_U95A | 12558 |

**Table 4.3** Frequency of gene occurrences in the 1000 GP classifiers for Latulippe et al prostate cancer study

| Probe Set | Gene Symbol | Gene Title | Count[a] | Z-Score |
|---|---|---|---|---|
| 37407_s_at | MYH11 | myosin, heavy polypeptide 11, smooth muscle | 112 | 147.59 |
| 32582_at | MYH11 | myosin, heavy polypeptide 11, smooth muscle | 101 | 133.02 |
| 767_at | MYH11 | myosin, heavy polypeptide 11, smooth muscle | 96 | 126.40 |
| 1197_at | ACTG2 | actin, gamma 2, smooth muscle, enteric | 93 | 122.42 |
| 36931_at | TAGLN | Transgelin | 77 | 101.23 |
| 32755_at | ACTA2 | actin, alpha 2, smooth muscle, aorta | 65 | 85.34 |
| 37576_at | PCP4 | Purkinje cell protein 4 | 62 | 81.36 |
| 774_g_at | MYH11 | myosin, heavy polypeptide 11, smooth muscle | 61 | 80.04 |
| 34203_at | CNN1 | calponin 1, basic, smooth muscle | 31 | 40.30 |
| 36834_at | MOXD1 | monooxygenase, DBH-like 1 | 28 | 36.33 |
| 39333_at | COL4A1 | collagen, type IV, alpha 1 | 27 | 35.01 |
| 773_at | MYH11 | myosin, heavy polypeptide 11, smooth muscle | 24 | 31.03 |
| 38834_at | TOPBP1 | topoisomerase (DNA) II binding protein 1 | 23 | 29.71 |
| 685_f_at | LOC112714 | similar to alpha tubulin | 23 | 29.71 |
| 34878_at | SMC4L1 | SMC4 structural maintenance of chromosomes 4-like 1 | 22 | 28.38 |
| 35970_g_at | MPHOSPH9 | M-phase phosphoprotein 9 | 22 | 28.38 |
| 41137_at | PPP1R12B | protein phosphatase 1, regulatory (inhibitor) subunit 12B | 21 | 27.06 |
| 1884_s_at | PCNA | proliferating cell nuclear antigen | 20 | 25.73 |
| 40407_at | KPNA2 | karyopherin alpha 2 (RAG cohort 1, importin alpha 1) | 20 | 25.73 |
| 32662_at | MDC1 | Mediator of DNA damage checkpoint 1 | 19 | 24.41 |
| 34376_at | PKIG | protein kinase (cAMP-dependent, catalytic) inhibitor gamma | 19 | 24.41 |
| 35742_at | C16orf45 | chromosome 16 open reading frame 45 | 19 | 24.41 |
| 36987_at | LMNB2 | lamin B2 | 19 | 24.41 |
| 39145_at | MYL9 | myosin, light polypeptide 9, regulatory | 18 | 23.09 |
| 38430_at | FABP4 | fatty acid binding protein 4, adipocyte | 17 | 21.76 |
| 1599_at | CDKN3 | cyclin-dependent kinase inhibitor 3 | 16 | 20.44 |
| 2012_s_at | PRKDC | protein kinase, DNA-activated, catalytic polypeptide | 16 | 20.44 |
| 32305_at | COL1A2 | collagen, type I, alpha 2 | 16 | 20.44 |
| 418_at | MKI67 | antigen identified by monoclonal antibody Ki-67 | 16 | 20.44 |
| 651_at | RPA3 | replication protein A3, 14kDa | 16 | 20.44 |
| 35474_s_at | COL1A1 | collagen, type I, alpha 1 | 15 | 19.11 |
| 37749_at | MEST | mesoderm specific transcript homolog (mouse) | 15 | 19.11 |
| 38031_at | DDX48 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 48 | 15 | 19.11 |
| 39990_at | ISL1 | ISL1 transcription factor, LIM/homeodomain, (islet-1) | 15 | 19.11 |
| 1505_at | TYMS | thymidylate synthetase | 14 | 17.79 |
| 33924_at | RAB6IP1 | RAB6 interacting protein 1 | 14 | 17.79 |
| 35694_at | MAP4K4 | mitogen-activated protein kinase kinase kinase kinase 4 | 14 | 17.79 |
| 32306_g_at | COL1A2 | collagen, type I, alpha 2 | 13 | 16.46 |
| 32847_at | MYLK | myosin, light polypeptide kinase | 13 | 16.46 |

|  |  |  |  |  |
|---|---|---|---|---|
| 37305_at | EZH2 | enhancer of zeste homolog 2 (Drosophila) | 13 | 16.46 |
| 41081_at | BUB1 | BUB1 budding uninhibited by benzimidazoles 1 homolog | 13 | 16.46 |
| 32272_at | K-ALPHA-1 | tubulin, alpha, ubiquitous | 12 | 15.14 |
| 36627_at | SPARCL1 | SPARC-like 1 (mast9, hevin) | 12 | 15.14 |
| 37347_at | CKS1B | CDC28 protein kinase regulatory subunit 1B | 12 | 15.14 |
| 39519_at | KIAA0692 | KIAA0692 protein | 12 | 15.14 |
| 40845_at | ILF3 | interleukin enhancer binding factor 3, 90kDa | 12 | 15.14 |

[a] Count is the number of occurrences of each gene in 1000 rules.

**Table 4.4** GP classifiers that distinguish different cancer classes of SRBCT or subtypes of prostate cancer[a]

| Analysis | Classifier | Training Errors | | Test Set Errors | |
|---|---|---|---|---|---|
| | | FN[b] | FP | FN | FP |
| Small ,Round Blue-Cell Tumor | **IF** (HCLS1 – GSTA4 > XPO6) **THEN** BL | 0 | 0 | 0 | 0 |
| | **IF** (PTPN13 / COX8A > CDK6) **THEN** EWS | 0 | 0 | 0 | 0 |
| | **IF** (SATB1 > CSDA^2) **THEN** NB | 0 | 0 | 0 | 0 |
| | **IF** (CDH17/FGFR4 <= MYL4) **THEN** RMS | 0 | 0 | 0 | 0 |
| Primary Prostate Cancer vs. Metastatic Prostat Cancer | **IF** (ARL6IP> MYH11) **THEN** MET | 0 | 0 | 0 | 4 |
| | **IF** (MYH11 < MYH11) **THEN** MET | 0 | 0 | 0 | 4 |
| Lung Cancer (High-risk vs. Low-risk) | **IF** (LTBP2 - IARS) <= (ADM + CCT2 * FCGR2A) **THEN** High-Risk | 0 | 1 | 1 | 0 |
| | **IF** (GYPB - MN1) < (ADM + (MCFD2 + CKS2)) **THEN** High-Risk | 1 | 0 | 3 | 0 |

[a] Only one or two classifiers per class per analysis are listed in the table.

[b] FN: the number of false negatives; FP: the number of false positives

**Table 4.5** GP classifiers that classify benign prostate and primary prostate cancer

| Classifier | The Area Under the ROC Curve (AUC) and Its 95% Confidence Interval | | |
|---|---|---|---|
| | Lapointe et al. Validation Set (Stanford) | Dhanasekaran et al. (UM) | Yu et al. (Pittsburgh) |
| **IF** (ENC1 + GJB1) >= -0.8902 **THEN** PCA | 0.95 (0.87 - 1.00) | 0.95 (0.90 - 1.00) | 0.92 (0.85 - 1.00) |
| **IF** (MYO6 + AMACR) >= -2.6776 **THEN** PCA | 0.95 (0.88 - 1.00) | 0.99 (0.97 - 1.00) | 0.95 (0.90 - 1.00) |
| **IF** (TSPAN13 + PRKCBP1 >= -0.4172 **THEN** PCA | 0.94 (0.85 - 1.00) | 0.88 (0.78 - 0.98) | 0.94 (0.90 - 0.99) |
| **IF** (C20ORF74 + DAPK1) >= -0.7765 **THEN** PCA | 0.91 (0.80 - 1.00) | 0.64 (0.49 - 0.80) | 0.87 (0.79 - 0.95) |
| **IF** (IMAGE:396839 + ENC1) >= -0.5513 **THEN** PCA | 0.97 (0.91 - 1.00) | 0.82 (0.70 - 0.94) | 0.89 (0.81 - 0.98) |
| Meta-classifier | 0.96 (0.87 - 1.00) | 0.99 (0.96 - 1.00) | 0.99 (0.98 - 1.00) |

**Table 4.6** The misclassification error rates of genetic programming (GP) and other common classification models

| Error Rate (%)<br>Algorithm | SRBCT (BL vs. NON-BL) | | Lung Cancer (High-Risk vs. Low-Risk) | |
|---|---|---|---|---|
| | 5 Genes | 10 Genes | 5 Genes | 10 Genes |
| Genetic Programming | 1.5 | N/A | 16 | N/A |
| Compound Covariate Predictor | 5 | 5 | 20 | 25 |
| 3-Nearest Neighbors | 5 | 5 | 15 | 30 |
| Nearest Centroid | 5 | 5 | 20 | 25 |
| Support Vector Machines | 0 | 5 | 20 | 25 |
| Diagonal Linear Discriminant Analysis | 5 | 5 | 10 | 20 |

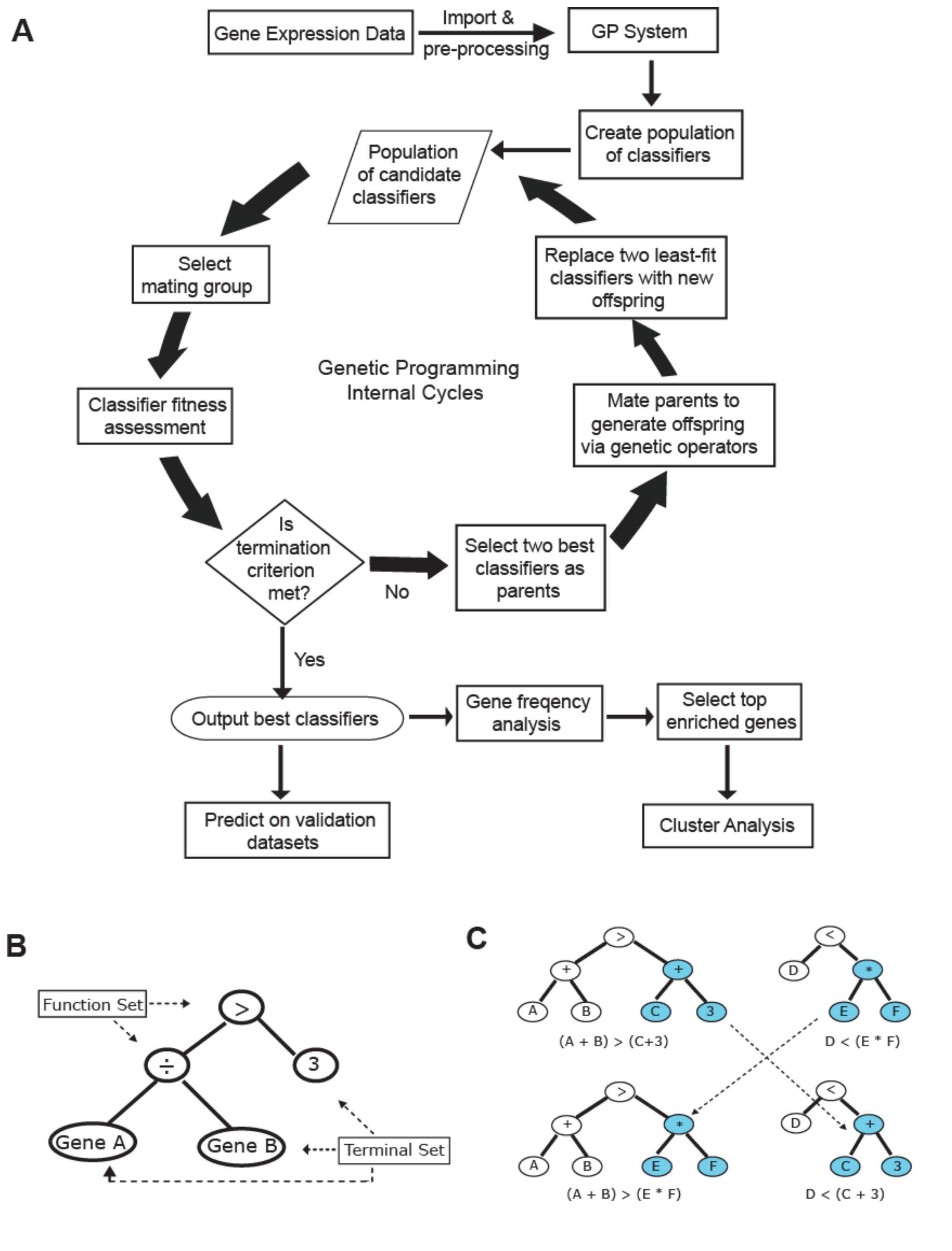

A
Gene Expression Data
Import &
pre-processing
GP System
Create population of classifiers
Population of candidate classifiers
Select mating group
Classifier fitness assessment
Genetic Programming Internal Cycles
Is termination criterion met?
No
Select two best classifiers as parents
Mate parents to generate offspring via genetic operators
Replace two least-fit classifiers with new offspring
Yes
Output best classifiers
Gene freqency analysis
Select top enriched genes
Predict on validation datasets
Cluster Analysis
B
Function Set
>
÷
3
Gene A
Gene B
Terminal Set
C
(A + B) > (C+3)
D < (E * F)
(A + B) > (E * F)
D < (C + 3)

**Figure 4.1**. A flowchart of the genetic programming (GP) process. **A.** Briefly, a population of tree-based classifiers is first created by randomly choosing gene expression data or constant values and combining with arithmetic or Boolean operators. An example of tree-based classifiers is represented in **B**. A small subgroup of classifiers is then selected as a “mating group” and each classifier in this “mating group” is assessed by a fitness function, which is defined as the area under the receiver-operator characteristic curve (ROC-AUC) in this study. The two fittest classifiers are then selected as ‘mating’ parents and ‘mated’ to produce ‘offspring’ by genetic operators (*crossover*, or *mutation*). The generated offspring then replace the least-fit parent classifiers within the population. A new generation of population is generated once the offspring fully replaced ‘parent’ classifiers in the population. This process of mating pool selection, fitness assessment, mating and replacement is repeated over generations, progressively creating better classifiers until a completion criterion is met. After the best classifiers are outputted, post-GP analyses are carried out to compute gene occurrence in the classifiers as well as to predict on new unknown samples. **B.** The representation of a genetic programming (GP) tree structure for an exemplified classifier, Gene[A] / Gene[B] >3. In general, a GP classifier is represented as a tree-based structure composed of the terminal set and function set. The terminal set, in tree terminology, are leaves (nodes without branches) and may represent as genes or constants. The function set is a set of operators such as arithmetic operators (+, −, ×, ÷) or Boolean operators (AND, OR, NOT), acting as the branch points in the tree, linking other functions or terminals. **C.** The representation of a *crossover* operator of GP tree.

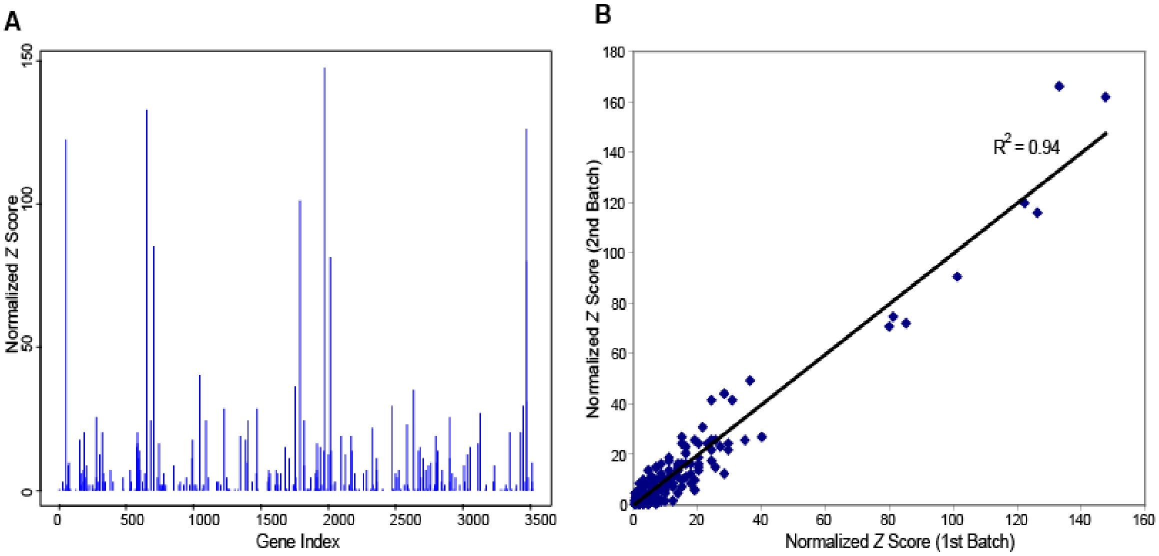


**Figure 4.2**. Feature selection in genetic programming. A. The statistical z-score of each of the 3547 genes occurring in the 1000 classifiers generated from LaTulippe et al. prostate cancer study by GP based on the parameters listed in Table 2. Let $Z=[X_i-E(X_i)]/\sigma$, where $X_i$ is the frequency times gene $i$ is selected, $E(X_i)$ is the expectation of frequency times gene $i$ is selected, $\sigma$ is the standard deviation of this binomial model. Let, n=1000, p, the probability of gene $i$ being selected randomly, is approximately equal to the total counts of frequency in 1000 classifiers divided by the number of classifiers (1000), then divided by the total number of genes (3547), then $E(X_i)$=np, and $\sigma = \sqrt{}$[np(1-p)]. B. Correlation between commonly occurring genes on two independent sets of classifiers. Each set contains 1000 classifiers.

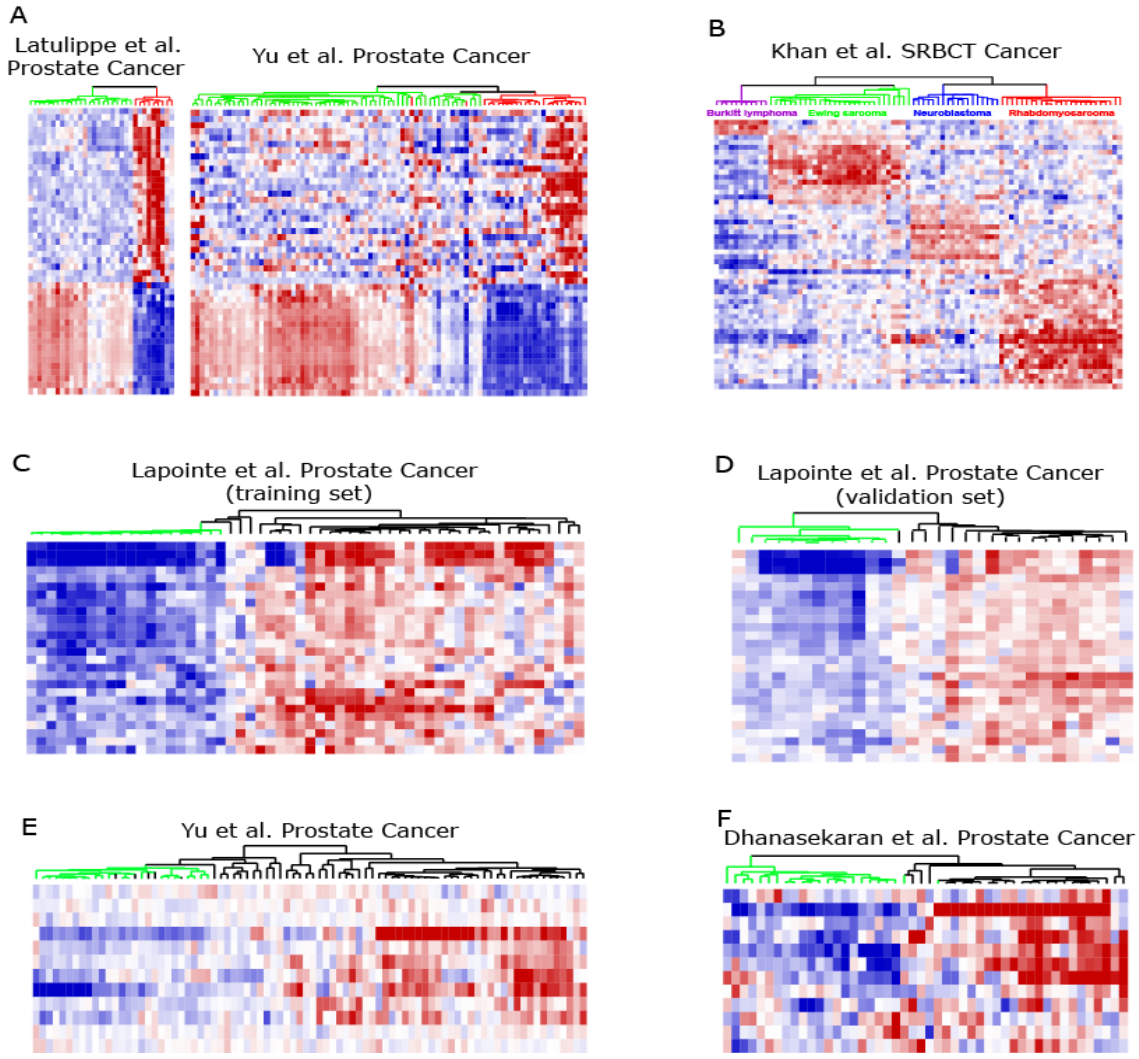


**Figure 4.3**. Top feature genes derived from GP separate tumors into their corresponding diagnostic classes. **A.** Hierarchical clustering using the top 46 most frequent genes derived from the 1000 classifiers generated for LaTulippe et al. (MSKCC) dataset. Genes were ranked by the frequency of their occurrences in the classifiers. The top 46 frequent genes (*z* score >=15, see **Figure 4.2**) were selected for hierarchical clustering. The left panel is the clustering of metastatic samples and primary prostate cancer samples for the MSKCC data, and the right panel is for the Yu et al. (Pittsburgh) validation dataset. Rows represent genes and columns represent samples. The green lines in the dendrogram indicate primary prostate cancer and the red lines represent metastatic prostate cancer samples. **B.** Hierarchical clustering of the entire SRBCT dataset using the top 54 feature genes obtained from the training set. **C-F**. The top 26 most frequent genes from the 2000 classifiers generated from Lapointe et al. (Stanford) training set was used to separate benign/control prostate samples from primary prostate cancer samples in the Stanford training set (**C**), Stanford validation set (**D**), Yu et al. (Pittsburgh) validation set **(E)**, and Dhanasekaran et al. (UM) validation set (**F**) respectively. The green lines in the dendrogram indicate benign/control prostate samples and the black lines represent primary prostate cancer samples.

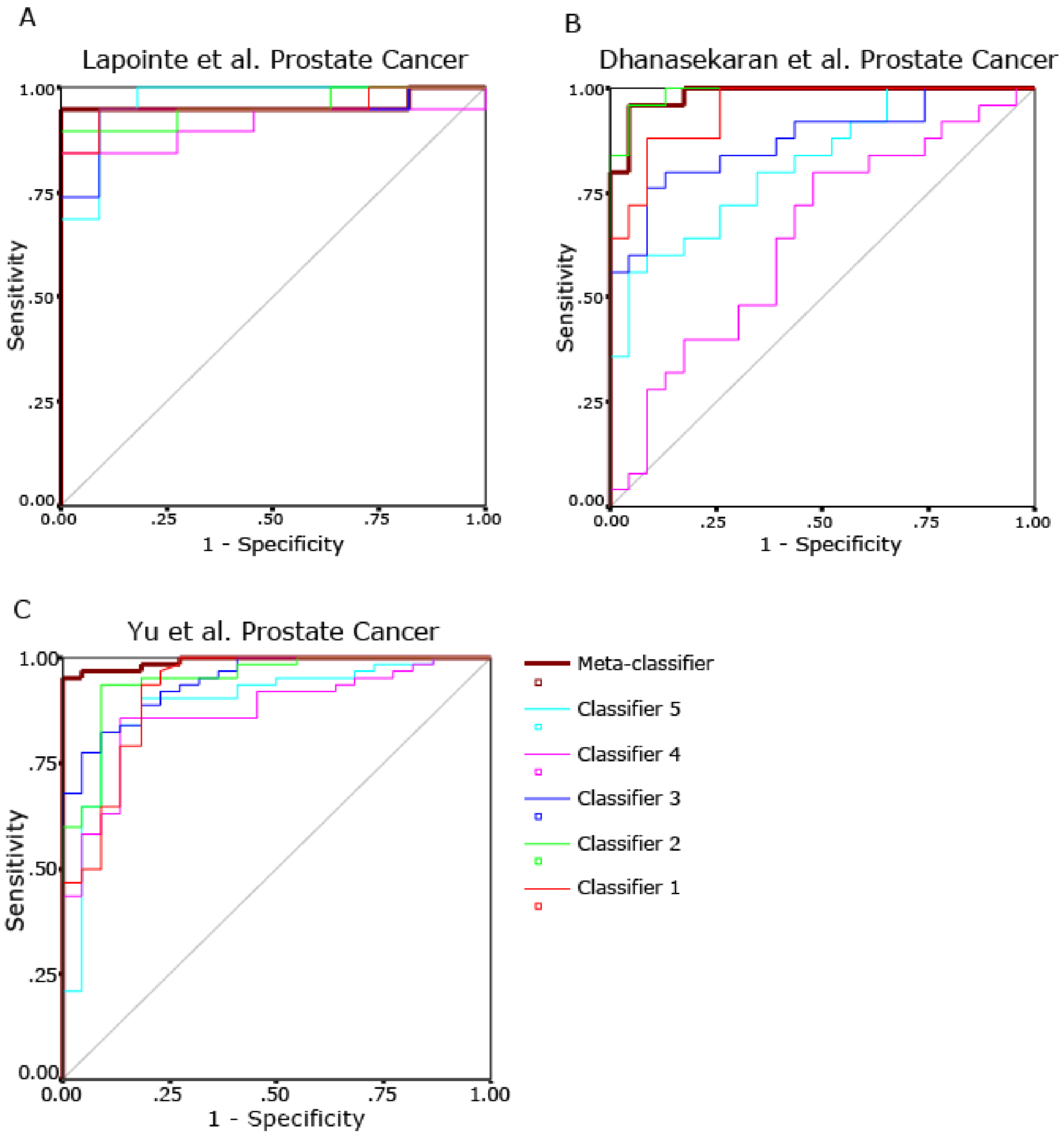


**Figure 4.4**. The receiver-operator characteristic curves (ROCs) of five classifiers and one meta-classifier for three prostate cancer validation sets. The classifiers were generated from the Lapointe et al. (Stanford) training set to distinguish benign prostate from primary prostate cancer. The ROCs are based on continuous prediction scores computed from the left side of the classifier inequality (see Methods). The scores of the meta-classifier are the mean values of prediction scores from each individual classifier. **A.**, **B.**, **C.** represents the ROC curve for Lapointe et al. (Stanford), Dhanasekaran et al. (UM), and Yu et al. (Pittsburgh) validation set respectively.

# REFERENCES

1. Alizadeh, AA, Eisen, MB, Davis, RE, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

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

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

4. Hedenfalk, I, Duggan, D, Chen, Y, et al. (2001). Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344**, 539-548.

5. Perou, CM, Sorlie, T, Eisen, MB, et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747-752.

6. Koza, JR. (1992). Genetic Programming: on the Programming of Computers by Means of Natural Selection.: Cambridge: MIT Press).

7. Bojarczuk, CC, Lopes, HS, and Freitas, AA (2001). Data mining with constrained-syntax genetic programming: applications to medical data sets. *Proceedings Intelligent Data Analysis in Medicine and Pharmacology ({IDAMAP}-2001).*

8. Tan, KC, Yu, Q, Heng, CM, and Lee, TH (2003). Evolutionary computing for knowledge discovery in medical diagnosis. *Artif Intell Med* **27**, 129-154.

9. Hong, JH, and Cho, SB (2006). The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. *Artif Intell Med* **36**, 43-58.

10. Langdon, WB, Buxton, B.F. (2004). Genetic Programming for Mining DNA Chip Data from Cancer Patients. *Genetic Programming and Evolvable Machines* **5**, 251-257.

11. Mitra, AP, Almal, AA, George, B, et al. (2006). The use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. *BMC Cancer* **6**, 159.

12. Moore, JH, Parker, JS, and Hahn, LW (2001). Symbolic discriminant analysis for mining gene expression patterns. *Lecture Notes in Artificial Intelligence* **2167**, 372-381.

13. Ho, SY, Hsieh, CH, Chen, HM, and Huang, HL (2006). Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. *Biosystems* **85**, 165-176.

14. Moore, JH, Parker, JS, Olsen, NJ, and Aune, TM (2002). Symbolic discriminant analysis of microarray data in autoimmune disease. *Genet Epidemiol* **23**, 57-69.

15. LaTulippe, E, Satagopan, J, Smith, A, et al. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* **62**, 4499-4506.

16. Li, L, Weinberg, CR, Darden, TA, and Pedersen, LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **17**, 1131-1142.

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

18. Kleer, CG, Cao, Q, Varambally, S, et al. (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A* **100**, 11606-11611.

19. Varambally, S, Dhanasekaran, SM, Zhou, M, et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629.

20. Khan, J, Wei, JS, Ringner, M, et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7**, 673-679.

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

22. Lapointe, J, Li, C, Higgins, JP, et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816.

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

24. Armstrong, SA, Staunton, JE, Silverman, LB, et al. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **30**, 41-47.

25. Gruvberger, SK, Ringner, M, Eden, P, et al. (2002). Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res* **5**, 23-26.

26. Mukherjee, S, Tamayo, P, Mesirov, JP, et al. (1999). Support vector machine classification of microarray data: MIT, CBCL).

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

28. Bhattacharyya, C, Grate, LR, Jordan, MI, El Ghaoui, L, and Mian, IS (2004). Robust sparse hyperplane classifiers: application to uncertain molecular profiling data. *J Comput Biol* **11**, 1073-1089.

29. Rissanen, J (1978). Modeling by shortest data description. *Automatica* **14**, 465-471.

30. Vapnik, VN (1995). The Nature of Statistical Learning Theory (Berlin: Springer-Verlag).

31. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

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

# PART 3: NON-INVASIVE TESTS FOR PROSTATE CANCER DIAGNOSIS

## CHAPTER 5

### Autoantibody Signatures in Prostate Cancer

New biomarkers, such as autoantibody signatures, may improve the early detection of prostate cancer. With a phage-display library derived from prostate-cancer tissue, we developed a phage protein microarray platform to analyze serum samples from 119 patients with prostate cancer and 138 controls, with the samples equally divided into training and validation sets. A 22-phage-peptide detector that was constructed from the training set was evaluated on an independent validation set of 128 serum samples (60 from patients with prostate cancer and 68 from controls). This phage-peptide detector resulted in 88.2% specificity (95% Confidence Interval [CI], 0.78 to 0.95) and 81.6% sensitivity (95% CI, 0.70 to 0.90) in discriminating between the group with prostate cancer and the control group in the validation set. This panel of peptides also performed better than did prostate-specific antigen (PSA) in term of classification accuracy (area under the ROC-curve for the autoantibody signature, 0.93; 95% CI, 0.88 to 0.97; area under the curve for PSA, 0.80; 95% CI, 0.71 to 0.88). Logistic-regression analysis revealed that the phage-peptide panel provided additional discriminative power over PSA ($P<0.001$). Taken together, this study demonstrated that autoantibodies against peptides derived from prostate-cancer tissue could be used as the basis for a screening test for prostate cancer.

Limitations of the prostate-specific antigen (PSA) test for the early detection of prostate cancer (1) indicate the need for other means of screening for this neoplasm. The finding that patients with cancer produce autoantibodies against antigens in their tumors (2-7) suggests that such autoantibodies could have diagnostic and prognostic value (2,6,8-10). For example, mutant forms of the p53 protein elicit anti-p53 antibodies in 30 to 40% of patients with various types of cancers (11). Recently, we found that patients with prostate cancer produce antibodies against alpha-methylacyl-coenzyme A racemase (12), an overexpressed protein in epithelial cells of prostate cancer (13-15). This autoantibody had 72% specificity and 62% sensitivity in detecting prostate cancer (12), indicating that the use of additional prostate-cancer antigens could improve the sensitivity and specificity of an autoantibody-based screening test for prostate cancer.

Here we report the use of phage-display microarrays to identify and characterize new autoantibody-binding peptides derived from prostate-cancer tissue. A similar approach has been used to identify selected antigens for the diagnosis of breast cancer (16). To develop a phage-display library of prostate-cancer peptides, we isolated mRNAs of prostate cancer tissues and inserted the synthesized cDNA fragments into the T7 phage system. Peptides that were encoded by the prostate cancer cDNAs were expressed and displayed on the surface of the phage fused to the C-terminal of the capsid 10B protein of the phage. This surface complex functioned as bait to capture autoantibodies in serum. To enrich the library for peptides that bind specifically to autoantibodies in patients with prostate cancer, we carried out successive rounds of selection and purification, termed biopanning (**Figure 5.1**). Phage clones, each bearing a single fusion peptide derived from the prostate-cancer cDNA library, were then selected randomly from the purified library to generate protein microarrays on coated glass slides with the use of a robotic spotter. Once in a microarray format, the enriched phage clones were used to test serum for autoantibodies against prostate cancer peptides.

Initially, we constructed a high-density phage-display microarray containing 2304 individual phage clones. Five empty clones were also included as negative controls. To decrease the complexity of subsequent validation studies, we sought to develop a focused array on the basis of the initial high-density arrays. We randomly selected 20 serum samples from cancer patients and 11 control samples from the University of Michigan collections. After normalization of all values obtained by the scanner, we selected 186 phage-peptide clones that yielded a ratio of Cy5 to Cy3 greater than 1.2 in at least one of the serum samples from prostate cancer patients. These clones, along with negative-control phage clones, were used to construct a smaller, focused protein microarray for subsequent screening of serum samples.

**Figure 5.2** shows the training and validation phases of this study. A total of 257 serum samples from 119 patients with clinically localized prostate cancer and 138 controls, were tested on the 186-element focused arrays (**Table 5.1**). In the training phase, we analyzed 59 samples from patients with prostate cancer and 70 control samples

(**Figure 5.2**). To identify a best subset of clones for detection of prostate cancer, we used a nonparametric-pattern-recognition approach that consisted of a genetic algorithm combined with k-nearest neighbour to select a subgroup of "informative" phage peptides from the training set based on leave-one-out cross-validation (17). Unlike individual feature ranking methods, this approach allows us to find a group of clones that serve together to maximize the classification performance even though individual clones may not be highly correlated with the diagnostic status of prostate cancer. Using this approach, we identified a panel of 22 phage-peptide clones that best distinguished cancer patients from control subjects, with 97.1% specificity and 88.1% sensitivity in the training set. **Figure 5.3A** shows a heatmap of the results with the 22 phage-peptide clones in the training set.

We next sought to apply this panel of 22 phage peptide clones into an independent validation set. By using a weighted voting scheme, we applied these peptides as a class detector to classify samples in the independent validation set (128 patients) as either prostate cancer or control (see Methods). Notably, only 8 of 68 serum samples from controls and 11 of 60 samples from patients with prostate cancer were misclassified in this validation set (**Figure 5.3B**), resulting in a specificity of 88.2% (95% CI, 78% to 95%) and a sensitivity of 81.6% (95% CI, 70% to 90%). We also observed similar performance when using different prediction methods and resampled datasets.

To examine whether if the 22-phage-peptide detector performs better than the conventional PSA test, we next calculated receiver-operating-characteristic (ROC) curves for both of them in the validation set. For the 22-phage-peptide detector, different threshold values of weighted voting scores were used as cutoff points to plot the true positive rate against the false positive rate. For the entire validation set, the ability of the panel of 22 phage peptides to discriminate between prostate-cancer samples and control samples was significant (P<0.001), with an area under the curve equal to 0.93 (95% CI, 0.88 to 0.97) (**Figure 5.3C**). The area under the curve for PSA was 0.80 (P<0.001; 95% CI, 0.71 to 0.88), which was expected as these patients were identified primarily by elevated PSA levels. Among patients with PSA levels of 4 to 10 ng per milliliter in the validation set, the phage-peptide detector remained significant discriminatory power (P<0.001) as compared with PSA (P=0.50) in distinguishing serum samples from patients with prostate cancer from those of controls. The area under the curve was 0.93 (95% CI, 0.86 to 1.00) for the phage-display method and 0.56 (95% CI, 0.38 to 0.74) for PSA (**Figure 5.3D**). When the lower limit of PSA was decreased to 2.5 ng per milliliter, the discriminatory power of the phage-peptide profile was maintained (P<0.001), with an area under the curve of 0.94 (95% CI, 0.88 to 1.00), whereas that for PSA decreased slightly to 0.50 (95% CI, 0.33 to 0.66) (**Figure 5.3E**).

To compare the ROCs for the 22 phage-peptide predictor and PSA, a permutation test was performed based on the difference between the AUCs for the two diagnostic techniques accounting for the fact that the same samples were used for both assays. For the entire validation set, the difference in AUCs was significant (p < 0.001). In addition, the difference was also significant for the subjects with PSA level between 4-10 ng/ml (p = 0.004) and with PSA level between 2.5-10 ng/ml (p < 0.001).

To evaluate whether the 22 phage-peptides predictor is a useful supplement to PSA, we performed logistic regression on the validation set. We first used cancer diagnostic status (cancer/non-cancer) as the response and carried out univariate logistic regression for the standardized weighted voting scores and PSA respectively. We found that both tests are statistically significant (odds ratio [OR] for the voting scores = 74.22, 95% CI = 16.17-340.67, p<0.001; OR for PSA = 4.17, 95% CI = 2.05-8.47, p<0.001). Next, we performed multivariate logistic regression with disease as the response and fit both the voting scores and PSA as covariates. We found that the effect of voting scores was strongly significant (OR = 47.69; 95% CI = 9.47-240.21; p<0.001) after adjusting for the effect of PSA (OR = 2.91; 95% CI = 1.29-6.56; p=0.01), indicating that the 22 phage-peptide predictor provides additional predictive value over preoperative PSA level.

We next sequenced the panel of 22 phage-peptide clones. Of these, four were in-frame and within known expressed transcripts, including bromodomain-containing protein 2 (BRD2), eukaryotic translation initiation factor 4 gamma 1 (eIF4G1), ribosomal protein L22 (RPL22), and ribosomal protein L13a (RPL13a). The others were not present in peptide stretches in known proteins. These clones may be weakly homologous to known proteins or may have no distinct homology to the primary sequences of known proteins and thus may be "mimotopes" (i.e., stretches of amino acids that mimic an antigen but are not homologous at the sequence level). To examine whether the four in-frame phage-peptide clones (**Figure 5.4A**) are deregulated in prostate cancer at the transcript level and protein level, we performed a meta-analysis of publicly available gene expression datasets in prostate cancer (18-24) as well as a preliminary immunoblot analysis. These analyses suggested that the four in-frame phage epitopes are overexpressed in prostate cancer (**Figure 5.4B** and **Figure 5.4C**).

The use of PSA-based screening for prostate cancer has risen dramatically since its introduction in the late 1980s (25,26). However, reliance on PSA for the detection of early prostate cancer is still unsatisfactory, especially because of a high rate of false positive results (27) — as high as 80% (28,29). This rate results in many unnecessary prostate biopsies (30). To circumvent this and other problems of screening for prostate cancer, we have begun to evaluate the use of autoantibody signatures to detect prostate cancer.

In this study, we used protein microarrays to identify autoantibodies against tumor antigens in patients with prostate cancer. Specifically, we constructed phage-protein microarrays in which peptides derived from a prostate-cancer cDNA library were expressed as a prostate-cancer – phage fusion protein. The phage-protein microarrays were screened to identify phage-peptide clones that bind autoantibodies in serum samples from patients with prostate cancer but not in those from controls. By relying on multiple immunogenic prostate-cancer peptides, this approach may improve the accuracy of prostate cancer diagnosis over a single biomarker such as PSA.

Our results were consistent across a range of clinical and pathological features, including PSA level, Gleason grade, stage, and presence or absence of PSA, with sensitivities and specificities ranging from 80 to 90% in discriminating between patients with prostate cancer and controls. In addition, this diagnostic performance was maintained in the intermediate ranges of PSA (either 4 to 10 ng per milliliter or 2.5 to 10 ng per milliliter). In addition, our data revealed that the 22-phage-peptide detector significantly increased the diagnostic power of PSA alone ($P<0.001$), suggesting that our autoantibody signature may be useful in combination with initial PSA screening to improve decision making in biopsy of the prostate.

We have not tested the phage-microarray system for screening for prostate cancer; this requires extension and confirmation in community-based screening cohorts. Furthermore, it will be important to evaluate the autoantibody signatures associated with prostate cancer in patients with prostatitis, autoimmune conditions, and other diseases. Although the technique is promising, how it will perform in prospective and multi-institutional studies remains to be determined.

**Methods**

Populations and Samples. This study, which was approved by the institutional review board of the University of Michigan Medical School, started in March 2003 and ended in December 2004. It had discovery, training, and validation phases. All serum samples, unless otherwise indicated, were obtained from patients in the University of Michigan Health System. Written informed consent was obtained from all patients.

In the discovery phase (biopanning and 2304-element microarrays), 39 prostate-cancer samples and 21 control samples were used. The training phase involved the use of 59 prostate-cancer samples and 70 control samples. To evaluate the phage-peptide detectors that we developed in the discovery and training phases, we used an independent validation set of 60 prostate-cancer samples (48 from the University of Michigan and 12 from the Dana-Farber Cancer Institute) and 68 control samples. In the 257 prostate-cancer samples and control samples (which included the training and validation sets, **Table 5.1**), the median levels of PSA were 6.3 ng per milliliter (range, 0.1 to 46.3) and 1.7 ng per milliliter (range, 0.1 to 24.5), respectively.

Autoantibody Profiling. By iterative biopanning of a phage-display library derived from prostate-cancer tissues, we developed phage protein microarrays and used them to develop an autoantibody signature to distinguish samples with prostate cancer from those of controls. Details concerning the construction of phage-display libraries and preparation of the phage-protein microarrays are shown in **Figure 5.1**.

Normalization and Analysis of the Microarray Data. Slides were scanned and quantified using GenePix 4000B scanner (Axon Laboratories). The Cy5/Cy3 ratios were calculated for each phage spot, and values for duplicate spots were averaged. The difference between duplicates was <5% for 98% of the spots. Analyses of repeated experiments using same serum samples revealed that the results were very consistent with correlation coefficient greater than 0.9. According to the experimental design, Cy5/Cy3 was utilized so as to control the small variations in the amount of phage particles being spotted. The ratio of Cy5/Cy3 for each spot was subtracted by median of Cy5/Cy3 of the negative T7 empty spots with the observation that the signal for the T7 empty phage on each chip highly correlated with the signal intensity for the whole array. A Z-transformation was applied to data such that the mean of each clone was zero across arrays and the variance was 1.0. Normalized data was then subjected to two-way clustering analysis with use of Cluster and TreeView.

Development of Phage-Peptide Predictor. By employing 186-element phage-peptide microarray platform, 257 sera samples were tested. These samples were divided into training and validation set. Training set was used to build a class prediction model by a leave-one-out-cross-validation (LOOCV) strategy in Genetic Algorithm/K-Nearest Neighbor (GA/KNN) (k=3 in this study) method (17). The raw data was normalized as described above. The normalized array data was then applied to GA for selection of the clones and assessment of their relative predictive importance by ranking them based on their frequency of occurrence in GA solutions with the top-most clone assigned a rank of 1. Different numbers of the top-most clones were used to build different KNN prediction models. Misclassification error rates were calculated using LOOCV to evaluate the performance of the models. As few as 10 phage clones performed with similar accuracy, but to maintain a diversity of clones for validation, we used the 22 phage-peptide predictor, which yielded the minimal misclassification error ratessensitivity during LOOCV. For the validation sample set, a weighted voting scheme was adopted, similar to that described previously (31). Briefly, let class 0 and class 1 represent non-cancer and cancer samples, respectively. Each informative phage clone, derived from the training set, casts a weighted vote for a class 0 or 1: $v_x = T_x (e_x - b_x)$ where $e_x$ is the signal value of phage peptide *x* for each individual validation sample on array images, $T_x$ is the t-statistic for comparing the two class means of phage *x* in the training set, and $b_x$ is $(\mu_0 + \mu_1)/2$, where $\mu_0$, and $\mu_1$ denote the means of phage *x* for class 0 or 1 in the training set. A negative $v_x$ indicates a vote for class 0 and a positive value indicates a vote for class 1. The total vote $V_0$ for class 0 is obtained by summing the absolute values of the negative votes over the informative phage-peptides, while the total vote $V_1$ for class 1 is obtained by summing the absolute values of the positive votes. The final voting score $V_s$ is $V_1$-$V_0$ and the final vote for class 0 or 1 is sign ($V_s$) and the confidence in the prediction of the winning class is $|V_1-V_0| / (V_0+V_1)$, where $V_i$ is the vote for class *i*.

Sequence Analysis of 22 Phage Clones. The top 22 phage clones were amplified by PCR using T7 capsid forward and reverse primers (Novagen). Briefly, 2 µl of fresh phage lysate with titer of ~ 1010 pfu was incubated with 100 µl of 10 mM EDTA, pH 8.0 at 60 ℃ for 10 min. After centrifuging at 14,000 g for 3 min, 2 µl of denatured phage was used for PCR in 100 µl volume of reaction under standard condition. PCR products were confirmed on 1% agarose gel containing ethidium bromide. After purifying with MultiScreen-FB filter plate (Millipore) following manufacturer's protocol, PCR products were sequenced using T7 capsid forward primer to determine the cDNA inserts. DNA sequence and translated protein sequence were aligned using NCBI BLAST.

Meta-Analysis of Gene Expression. The gene expression level of four genes, namely BRD2, eIF4G1, RPL13a and RPL22, were studied using ONCOMINE (22). Briefly, each gene was searched on the database, and the results were filtered by selecting prostate cancer. The data from study classes of benign prostate, prostate cancer and / or metastatic prostate cancer with $p<0.05$ were used to plot the box plots with SPSS11.5. *P* values for each group were calculated using student t-test.

**Table 5.1** Clinical and pathology information for the training and validation samples.

| **Variable** | Training set | Validation set |
|---|---|---|
| **Clinically localized prostate cancer patients** | | |
| No. of patients | 59 | 60 |
| Mean age (yr) ± SD | 58.3 ± 7.7 | 60.81 ± 9.0 |
| Mean gland weight (g) ± SD | 49.55 ± 17.17 | 51.78 ± 19.57 |
| Dim. Of max tumor (cm) ± SD | 1.44 ± 0.75 | 1.62 ± 0.97 |
| PSA level (ng/ml) | | |
| Mean ± SD | 6.19 ± 4.58 | 10.45 ± 9.52 |
| 0 - 2.4 (%) | 17.2 | 7.7 |
| 2.5 -10 (%) | 67.2 | 53.8 |
| 4 - 10 (%) | 50 | 42.3 |
| > 10 (%) | 15.5 | 38.5 |
| ***Gleason grade (%)*** | | |
| <= 6 | 35.7 | 37.3 |
| >= 7 | 64.3 | 62.7 |
| ***Primary tumor identification (%)*** | | |
| T2a | 29.8 | 43.7 |
| T2b | 59.6 | 41.7 |
| T3a | 3.5 | 2.1 |
| T3b | 7 | 12.5 |
| **Control subjects with no known history of cancer** | | |
| No. of patients | 70 | 68 |
| Mean age (yr) ± SD | 62.8 ± 8.6 | 63.6 ± 9.3 |
| PSA level (ng/ml) | | |
| Mean ± SD | 2.88 ± 2.57 | 3.01 ± 2.68 |
| 0 - 2.4 (%) | 61.4 | 59.7 |
| 2.5 -10 (%) | 38.6 | 34.3 |
| 4 - 10 (%) | 32.9 | 29.9 |
| > 10 (%) | 0 | 5.9 |

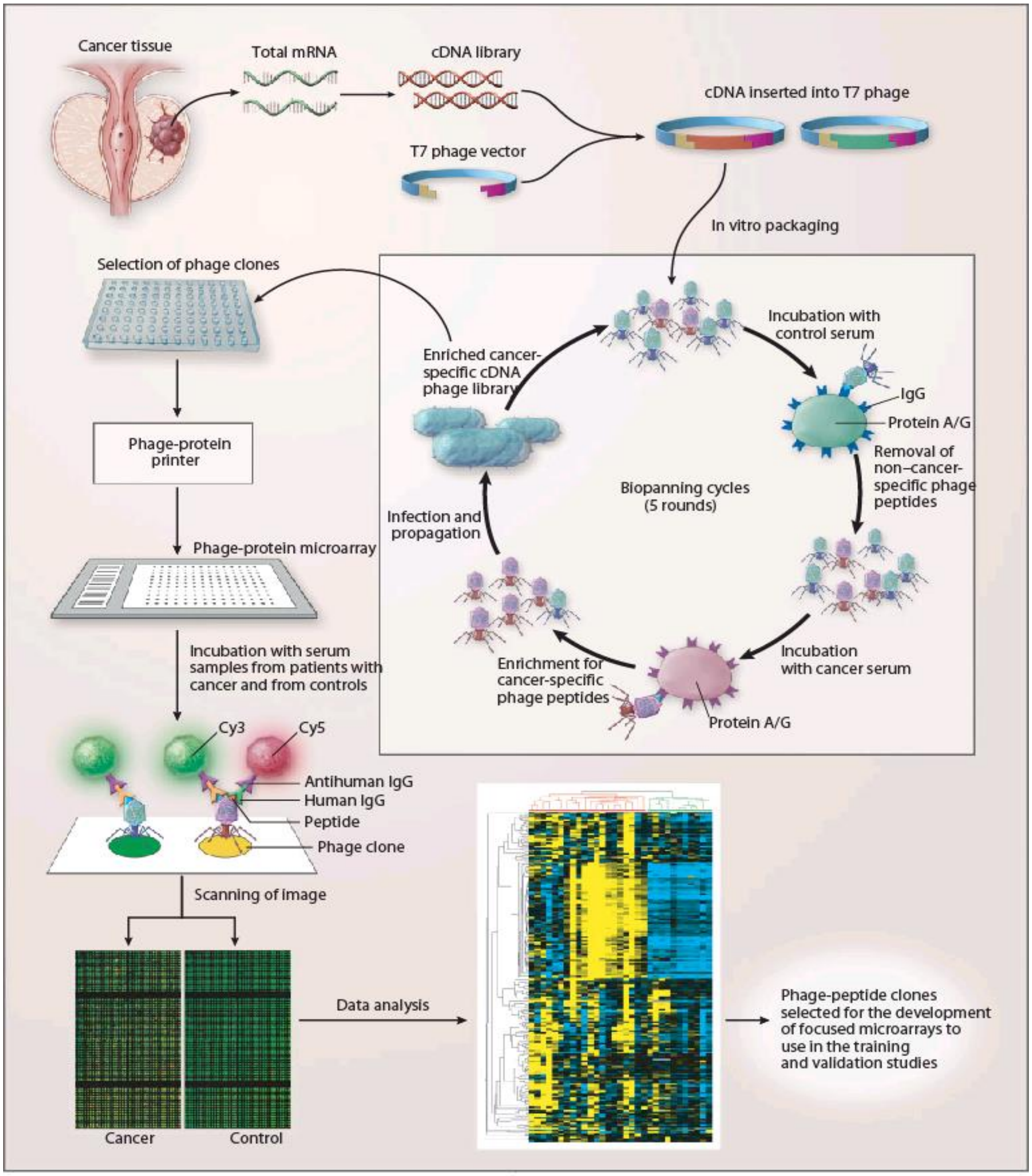

Cancer tissue
Total mRNA
cDNA library
cDNA inserted into T7 phage
T7 phage vector
In vitro packaging
Selection of phage clones
Incubation with control serum
Enriched cancer-specific cDNA phage library
IgG
Protein A/G
Removal of non–cancer-specific phage peptides
Phage-protein printer
Biopanning cycles (5 rounds)
Infection and propagation
Phage-protein microarray
Incubation with serum samples from patients with cancer and from controls
Incubation with cancer serum
Enrichment for cancer-specific phage peptides
Protein A/G
Cy3
Cy5
Antihuman IgG
Human IgG
Peptide
Phage clone
Scanning of image
Data analysis
Phage-peptide clones selected for the development of focused microarrays to use in the training and validation studies
Cancer
Control

**Figure 5.1**. Schematic representation of the development of phage-protein microarrays to characterize autoantibody signatures in prostate cancer. A cDNA library was constructed from a pool of total mRNA isolated from prostate-cancer tissue obtained from six patients. After digestion, the cDNA library was inserted into the T7 phage vector. The T7 fusion vectors were then packaged into T7 phages to generate a prostate-cancer cDNA T7 phage-display library. To enrich the library with clones of peptides reacting with human serum from patients with clinically localized prostate cancer and not with serum from controls, several cycles of affinity selections (biopanning) were performed. Briefly, the phage libraries were preadsorbed onto purified IgGs from the control pool of serum samples (from 10 patients) to remove nonspecific clones. Next, the precleared phage libraries were enriched for cancer-specific peptides with the use of a pool of IgGs purified from the serum of 19 patients with prostate cancer. The bound phages were eluted and propagated by infecting bacterial cells. After five rounds of biopanning, enriched prostate-cancer–specific peptide clones were cultured onto LB agar plates. A total of 2304 single colonies, including T7 empty phage clones as negative spots and antihuman IgG as positive spots, were randomly picked and propagated into 96-well plates. Phage-clone lysates were then printed onto coated glass slides with the use of a robotic spotter to create a phage-protein microarray. Cy5 (red fluorescent dye)–labeled antihuman antibody was used to detect IgGs in human serum that were reactive to peptide clones, and a Cy3 (green fluorescent dye)–labeled antibody was used to detect the phage capsid protein in order to normalize for spotting. Thus, if a phage clone carries a peptide that is reactive to human IgG, after scanning, this spot will be yellow in color; otherwise, the spot will remain green, representing an unreactive clone. A total of 31 samples (20 from patients with cancers and 11 from controls) were tested on the 2304 phage-peptide microarray. Analysis of these 31 samples identified 186 phage peptides with the highest level of differentiation between cancers and controls, which were then used to develop focused microarrays for analyses in the subsequent training and validation phases.

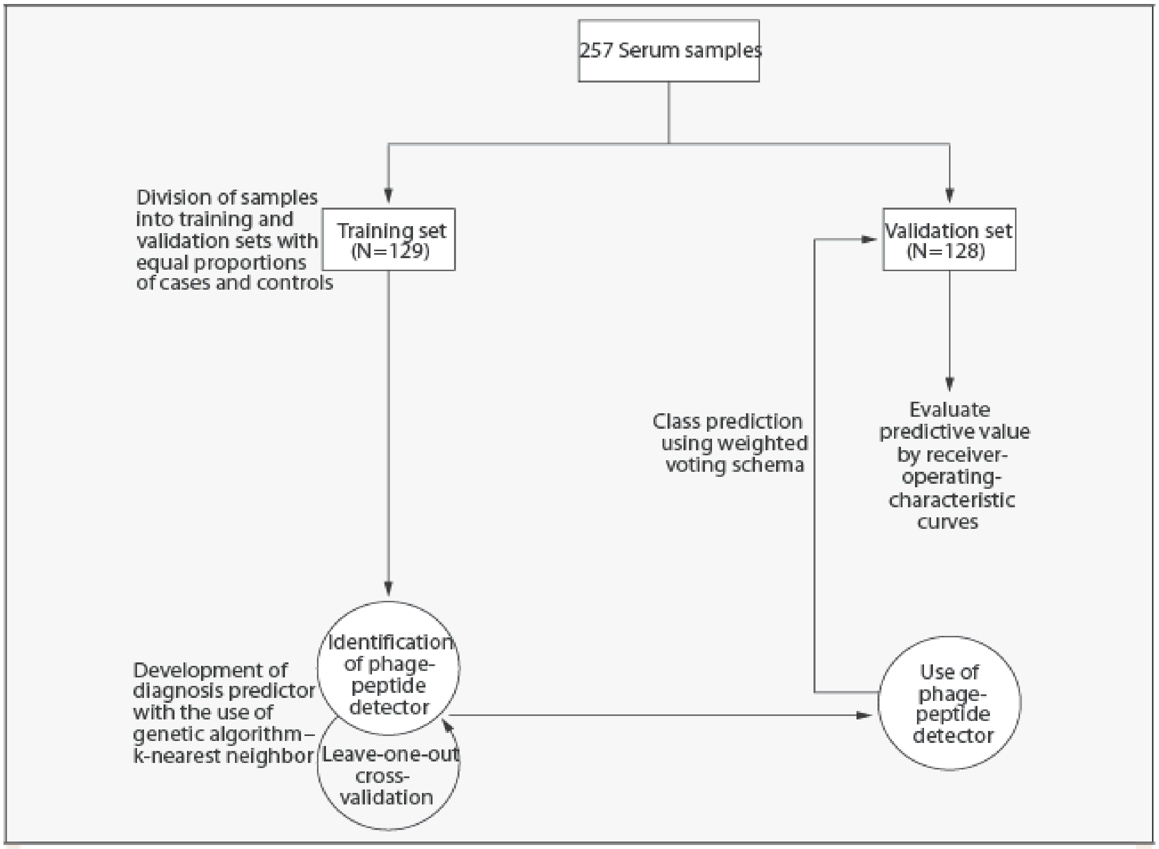

257 Serum samples
Division of samples into training and validation sets with equal proportions of cases and controls
Training set (N=129)
Validation set (N=128)
Class prediction using weighted voting schema
Evaluate predictive value by receiver-operating-characteristic curves
Development of diagnosis predictor with the use of genetic algorithm–k-nearest neighbor
Identification of phage-peptide detector
Leave-one-out cross-validation
Use of phage-peptide detector

**Figure 5.2**. Overview of the strategy used for the development and validation of autoantibody signatures to identify prostate cancer.

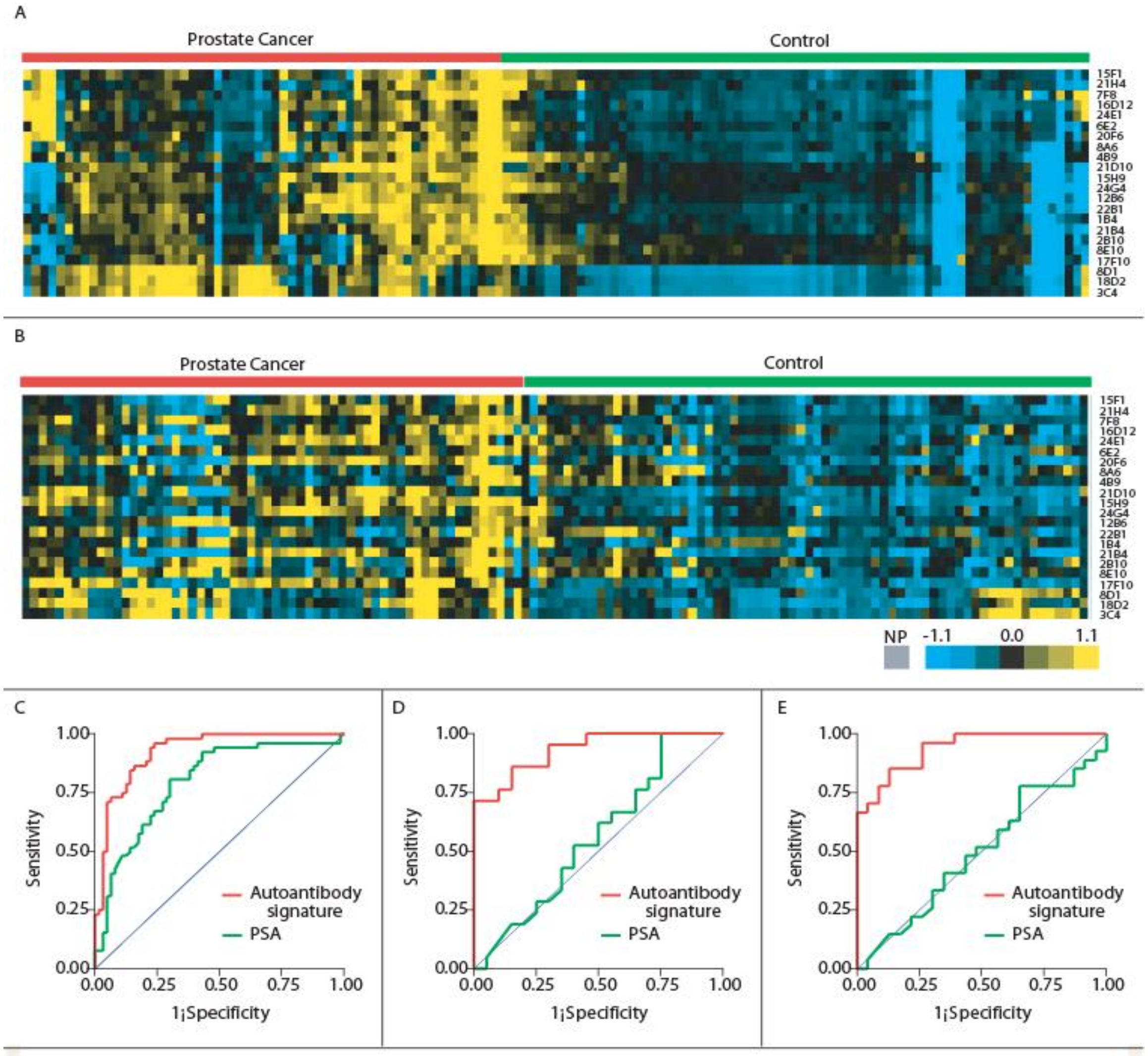


**Figure 5.3**. Supervised analyses and validation of autoantibody signatures in prostate cancer. **A**. Heatmap representation of the 22 phage-peptides analyzed for immunoreactivity across 129 training samples. **B**. Heatmap representation of the 22 phage-peptides for 128 independent validation set of sera from prostate cancer patients and controls. Individual peptide clones were represented in rows while serum samples were represented in columns. Intensities of yellow color represent positive immunoreactivity while intensities of black and blue represent no immunoreactivity. **C**, Performance of the 22 phage-peptide predictor as compared with PSA in the validation set. Receiver operating characteristic (ROC) curves are based on multiplex analysis of the 22 phage-peptide biomarkers and serum PSA (n=128; 60 prostate cancer patients and 68 control subjects). The red line represents the 22 phage-peptide predictor, and the green line represents the PSA test. **D**, Performance of the 22 phage-peptide predictor in patients with PSA levels between 4-10 ng/ml. The patients were from 128 validation samples with total number of 42 (22 cancers and 20 controls). See **C** for color label. **E**. Performance of the 22 phage-peptide predictor in patients with PSA levels between 2.5 and 10 ng/ml. Same as **D**, the samples were a subset of 128 validation group (n=51, 28 cancers and 23 controls). Color labels are same as **C**.

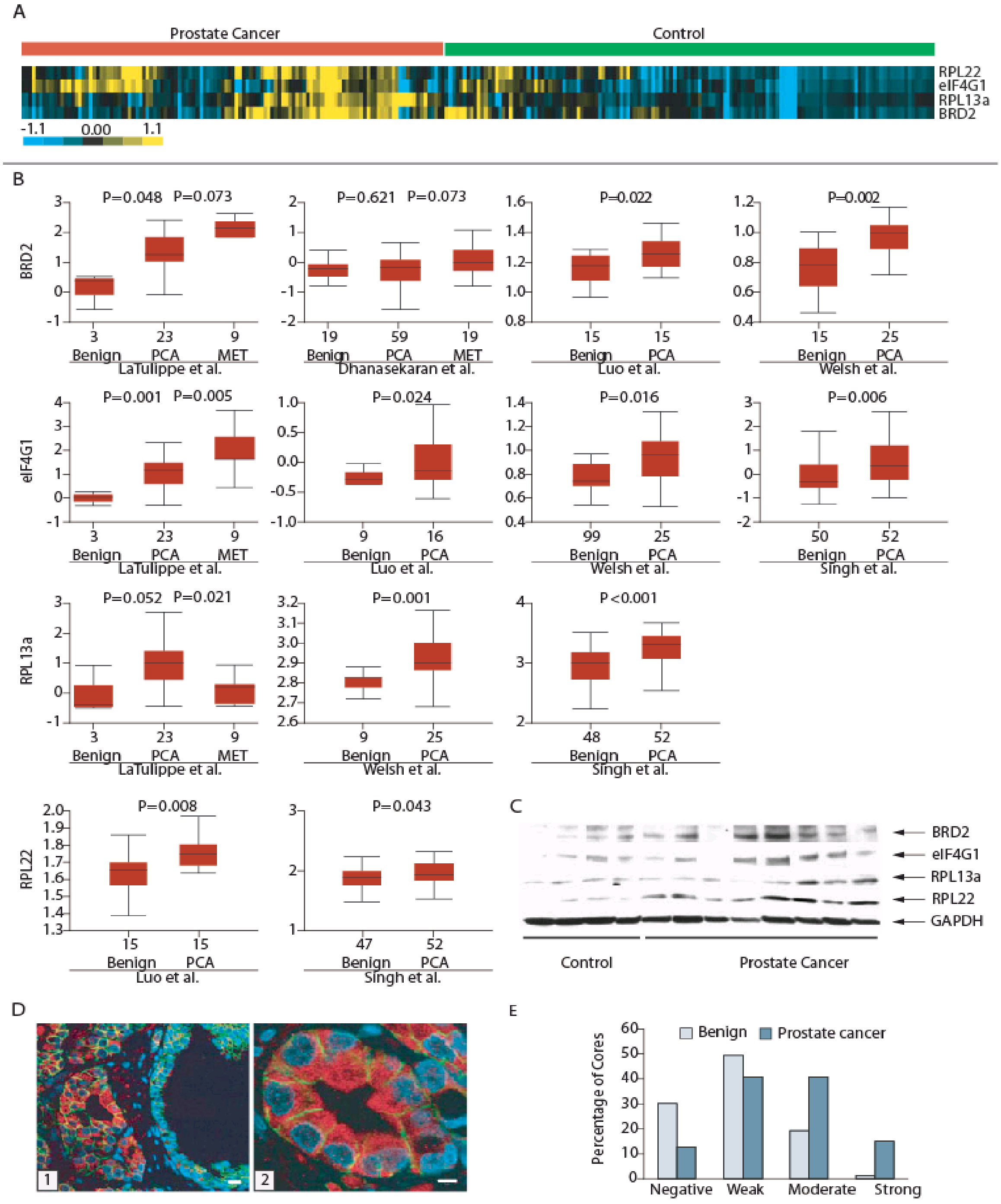


**Figure 5.4**. Gene expression meta-analysis of humoral immune response candidates. **A**, Heatmap representation of the immunoreactivity for four in-frame phage-peptide clones assessed across 257 serum samples (**Figure 5.2**). See **Figure 5.3B** for color scheme. **B**, Relative gene expression levels of in frame phage-peptide clones assessed using publicly available DNA microarray data housed in ONCOMINE (www.oncomine.org). First author of each DNA microarray study is provided. P values for each comparison made is provided (e.g., benign vs localized prostate cancer (PCA); PCA vs. metastatic prostate

cancer (MET)). **C**, Immunoblot validation of the overexpression of humoral response candidates at the protein level in prostate cancer. **D**. Expression of the humoral response candidate eIF4G1 in prostate cancer by immunofluorescence staining. **Panel 1** displays clinically localized prostate cancer (left) adjacent to a benign gland (right). **Panel 2** display magnifications of a single prostate cancer gland. Stains for eIF4G1 (red), E-cadherin (green) and nuclei (blue) were employed. Scale bar represents 5 µm. **E**. Histogram of staining intensity from immunohistochemistry. Open box represents benign tissue cores, while black box represent tumor cores. X-axis is the stain intensity, and y-axis is the percentage of tissue cores.

# REFERENCES

1. Thompson, IM, Pauler, DK, Goodman, PJ, et al. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* **350**, 2239-2246.

2. Brichory, FM, Misek, DE, Yim, AM, et al. (2001). An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer. *Proc Natl Acad Sci U S A* **98**, 9824-9829.

3. Chen, YT, Gure, AO, Tsang, S, et al. (1998). Identification of multiple cancer/testis antigens by allogeneic antibody screening of a melanoma cell line library. *Proc Natl Acad Sci U S A* **95**, 6919-6923.

4. Minenkova, O, Pucci, A, Pavoni, E, et al. (2003). Identification of tumor-associated antigens by screening phage-displayed human cDNA libraries with sera from tumor patients. *Int J Cancer* **106**, 534-544.

5. Sahin, U, Tureci, O, Schmitt, H, et al. (1995). Human neoplasms elicit multiple specific immune responses in the autologous host. *Proc Natl Acad Sci U S A* **92**, 11810-11813.

6. Stockert, E, Jager, E, Chen, YT, et al. (1998). A survey of the humoral immune response of cancer patients to a panel of human tumor antigens. *J Exp Med* **187**, 1349-1354.

7. Zhong, L, Peng, X, Hidalgo, GE, et al. (2003). Antibodies to HSP70 and HSP90 in serum in non-small cell lung cancer patients. *Cancer Detect Prev* **27**, 285-290.

8. Mintz, PJ, Kim, J, Do, KA, et al. (2003). Fingerprinting the circulating repertoire of antibodies from cancer patients. *Nat Biotechnol* **21**, 57-63.

9. Nilsson, BO, Carlsson, L, Larsson, A, and Ronquist, G (2001). Autoantibodies to prostasomes as new markers for prostate cancer. *Ups J Med Sci* **106**, 43-49.

10. Old, LJ, and Chen, YT (1998). New paths in human cancer serology. *J Exp Med* **187**, 1163-1167.

11. Soussi, T (2000). p53 Antibodies in the sera of patients with various types of cancer: a review. *Cancer Res* **60**, 1777-1788.

12. Sreekumar, A, Laxman, B, Rhodes, DR, et al. (2004). Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer. *J Natl Cancer Inst* **96**, 834-843.

13. Jiang, Z, Woda, BA, Rock, KL, et al. (2001). P504S: a new molecular marker for the detection of prostate carcinoma. *Am J Surg Pathol* **25**, 1397-1404.

14. Luo, J, Zha, S, Gage, WR, et al. (2002). Alpha-methylacyl-CoA racemase: a new molecular marker for prostate cancer. *Cancer Res* **62**, 2220-2226.

15. Rubin, MA, Zhou, M, Dhanasekaran, SM, et al. (2002). alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *Jama* **287**, 1662-1670.

16. Fernandez-Madrid, F, Tang, N, Alansari, H, et al. (2004). Autoantibodies to Annexin XI-A and Other Autoantigens in the Diagnosis of Breast Cancer. *Cancer Res* **64**, 5089-5096.

17. Li, L, Weinberg, CR, Darden, TA, and Pedersen, LG (2001). Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* **17**, 1131-1142.

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

19. LaTulippe, E, Satagopan, J, Smith, A, et al. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* **62**, 4499-4506.

20. Luo, J, Duggan, DJ, Chen, Y, et al. (2001). Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* **61**, 4683-4688.

21. Luo, JH, Yu, YP, Cieply, K, et al. (2002). Gene expression analysis of prostate cancers. *Mol Carcinog* **33**, 25-35.

22. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

23. Singh, D, Febbo, PG, Ross, K, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209.

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

25. Dennis, LK, and Resnick, MI (2000). Analysis of recent trends in prostate cancer incidence and mortality. *Prostate* **42**, 247-252.

26. Jemal, A, Tiwari, RC, Murray, T, et al. (2004). Cancer statistics, 2004. *CA Cancer J Clin* **54**, 8-29.

27. Schmid, HP, Prikler, L, and Semjonow, A (2003). Problems with prostate-specific antigen screening: a critical review. *Recent Results Cancer Res* **163**, 226-231; discussion 264-226.

28. Catalona, WJ, Smith, DS, Ratliff, TL, et al. (1991). Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med* **324**, 1156-1161.

29. Makinen, T, Tammela, TL, Stenman, UH, et al. (2004). Second round results of the Finnish population-based prostate cancer screening trial. *Clin Cancer Res* **10**, 2231-2236.

30. Cohen, L, Fouladi, RT, Babaian, RJ, et al. (2003). Cancer worry is associated with abnormal prostate-specific antigen levels in men participating in a community screening program. *Cancer Epidemiol Biomarkers Prev* **12**, 610-617.

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

# CHAPTER 6

## Multiplexed Urine Test for the Early Detection of Prostate Cancer

Prostate specific antigen (PSA) serum level is currently the standard of care for prostate cancer screening in the United States. The PSA test lacks specificity due to elevated levels in benign conditions such as benign prostatic hyperplasia (BPH) or prostatitis. Thus, additional biomarkers are needed to supplement or potentially replace the serum PSA test. Emerging evidence suggests that monitoring the non-coding RNA transcript *PCA3* in urine may be useful in detecting prostate cancer in patients with elevated PSA. Here we provide evidence that a multiplex panel of urine transcripts outperforms PCA3 transcript alone for the detection of prostate cancer. We measured the expression of 7 putative prostate cancer biomarkers, including *PCA3*, in sedimented urine using quantitative PCR on a cohort of 234 patients presenting for biopsy or radical prostatectomy. By univariate analysis, we found that increased *GOLPH2*, *SPINK1* and *PCA3* transcript expression, and *TMPRSS2:ERG* fusion status were significant predictors of prostate cancer. Multivariate regression analysis demonstrated that a multiplexed model including these biomarkers outperformed serum PSA or *PCA3* alone in detecting prostate cancer. The area under the receiver-operating characteristic curve was 0.758 for the multiplexed model versus 0.662 for *PCA3* alone, p = 0.003. The sensitivity and specificity for the multiplexed model were 65.9% and 76.0%, respectively, and the positive and negative predictive values were 79.8% and 60.8%, respectively. Taken together, these results provide the framework for the development of highly optimized, multiplex urine biomarker tests for the early detection of prostate cancer.

Prostate cancer is one of the leading causes of cancer-related death in American men. Prostate specific antigen (PSA) has been used extensively to screen for prostate cancer in the United States, based on early studies showing that PSA levels greater than 4 ng/ml have predictive value for detecting prostate cancer (1,2). While PSA testing has led to a dramatic increase in the detection of prostate cancer (3), PSA as a cancer biomarker has substantial drawbacks. For example, PSA is often elevated in benign conditions such as benign prostatic hyperlasia (BPH) and prostatitis, likely accounting for the poor specificity of the PSA test, which has been reported to be only 20% at a sensitivity of 80% (4). Further, a study investigating men in the Prostate Cancer Prevention Trial showed that even in patients with PSA levels lower than 4ng/ml, over 15% had biopsy-detectable prostate cancer (5). Taken together, this supports the value of identification and characterization of prostate cancer biomarkers that could supplement PSA.

Numerous genes have been identified as promising prostate cancer biomarkers, including genes specific for prostate cancer, such as *AMACR* (6) and *PCA3* (7), and markers based on recurrent fusions involving *TMPRSS2* and ETS family members (such as *TMPRSS2:ERG*) (8). As prostate cells can be detected in the urine of men with prostate cancer, urine based diagnostic tests have the advantage of being non-invasive. While urine-based testing for *PCA3* expression has already been documented in large screening programs (9), the feasibility of testing based on other markers has not been rigorously evaluated. Importantly, single marker tests, such as those based on *PCA3*, ignore the heterogeneity of cancer development, and may only capture a proportion of cancer cases. To overcome this limitation, multiplexing, or combining, biomarkers for cancer detection can improve testing characteristics (10,11). Thus, in this study, we sought to explore a multiplexed urine-based diagnostic test for prostate cancer. We reported a new outlier gene in prostate cancer which represents a subset of prostate cancer and then develop a multiplexed model for urinary dection of prostate cancer by combining this new outlier gene with known prostate cancer biomarkers and fusion genes.

Recently, our lab developed a bioinformatics approach termed COPA (cancer outlier profile analysis) to nominate candidate oncogenes from transcriptomic data based on high expression in a subset of cases (8). When applied to the Oncomine compendium of tumor profiling studies **(www.oncomine.org)** (12), COPA successfully identified the ETS family members *ERG* and *ETV1* as high-ranking outliers in multiple prostate cancer profiling studies, leading to the discovery of recurrent gene fusions involving the 5' untranslated region of the androgen regulated gene *TMPRSS2* with *ERG, ETV1*, or *ETV4* in prostate cancer cases that over-expressed the respective ETS family member1 (13).

ETS gene fusions occur in 40-80% of prostate specific antigen (PSA)-screened prostate cancers, leaving 20-60% of prostate cancers in which the key genetic aberration cannot be ascribed to ETS gene fusions. In order to detect new candidate oncogenes in ETS negative prostate cancers, our lab performed a COPA-based meta-analysis on 7 prostate cancer profiling studies (14-20). Eleven genes were nominated as outliers in at least 4 of the 7 datasets (**Figure 6.1A**). Consistent with the previous application of COPA filtered by causal cancer genes (8), both *ERG* and *ETV1* were high ranking meta-outliers; *ERG* ranked as the 1st meta-outlier (7 studies) and *ETV1* as the 5th meta-outlier (4 studies).

Interestingly, this analysis also identified *SPINK1* (serine peptidase inhibitor, Kazal type 1), the 2nd ranked meta-outlier, as showing over-expression in prostate cancer compared to benign prostate tissue and mutually exclusive over-expression with *ERG* and *ETV1* across multiple studies. The profile of *SPINK1* expression and scatter plots with *ERG* and *ETV1* for two studies (15,20) where *SPINK1* was identified as a top-100 outlier are shown in **Figure 6.1B**. We further evaluated the expression of SPINK1 protein in prostate cancers. By immunohistochemical (IHC) analysis on tissue microarrays (TMAs), we evaluated SPINK1 expression in two independent cohorts, (University of Michigan (UM) and Swedish Watchful Waiting (SWW)) representing a total of 392 cases of clinically localized prostate cancers that had been previously evaluated for TMRPSS2-ERG fusion status by fluorescence in situ hybridization (FISH) (21,22). As shown in **Figure 6.2A-B**, in the UM cohort, 10 and 36 of 75 cases were positive for SPINK1 expression (13.3%) and TMRPSS2-ERG fusions (48%), respectively, with all SPINK1 positive cases being TMRPSS2-ERG negative (one sided Fisher's exact test, p = 0.0008). In the SWW cohort, 23 and 57 of 312 cases were positive for SPINK1 expression (7.4%) and TMRPSS2-ERG fusions (18.3%), respectively, again with all SPINK1 positive cases being TMRPSS2-ERG negative (one sided Fisher's exact test, p = 0.008).

We also examined if *SPINK1* outlier status was associated with biochemical recurrence after surgical resection. In the Glinsky et al. (15) gene expression dataset, which contained tumors from 79 patients (with 37 recurrences), 10 of which showed outlier mRNA transcript expression of *SPINK1* (>0.5 normalized expression units). These patients had a significantly higher risk of recurrence than patients without outlier *SPINK1* expression (hazard ratio: 2.65, 95% CI: 1.16-6.07, log rank p = 0.02) by Kaplan-Meier analysis (**Figure 6.2C**). A similar analysis was also conducted on the UM cohort (75 cases, 28 recurrences) evaluated for SPINK1 status by IHC. By Kaplan-Meier analysis, SPINK1 positive staining was significantly associated with biochemical recurrence (hazard ratio: 2.49, 95% CI: 1.01-6.18, p = 0.04, **Figure 6.2D**). As a final validation, we performed IHC for SPINK1 status on an independent cohort of 817 evaluable prostate cancers (200 recurrences) from the Memorial Sloan Kettering Cancer Center (MSKCC). In this MSKCC cohort, we defined SPINK1 positive cases in the MSKCC cohort as those with at least one core showing greater than 80% of cells showing positive SPINK1 immunoreactivity, resulting in 75 SPINK positive cases (9%), consistent with the other analyses. By Kaplan-Meier analysis, SPINK1 positive cases in the MSKCC cohort showed significantly shorter time to biochemical recurrence (hazard ratio: 2.32, 95% CI: 1.59-3.39, P = 6.96E-06, **Figure 6.2E**). While the above survival analyses were based on univariate analysis, we also performed multivariate Cox proportional-hazards regression analyses on the above three datasets and confirmed that SPINK1 status predicted recurrence independently of other clinical parameters such as Gleason score, lymph node status, surgical margin status and pre-operative PSA.

In the above section, we have demonstrated by analyzing 971 cancers from three cohorts that SPINK1 outlier status identifies an aggressive subset of ETS-negative prostate cancers. In the next section, we sought to determine whether such outlier genes as *ERG* and *SPINK1*, combining with known prostate cancer biomarkers can improve the diagnosis of prostate cancer.

We set out to assess expressions of seven putative prostate cancer biomarkers using qRT-PCR technique. These biomarkers included those generally over-expressed in prostate cancer, such as *PCA3*, *AMACR* and *GOLPH2* (6,7), as well as those over-expressed in subsets of prostate cancers, such as *ERG* and *TMPRSS2:ERG*, *TFF3*, and *SPINK1* (8,23,24). To develop a multiplexed qPCR based test for prostate cancer, we profiled a cohort of 138 patients with prostate cancer (86 positive needle biopsy and 52 radical prostatectomy patients) and 96 patients with negative needle biopsies from the University of Michigan.

All genes were first tested by univariate analysis, with *GOLPH2* (P = 0.0002), *SPINK1* (P = 0.0002), *PCA3* (P = 0.001) and *TMPRSS2:ERG* fusion (P = 0.034) showing significant association for discriminating patients with prostate cancer from patients with negative needle biopsies (**Figure 6.3** and **Table 6.1**). Both *AMACR*, which has previously been shown to be a sensitive and specific biomarker for prostate cancer in tissues (6) and *TFF3*, which shows high expression in a subset of prostate cancers(23,24), were not statistically significant predictors of prostate cancer using urine samples (P = 0.450 and

0.189, respectively). The lack of specificity of *AMACR* and *TFF3* in urine may be due to expression of these transcripts in urothelial or kidney derived cellular material which can also be shed in the urine. While *TMPRSS2:ERG* fusion was significantly associated with the presence of prostate cancer (**Figure 6.3** and **Table 6.1**), *ERG* overexpression was not associated with cancer presence on univariate analysis (P = *0.166*), suggesting that cells from other tissues may be contributing *ERG* transcripts to the urine. Additionally, serum PSA levels prior to biopsy or prostatectomy were also not associated with cancer presence in this cohort (P = *0.376*). When tested as individual variables for the ability to detect prostate cancer based on the receiver-operating-characteristic curves (ROC), *GOLPH2* (area under the curve (AUC) = 0.664, P = 2.01E-5), *PCA3* (AUC = 0.661, P = 2.84E-5), and *SPINK1* (AUC = 0.642, P = 0.0002) outperformed serum PSA (AUC=0.508, p=0.837) (**Figure 6.3**). Thus, in this analysis we have identified a number of novel biomarkers for the non-invasive detection of prostate cancer using patient urine instead of biopsy samples. Of the seven markers utilized in this study, only *PCA3* was previously reported as urinary diagnostic biomarker (9).

To determine if a multiplex model could improve on the performance of these single biomarkers, the analyzed prostate cancer biomarkers were next tested in a multivariate regression analysis using Akaike Information Criterion (AIC)-based backward selection (25) to drop insignificant terms from the model. This analysis resulted in a final model that included *SPINK1* (P = 7.41E-5), *PCA3* (P = 0.003), *GOLPH2* (P = 0.004) and *TMPRSS2:ERG* (P = 0.006) (**Table 6.1**). To evaluate the performance of this model for diagnosing prostate cancer, we then performed ROC analysis based on the predicted probabilities derived from the final model. For our cohort, we compared the ROC curves from the multiplexed model and *PCA3* alone, as urine based detection of *PCA3* has previously been evaluated in similar cohorts as a single biomarker using alternative detection technologies (9,26-29). For example, Van gils *et al*. demonstrated that in a cohort of 534 men presenting for prostate biopsy with serum PSA between 3-15 ng/mL, urinary *PCA3* detection expression had an area under the ROC curve (AUC) of 0.66 compared to 0.57 for serum PSA (9). As shown in **Figure 6.4A**, in our cohort, the AUC for the multiplexed model (0.758, P = 1.91E-11) was significantly improved (P=0.003 (30)) compared to the AUC for *PCA3* alone (0.662, P = 2.58E-5). At the point on the multiplex model ROC with the maximum sum of sensitivity and specificity (65.9% and 76.0%, respectively), the positive predictive value was 79.8% and the negative predictive value was 60.8% (**Figure 6.4A**). As we and previous studies used different methodologies to detect *PCA3* transcripts in patient urine, directly comparing AUCs is inappropriate; however, we demonstrate that *PCA3* shows improved AUC compared to serum PSA, consistent with previous reports (9,26-29). Importantly, we demonstrate that a multiplex model including *PCA3* significantly improves the predictive ability of *PCA3* alone, suggesting the ability to improve *PCA3* and other single-gene based diagnostic tests. The rationale for the multiplex approach is consistent with tests offered to breast cancer patients to identify patients at high risk for disease recurrence (10,31).

As all samples were used to select the best subset of variables for regression analysis, there is a potential to over-optimize the reported AUC. Thus, we used a leave-one-out-cross validation (LOOCV) strategy to generate an unbiased AUC. As shown in **Figure 6.4B**, the AUC for the LOOCV multiplex model (0.736) is again significantly better (P = 0.006) than that for LOOCV PCA3 alone (0.645). At the point on the LOOCV multiplex model ROC with the maximum sum of sensitivity and specificity (62.3% and 75.0%, respectively), the positive predictive value was 78.2% and the negative predictive value was 58.1% (**Figure 6.4B**).

Lastly, we tested the ability of these genetic markers to predict clinical risk groups based on patient parameters. Clinical risk groups were determined by clinical patient data that direct the decision to pursue biopsy, to determine treatment, or to stratify patients for surveillance regimens (see **Methods**). We observed only limited associations between these prostate cancer biomarkers and clinical risk groups, with *GOLPH2*, *SPINK1* and *TMPRSS2:ERG* status showing marginal correlates with clinical stage, and major gleason. As the biomarkers in this study were chosen based on their ability to differentiate benign prostate tissue and prostate cancer, it is not surprising that they did not show strong association with risk stratification measures, such as Kattan nomogram prediction of recurrence or organ confined status. Thus, the ideal marker panel would be designed to enable risk stratification based on pre-biopsy urine samples while incorporating markers designed to predict cancer presence. Similar to the previously described PCR based test for breast cancer recurrence risk, a prostate cancer risk test could drive high risk patients to therapies more suited for their disease course (10).

In summary, we demonstrate that a multiplexed qRT-PCR based assay on sedimented urine collected from patients presenting for prostate biopsy or prostatectomy exhibits superior performance relative to serum PSA or *PCA3* alone. Of note, the multiplex urine test that we present here, which is a combination of *PCA3*, *SPINK1*, *GOLPH2* and *TMPRSS2:ERG* gene fusion status achieves a specificity and positive predictive value of >75%, making it a potentially useful test to complement serum PSA, which has poor specificity in detecting prostate cancer. This study establishes a basic framework for the development of a urine multiplex test for the early detection and prognosis of prostate cancer. Future studies will be directed at improving the performance of this first generation urine multiplex test by evaluating additional markers for inclusion as well as allow for improved risk stratification and patient counseling prior to treatment decision making.

### Methods

Cancer Outlier Profile Analysis (COPA). COPA analysis was performed on 7 prostate cancer gene expression data sets (14-20) in Oncomine **(www.oncomine.org)** as described (8). Briefly, for each data set, gene expression values are median-centered, setting each gene's median expression value to zero. Second, the median absolute deviation (MAD) is calculated and scaled to 1 by dividing each gene expression value by its MAD. Of note, median and MAD are used for transformation as opposed to mean and standard deviation so that outlier expression values do not unduly influence the distribution estimates, and are thus preserved post-normalization. In each dataset, genes are rank-ordered by their COPA scores at three percentile cutoffs: 75th, 90th and 95th. For each dataset, we defined outlier genes as those that ranked in the top 100 COPA scores at any one of the percentile the cutoffs. To identify meta-outlier genes, we ranked genes by the number of studies where the gene was identified as a top 100 outlier. Genes identified as outliers in the same number of studies were further ranked by their average outlier rank across those studies.

Immunohistochemsitry (IHC) and fluorescence in situ hybridization (FISH). IHC for the University of Michigan (UM) and Swedish Watchful Waiting (SWW) cohorts was performed using a mouse monoclonal antibody against SPINK1 (H00006690-M01, Abnova, Taipei City, Taiwan) on tissue microarrays (TMA) containing cores from 75 (UM) and 312 (SWW) evaluable cases of localized prostate cancer. Staining in greater than 1% of cancerous epithelial cells was deemed positive. Previously, we have evaluated cases on these tissue microarrays for *TMRPSS2-ERG* fusion status by FISH using break apart *ERG* assays as previously described (21,22). A one-sided Fisher's exact test was used to evaluate the relationship between SPINK1 and fusion status, as these studies were performed with the prior hypothesis that there was an inverse correlation between SPINK1 expression and fusion status.

MSKCC Immunohistochemistry. IHC for the MSKCC cohort was performed using an in house mouse monoclonal antibody against SPINK1 on tissue microarrays containing triplicate cores from 817 evaluable cases of localized prostate cancer. The percentage of positive tumor cells in each core was estimated and assigned values of 0%, 5%, or multiples of 10%. The intensity of the expression was assigned a value of 0, 1, 2, or 3.

Triplicate cores from each specimen were scored separately and the presence of tumorous tissue in at least two interpretable cores was required to include a case for analysis. We considered cases as SPINK1 positive if any of the three cores showed >80% of cancerous cells showing positive SPINK1 immunoreactivity (intensity 1-3).

Urine Collection, RNA isolation, amplification and quantitative PCR. This study was approved by the Institutional Review Board (IRB) of the University of Michigan Medical School and samples were obtained from 276 patients with informed consent following a digital rectal exam before either needle biopsy (n=216) or radical prostatectomy (n=60) at the University of Michigan Health System (UMHS). Urine was voided into urine collection cups containing DNA/RNA preservative (Sierra Diagnostics LLC, Sonora, CA). Isolation of RNA from urine and whole transcriptome amplification (WTA) were as described in (32). Quantitative PCR (qPCR) was used to detect seven prostate cancer biomarkers (*AMACR*, *ERG*, *GOLPH2*, *PCA3*, *SPINK1*, *TFF3*, and *TMPRSS2:ERG* fusions) and the control transcripts *PSA* and *GAPDH* from WTA amplified cDNA essentially as described(32,33). The primer sequences for *ERG* (exon5_6)(8), *GAPDH*, (34) *AMACR* (35) and *PSA*(36) were previously described and for other biomarkers were as follows:

*GOLPH2*-f: CTGGTGGCCTGCATCATCGTCTTG,

*GOLPH2*-r: GCTGCTCCCGCTGCTTCTCCA,

*PCA3*-f: CATGGTGGGAAGGACCTGATGATAC,

*PCA3*-r: GATGTGTGGCCTCAGATGGTAAAGTC,

*SPINK1*-f: CAAAAATCTGGGCCTTGCTGAGAAC,

*SPINK1*-r: AGGCCTCGCGGTGACCTGAT,

*TFF3*-f: AACCGGGGCTGCTGCTTTGACTC,

*TFF3*-r: TCCTGCAGGGGCTTGAAACACCA.

*TMPRSS2:ERG* fusions were detected using Taqman primers/probe, with the following sequences:

*TM-ERGa3*-f: CTGGAGCGCGGCAGGAA,

*TM-ERGa3* -r: CCGTAGGCACACTCAAACAACGA,

*TM-ERGa3*_MGB-probe: 5’-MGB-TTATCAGTTGTGAGTGAGGAC-3’.

Threshold levels were set during the exponential phase of the qPCR reaction using Sequence Detection Software version 1.2.2 (Applied Biosystems, Foster City, CA), with the same baseline and threshold set for each plate, to generate threshold cycle ($C_t$) values for all genes for each sample.

Outcome Analysis. For Kaplan-Meier analysis of the Glinsky et al.(15) and UM datasets, biochemical recurrence was defined as a 0.2 ng/ml increase in PSA or recurrence of disease after prostatectomy, such as development of metastatic cancer, if biochemical recurrence information was not available. For the MSKCC cohort, only biochemical recurrence, defined as PSA > 0.2 ng/ml after surgical resection with a second confirmatory PSA-measurement > 0.2 ng/ml, was considered, as all patients with a clinical failure had previously had a biochemical recurrence. For outcome analysis from the Glinsky *et al*. dataset, samples positive for outlier expression of *SPINK1* were defined as those with greater than 0.5 normalized expression units (as shown in **Figure 1B**). For the IHC analysis of the UM and MSKCC cohorts, positive cases were defined as described above. Kaplan-Meier analysis and multivariate Cox proportional-hazards regression were then used to examine the association of *SPINK1* with biochemical PSA recurrence.

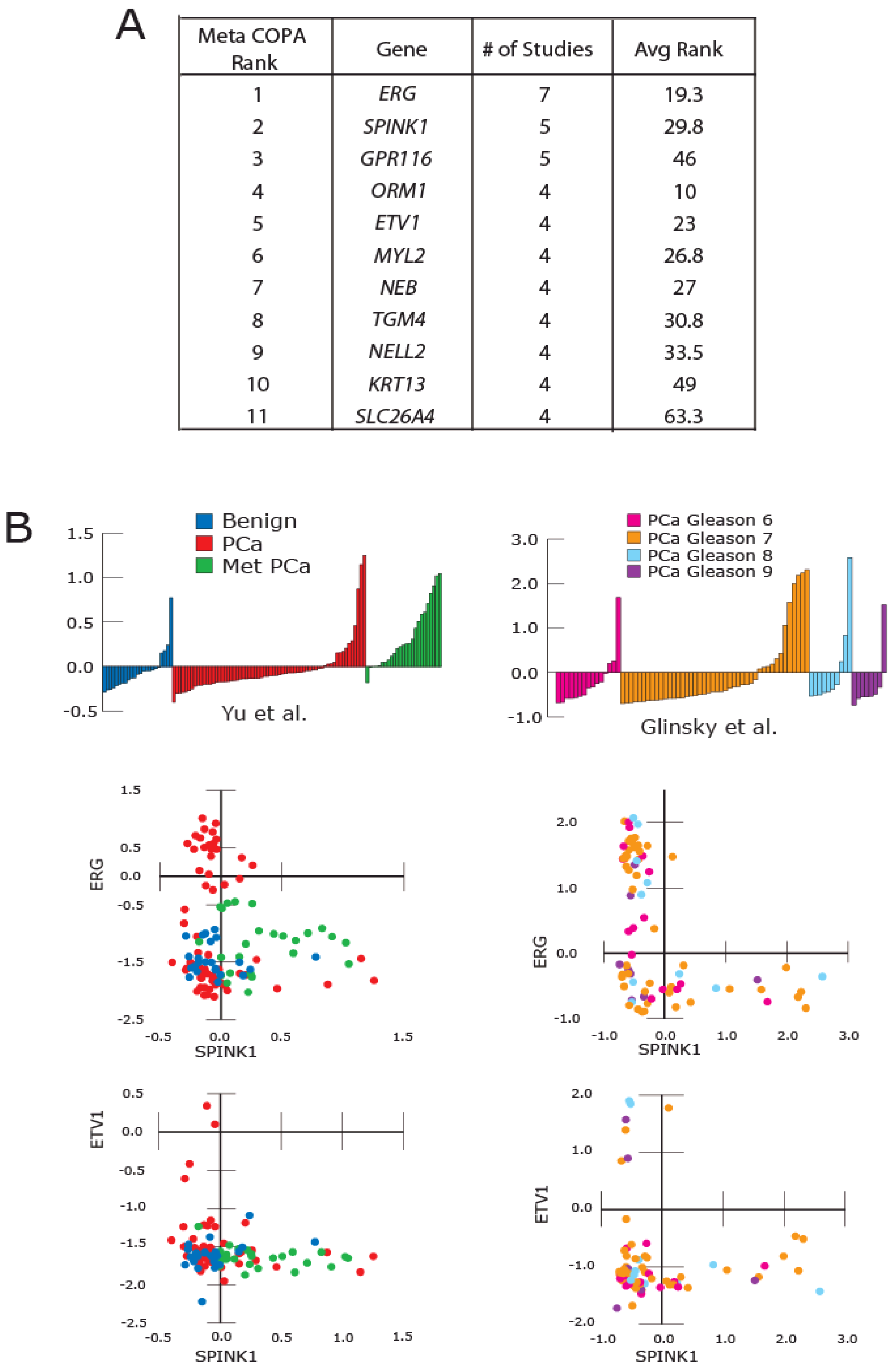
Urinary Data Analysis. qPCR was performed on WTA cDNA from urine collected from 111 biopsy-negative patients and 165 patients with prostate cancer (105 biopsy positive patients and 60 prostatectomy patients). Samples that had PSA $C_t$ values greater than 27 were excluded to ensure sufficiency of the amount of prostate cells in the samples, leading to 105 biopsy-negative and 152 samples from patients with prostate cancer in the analysis. For qPCR analysis, we used raw $–\Delta C_t$ (to stabilize the variance of testing variables) as opposed to testing markers against control ($2^{–\Delta Ct}$). *TMPRSS2:ERG* was dichotomized as a binary variable to reflect the fusion positive or negative status observed in tissue samples(8,37), with positive samples defined as those with $C_t$ values less than 37. As *PCA3* has been reported to be a prostate tissue-specific marker(7), it was normalized against urine *PSA* ($C_{tPSA}$-$C_{tPCA3}$). All other testing variables were adjusted against their mean urine *PSA* and *GAPDH* values (($C_{tPSA}$+$C_{tGAPDH}$) / 2-$C_{tVariable}$). We excluded 23 samples showing outlier values, as at least one testing variable in those samples showed an adjusted value below 3 standard deviations from its sample mean, suggesting qPCR failure. This resulted in a final data set of samples from 138 patients with prostate cancer (86 positive needle biopsy and 52 radical prostatectomy patients) and 96 biopsy-negative patients.

Statistical Analysis. Univariate and multivariate logistic regressions were used to examine associations between prostate cancer diagnostic status and testing variables. For multivariate logistic regression, the Akaike Information Criterion (AIC)-based backward selection was used to drop insignificant terms (25). All testing markers were included in the initial regression model which was further refined by the AIC-based backward selection. After the final model was determined, the predicted probability for each sample was used as input to generate the receiver operating characteristic (ROC) curve and the area under the curve (AUC) was calculated. As all samples were used for regression model generation, the estimated AUC may be over-optimized. To correct this bias, we further performed a leave-one-out cross validation. Briefly, one sample was omitted while the regression model was trained on the remaining samples to select optimal markers and estimate their coefficients. The prediction probability for the left-out sample is then calculated based on the model prediction. This procedure was repeated until every sample was left out once and the derived prediction probability values were then used for ROC analysis. Similarly, PCA3 alone was fitted in a logistic regression model to generate an AUC. The difference of AUCs was examined as described previously (30). All analyses were performed in R (**http://www.r-project.org)** and ROC curves were plotted in SPSS 11.5 (SPSS Inc., Chicago, IL, USA).

**Table 6.1**. Univariate and multivariate logistic regression analyses used to identify urine biomarkers for the detection of prostate cancer. For the multivariate analysis, Akaike Information Criterion (AIC)-based backward selection was used to drop insignificant terms.

| **Univariate Logistic Regression Analysis** | | |
|---|---|---|
| **Variable** | **Coefficient** | **P-value** |
| *GOLPH2* | 0.4444 | 0.0002 |
| *SPINK1* | 0.25 | 0.0002 |
| *PCA3* | 0.187 | 0.001 |
| *TMPRSS2:ERG* | 0.609 | 0.034 |
| *ERG* | 0.043 | 0.166 |
| *TFF3* | 0.11 | 0.189 |
| PSA (serum) | 0.0151 | 0.376 |
| *AMACR* | 0.049 | 0.45 |

| **Multivariate Logistic Regression Analysis** | | |
|---|---|---|
| **Variable** | **Coefficient** | **P-value** |
| *SPINK1* | 0.308 | 7.41E-05 |
| *PCA3* | 0.191 | 0.003 |
| *GOLPH2* | 0.372 | 0.004 |
| *TMPRSS2:ERG* | 0.924 | 0.006 |



| Meta COPA Rank | Gene | # of Studies | Avg Rank |
|---|---|---|---|
| 1 | *ERG* | 7 | 19.3 |
| 2 | *SPINK1* | 5 | 29.8 |
| 3 | *GPR116* | 5 | 46 |
| 4 | *ORM1* | 4 | 10 |
| 5 | *ETV1* | 4 | 23 |
| 6 | *MYL2* | 4 | 26.8 |
| 7 | *NEB* | 4 | 27 |
| 8 | *TGM4* | 4 | 30.8 |
| 9 | *NELL2* | 4 | 33.5 |
| 10 | *KRT13* | 4 | 49 |
| 11 | *SLC26A4* | 4 | 63.3 |

**Figure 6.1**. Meta COPA identifies *SPINK1* as a mutually exclusive outlier with *ERG* and *ETV1* in prostate cancer. Meta-COPA analysis of 7 prostate cancer gene expression profiling datasets in Oncomine. **A**. Genes were ranked by the number of studies in which they scored in the top 100 outliers (ranked by COPA) at any of the three pre-defined percentile cutoffs (75th, 90th, 95th). Genes were further ranked by their average COPA rank (Avg. Rank) in studies where they ranked in the top 100. **B**. The expression of *SPINK1* and scatter plots of *ERG* vs. *SPINK1* and *ETV1* vs. *SPINK1* expression are shown from two studies where *SPINK1* ranked as a top 100 COPA outlier. The expression of *SPINK1*, in normalized expression units, for all profiled samples including benign prostate tissue (blue), clinically localized prostate cancer (PCa, red) and metastatic PCa (Met PCa, green), as well as Gleason pattern 6, 7, 8 or 9 prostate cancer (magenta, orange, light blue and purple, respectively) are shown in the top panels. Scatter plots are shown for *ERG* vs. *SPINK1* (middle panels) and *ETV1* vs. *SPINK1* (lower panels) for all samples in both studies.

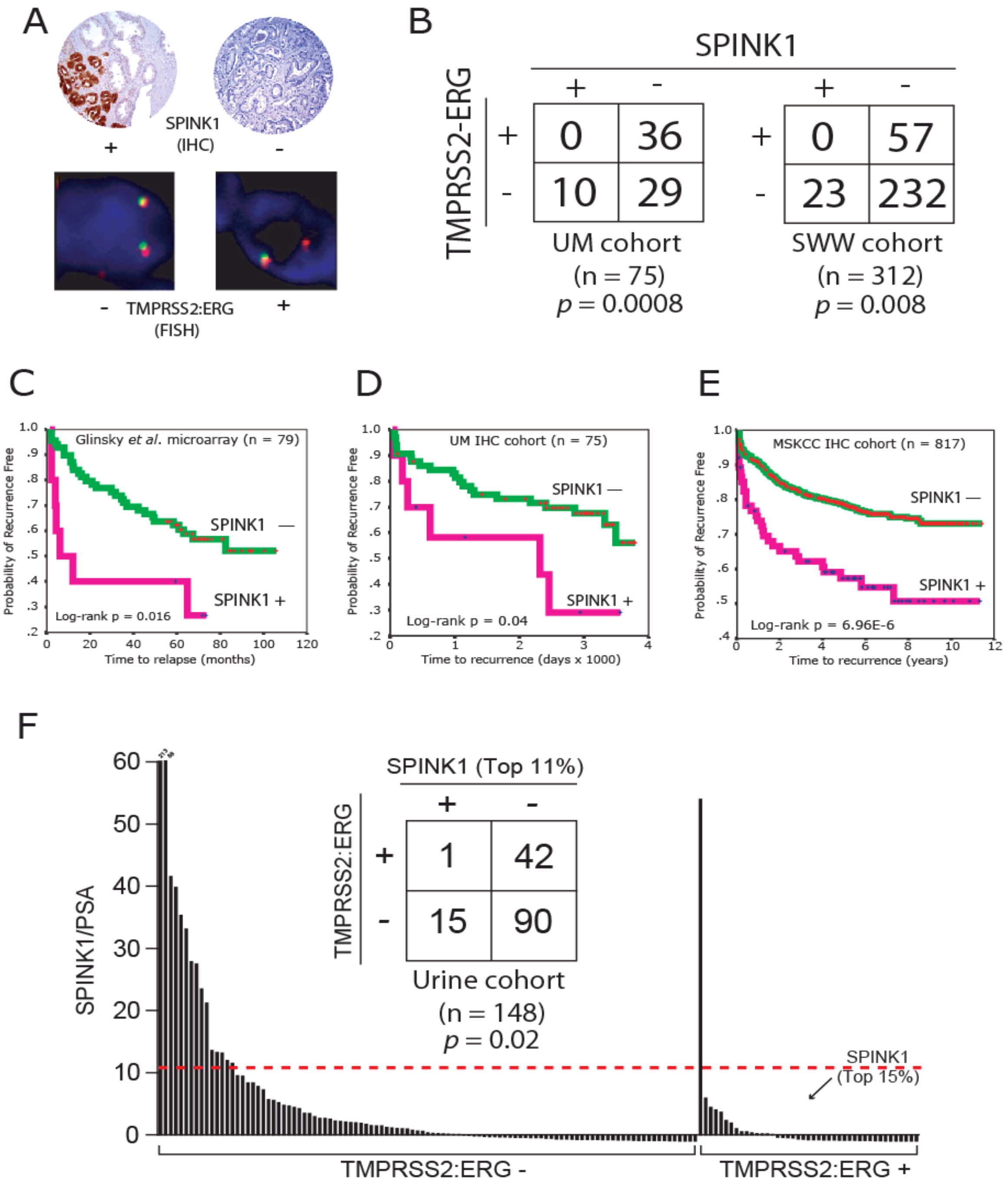
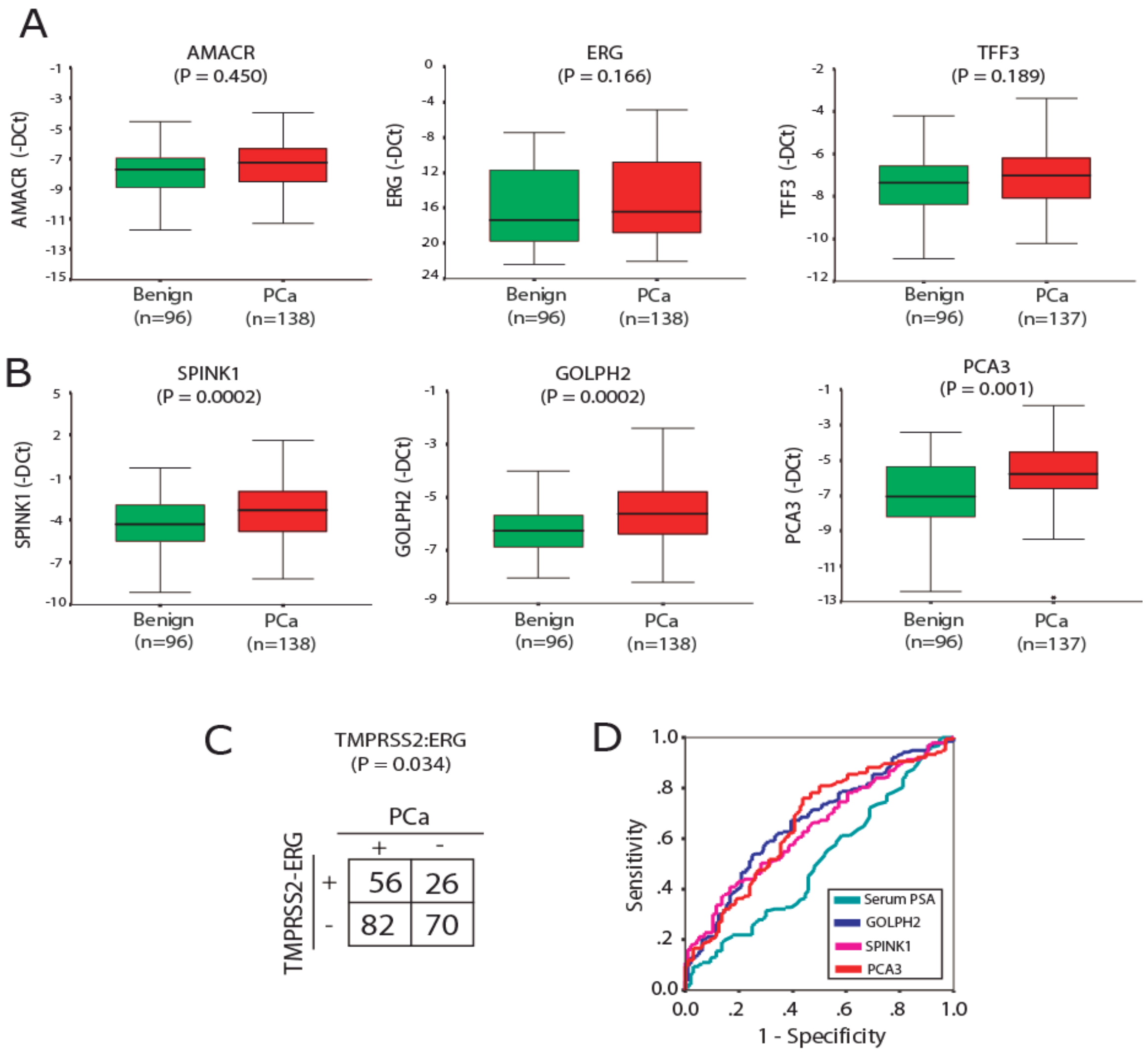


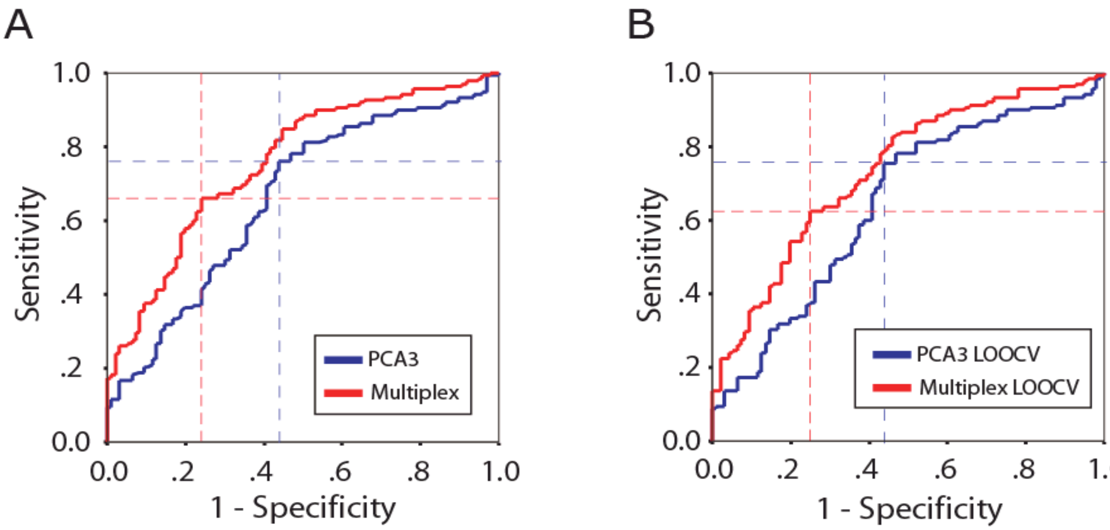
**Figure 6.2**. SPINK1 over-expression identifies an aggressive subset of ETS negative prostate cancers and can be detected non-invasively. **A-B**. SPINK1 expression was evaluated in two cohorts (University of Michigan (UM) and Swedish Watchful Waiting (SWW)) using immunohistochemsitry (IHC) on tissue microarrays that have previously been evaluated for TMRPSS2-ERG status by fluorescence in situ hybridization (FISH). **A.** Representative SPINK1 positive and negative cores are shown, along with cells from the same cores negative and positive for TMRPSS2-ERG rearrangement by FISH. A

TMRPSS2-ERG rearrangement through intrachromosomal deletion is indicated by loss of one 5' (green) ERG signal. **B**. Contingency tables for SPINK1 expression and TMRPSS2-ERG status and p-values for Fisher's exact tests for both cohorts are indicated. **C**-**E**. Relationship between SPINK1 outlier expression and biochemical recurrence after surgical resection. Kaplan-Meier analyses of outlier *SPINK1* expression from the (**C**) Glinsky et al. DNA microarray dataset and SPINK1 IHC from the (**D**) UM and (**E**) Memorial Sloan Kettering Cancer Center (MSKCC) cohorts and biochemical recurrence after surgical resection are shown. **F**. Non-invasive detection of *SPINK1* outlier-expression in men with *TMRPSS2:ERG* negative prostate cancer. Total RNA was isolated from the urine of 148 men with prostate cancer and assessed for *TMRPSS2:ERG* and *SPINK1* expression by quantitative PCR. Samples above the dashed red line show *SPINK1* outlier expression (See Methods). Contingency table for *SPINK1* outlier expression and *TMPRSS2:ERG* status and the Fisher's exact test p-value is shown.



|  | PCa |  |
|---|---|---|
| TMPRSS2-ERG | + | - |
| + | 56 | 26 |
| - | 82 | 70 |

**Figure 6.3**. Characterization of candidate urine-based biomarkers of prostate cancer. **A-C**, Quantitative PCR (qPCR) was performed on whole transcriptome amplified (WTA) cDNA from urine obtained from patients presenting for needle biopsy or prostatectomy. Biomarker expression in patients with negative needle biopsies (green) or patients with prostate cancer (positive needle biopsy or prostatectomy, red) are shown. Normalization was performed using –ΔCt, with *PCA3* normalized to urine *PSA* expression as performed previously(26). *AMACR*, *ERG*, *GOLPH2*, *SPINK1* and *TFF3* were normalized to the average of urine sediment *PSA* and *GAPDH* expression. *TMPRSS2:ERG* gene fusion expression was dichotomized as positive or negative. The –ΔCt values of genes that were not significant predictors of prostate cancer by univariate analysis (see **Table 6.1**) are shown in **A**, and the expression of those that were significant predictors are shown in **B** & **C**. *P* values from the univariate analysis for the detection of prostate cancer are indicated. **D**, Receiver operator characteristic (ROC) curves for individual variables for the diagnosis of prostate cancer. The area under the curves (AUC) for *GOLPH2*, *PCA3*, *SPINK1* and serum PSA are 0.664, 0.661, 0.642 and 0.508, respectively.



| Variable | Sens | Spec | PPV | NPV |
|---|---|---|---|---|
| Multiplex | 65.9% | 76.0% | 79.8% | 60.8% |
| PCA3 | 75.4% | 56.3% | 71.2% | 61.4% |

| Variable | Sens | Spec | PPV | NPV |
|---|---|---|---|---|
| Multiplex LOOCV | 62.3% | 75.0% | 78.2% | 58.1% |
| PCA3 LOOCV | 75.4% | 56.3% | 71.2% | 61.4% |

**Figure 6.4**. A multiplexed set of urine biomarkers outperforms PCA3 alone in the detection of prostate cancer. **A.** Multivariate regression analysis resulted in a multiplexed model including *SPINK1*, *PCA3*, *GOLPH2* and *TMPRSS2:ERG* as significant predictors of prostate cancer (see **Table 6.1**). ROC analysis was then performed based on the predicted probabilities derived from the final model. The multiplexed model (red) showed significantly greater AUC than *PCA3* (blue) alone (0.758 vs 0.662, P = 0.003) for the detection of prostate cancer. The point on the ROC curve with the maximum sum of sensitivity (Sens) and specificity (Spec) is indicated by the dashed line, and the positive and negative predictive values (PPV and NPV, respectively) are given. **B**. As in **A**, except a leave-one-out cross validation (LOOCV) strategy was used to generate unbiased AUCs. The AUC for the LOOCV multiplex model is significantly better than LOOCV of *PCA3* alone (0.736 vs. 0.645, P = 0.006).

# REFERENCES

1. Brawer, MK, Chetner, MP, Beatie, J, et al. (1992). Screening for prostatic carcinoma with prostate specific antigen. *J Urol* **147**, 841-845.

2. Catalona, WJ, Smith, DS, Ratliff, TL, et al. (1991). Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med* **324**, 1156-1161.

3. Brawley, OW, Knopf, K, and Merrill, R (1998). The epidemiology of prostate cancer part I: descriptive epidemiology. *Semin Urol Oncol* **16**, 187-192.

4. Catalona, WJ, Hudson, MA, Scardino, PT, et al. (1994). Selection of optimal prostate specific antigen cutoffs for early detection of prostate cancer: receiver operating characteristic curves. *J Urol* **152**, 2037-2042.

5. Thompson, IM, Pauler, DK, Goodman, PJ, et al. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter. *N Engl J Med* **350**, 2239-2246.

6. Rubin, MA, Zhou, M, Dhanasekaran, SM, et al. (2002). alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *Jama* **287**, 1662-1670.

7. de Kok, JB, Verhaegh, GW, Roelofs, RW, et al. (2002). DD3(PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res* **62**, 2695-2698.

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

9. van Gils, MP, Hessels, D, van Hooij, O, et al. (2007). The time-resolved fluorescence-based PCA3 test on urinary sediments after digital rectal examination; a Dutch multicenter validation of the diagnostic performance. *Clin Cancer Res* **13**, 939-943.

10. Paik, S, Shak, S, Tang, G, et al. (2004). A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* **351**, 2817-2826.

11. Schmidt, U, Fuessel, S, Koch, R, et al. (2006). Quantitative multi-gene expression profiling of primary prostate cancer. *Prostate* **66**, 1521-1534.

12. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

13. Tomlins, SA, Mehra, R, Rhodes, DR, et al. (2006). TMPRSS2:ETV4 gene fusions define a third molecular subtype of prostate cancer. *Cancer Res* **66**, 3396-3400.

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

15. Glinsky, GV, Glinskii, AB, Stephenson, AJ, Hoffman, RM, and Gerald, WL (2004). Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* **113**, 913-923.

16. Lapointe, J, Li, C, Higgins, JP, et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816.

17. LaTulippe, E, Satagopan, J, Smith, A, et al. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* **62**, 4499-4506.

18. Vanaja, DK, Cheville, JC, Iturria, SJ, and Young, CY (2003). Transcriptional silencing of zinc finger protein 185 identified by expression profiling is associated with prostate cancer progression. *Cancer Res* **63**, 3877-3882.

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

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

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

22. Mehra, R, Tomlins, SA, Shen, R, et al. (2007). Comprehensive assessment of TMPRSS2 and ETS family gene aberrations in clinically localized prostate cancer. *Mod Pathol* **20**, 538-544.

23. Faith, DA, Isaacs, WB, Morgan, JD, et al. (2004). Trefoil factor 3 overexpression in prostatic carcinoma: prognostic importance using tissue microarrays. *Prostate* **61**, 215-227.

24. Garraway, IP, Seligson, D, Said, J, Horvath, S, and Reiter, RE (2004). Trefoil factor 3 is overexpressed in human prostate cancer. *Prostate* **61**, 209-214.

25. Venables, WNaR, B. D. (2002). Modern Applied Statistics with S, 4th Edition: New York: Springer).

26. Hessels, D, Klein Gunnewiek, JM, van Oort, I, et al. (2003). DD3(PCA3)-based molecular urine analysis for the diagnosis of prostate cancer. *Eur Urol* **44**, 8-15; discussion 15-16.

27. Fradet, Y, Saad, F, Aprikian, A, et al. (2004). uPM3, a new molecular urine test for the detection of prostate cancer. *Urology* **64**, 311-315; discussion 315-316.

28. Groskopf, J, Aubin, SM, Deras, IL, et al. (2006). APTIMA PCA3 molecular urine test: development of a method to aid in the diagnosis of prostate cancer. *Clin Chem* **52**, 1089-1095.

29. Marks, LS, Fradet, Y, Deras, IL, et al. (2007). PCA3 molecular urine assay for prostate cancer in men undergoing repeat biopsy. *Urology* **69**, 532-535.

30. DeLong, ER, DeLong, DM, and Clarke-Pearson, DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845.

31. van de Vijver, MJ, He, YD, van't Veer, LJ, et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009.

32. Laxman, B, Tomlins, SA, Mehra, R, et al. (2006). Noninvasive detection of TMPRSS2:ERG fusion transcripts in the urine of men with prostate cancer. *Neoplasia* **8**, 885-888.

33. Tomlins, SA, Mehra, R, Rhodes, DR, et al. (2006). Whole transcriptome amplification for gene expression profiling and development of molecular archives. *Neoplasia* **8**, 153-162.

34. Vandesompele, J, De Preter, K, Pattyn, F, et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* **3**, RESEARCH0034.

35. Kumar-Sinha, C, Shah, RB, Laxman, B, et al. (2004). Elevated alpha-methylacyl-CoA racemase enzymatic activity in prostate cancer. *Am J Pathol* **164**, 787-793.

36. Specht, K, Richter, T, Muller, U, et al. (2001). Quantitative gene expression analysis in microdissected archival formalin-fixed and paraffin-embedded tumor tissue. *Am J Pathol* **158**, 419-429.

37. Perner, S, Mosquera, JM, Demichelis, F, et al. (2007). TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion. *Am J Surg Pathol* **31**, 882-888.

# PART 4: AN INTEGRATIVE APPRAOCH TO MODEL PROSTATE CANCER PROGRESSION

## CHAPTER 7

### Integrative Genomic and Proteomic Analysis of Prostate Cancer Reveals Signatures of Metastatic Progression

Molecular profiling of cancer at the transcript level has become routine. Large scale analysis of proteomic alterations during cancer progression has been a more daunting task. Here, we employed high-throughput immunoblotting in order to interrogate tissue extracts derived from prostate cancer. We identified 64 proteins that were altered in prostate cancer relative to benign prostate and 156 additional proteins that were altered in metastatic disease. An integrative analysis of this compendium of proteomic alterations and transcriptomic data was performed revealing only 48-64% concordance between protein and transcript levels. Importantly, differential proteomic alterations between metastatic and clinically localized prostate cancer that mapped concordantly to gene transcripts served as predictors of clinical outcome in prostate cancer as well as other solid tumors.

Prostate cancer is a highly prevalent disease in older men of the Western world (1,2). Unlike other cancers, more men die with prostate cancer than from the disease (3,4). Deciphering the molecular networks that distinguish progressive disease from non-progressive disease will shed light into the biology of aggressive prostate cancer as well as lead to the identification of biomarkers that will aid in the selection of patients that should be treated (5). To begin to understand prostate cancer progression with a systems perspective, we need to characterize and integrate the molecular components involved (6-9). A number of groups have employed gene expression microarrays to profile prostate cancer tissues (10-18) as well as other tumors (19-22) at the transcriptome level but much less work has been done at the protein level. Proteins, as opposed to nucleic acids, represent the functional effectors of cancer progression and thus serve as therapeutic targets as well as markers of disease.

In the present study, we utilized a high-throughput immunoblot approach to characterize proteomic alterations in human prostate cancer progression focusing on the transition from clinically localized prostate cancer to metastatic disease. Using an integrative approach we were able analyze proteomic profiles with mRNA transcript data from several laboratories. Our analyses also indicated the proteins that were qualitatively concordant with gene expression could be used to define a multiplex gene predictor of clinical outcome.

In order to derive a first approximation of the prostate cancer proteome, we employed high-throughput immunoblot analysis. This method, while not feasible for use on many individual samples, allowed us to screen pooled tissue extracts for qualitative levels of hundreds of proteins (and post-translational modifications) using commercially available antibodies. The basic approach is illustrated in **Figure 7.1A**. Extracts from five tissue specimens of benign prostate, clinically localized prostate cancer and metastatic prostate cancer from distinct patients were pooled. Each of the 3 pools of tissue extracts were run on preparative SDS-PAGE gels, transferred to PVDF, and incubated with different antibodies using a miniblot apparatus. **Figure 7.1B** displays representative data using the high-throughput immunoblot approach. Known proteomic alterations in prostate cancer progression such as EZH2 (23) and AMACR (24-26) are highlighted in red while novel associations such as GSK-3beta and IRAK1 are highlighted in green. To further increase the number of proteins analyzed, we used an analogous high-throughput immunoblot methodology provided by commercial services (see Methods). Thus, in total we assessed 1484 antibodies against 1354 distinct proteins or post-translational modifications. Of these antibodies, 521 detected a band of the expected molecular weight in at least one of the pooled extracts. Antibodies that did not detect the correct molecular weight protein product may represent lack of antibody sensitivity (or poor quality antibody) or absence of protein expression in prostate tissues.

To validate the proteomic alterations identified by this screen in individual tissue extracts (as opposed to pooled extracts), we analyzed 86 proteins and 2 post-translational modifications by conventional immunoblot analysis using 4-5 tissue extracts per class. As with most gene expression studies done in prostate, our proteomic screen employed grossly dissected tumor specimens. Thus, the proteomic alterations that we detected could be due to differences in the stromal-eptihelial ratio of the tissues in addition to actual alterations in the epithelial cells. In order to evaluate the proteomic alterations in situ, we employed high-density tissue microarrays (27). As only a subset of the identified proteins have antibodies that are compatible with immunohistochemical analysis, a single tissue microarray containing 216 specimens from 51 cases was stained using twenty of these IHC-compatible antibodies. Representative tissue microarray elements are shown in **Figure 7.2A**. Each tissue microarray element was evaluated by a pathologist and scored for staining (scale of 1-4) as per cell type considered (e.g., epithelial, stromal etc…). Using an *in situ* technique such as evaluation by immunohistochemistry allowed us to distinguish stromal versus epithelial expressed proteins. In general, proteins that demonstrated a decrease in expression in the metastatic tumors most often were stromally expressed proteins. As the amount of stroma per unit area decreases with tumor progression, metastatic samples demonstrated a parallel significant decrease in protein expression of paxillin and ABP-280, among others. In order to visualize and cluster the tissue microarray data (28), the qualitative evaluations were normalized (See Methods). Similar to gene expression analyses (21,29), unsupervised hierarchical clustering of the data revealed that the in situ protein levels could be used to accurately classify prostate samples as benign, clinically localized prostate cancer, or metastatic disease (**Figure 7.2B**).

This high-throughput immunoblotting of prostate extracts led to the identification of a several known and previously unknown proteomic alterations in prostate cancer. For example, previous studies have shown that the anti-apoptosis protein, XIAP (30), the racemase AMACR (24-26) and the Polycomb Group protein EZH2 (23) are dysregulated in prostate cancer progression. Novel associations (increases or decreases in protein expression) with prostate cancer progression identified by this screen include the E2 ubiquitin ligase UBc9, the cytosolic phosphoprotein stathmin, the death receptor DR3, and the Aurora A kinase (STK15), among others.

Having amassed this compendium of proteomic alterations in prostate cancer progression, we next examined the general concordance with the prostate cancer transcriptome. To this end we developed an integrative model to incorporate qualitative proteomic alterations as assessed by high-throughput immunoblotting (but applicable to other proteomic technologies), with transcriptomic data derived from 8 prostate cancer gene expression studies (**Figure 7.3**). As both the genomic and proteomic approach involve analysis of grossly dissected tissues, this facilitates molecular comparisons to be made. The high-throughput immunoblot analysis of benign prostate, clinically localized prostate cancer and metastatic disease yielded 521 proteins of the expected molecular weight. Immunoreactive bands in each of the three tissue extracts were assessed and comparisons were made between benign tissue and clinically localized prostate cancer (**Figure 7.3A**) and between clinically localized prostate cancer and metastatic disease (**Figure 7.3B**). Visually qualified proteins that were over-expressed were coded red, under-expressed proteins were coded blue, and unchanged proteins were coded white. Based on this analysis, 64 proteins were dysregulated in clinically localized prostate cancer relative to benign prostate tissue, while 156 proteins were dysregulated between metastatic disease relative to clinically localized prostate cancer. As might be expected, most of the proteins analyzed were unchanged in the context of prostate cancer progression (i.e., 87.7% (457/521) of the proteins were unchanged between clinically localized prostate cancer and benign and 70.1% (365/521) of the proteins were unchanged between clinically localized and metastatic disease).

The set of quantifiable proteins (n=521) was then mapped to the NCBI Locus link and UniGene databases to identify each corresponding gene. Data for mRNA was extracted for these genes using 8 publicly available prostate cancer gene expression data sets (See methods). Over 90% of the genes were represented in at least one microarray study, allowing for integrative analysis to be performed. All eight of the prostate profiling studies made a comparison between clinically localized prostate cancer and benign tissue, while only four of these studies made a comparison between clinically localized disease and metastatic disease. Genes which could only be found in one-fourth of studies or less were excluded, leading to 481 genes involved in the former comparison and 492 involved in the latter comparison. Since we assessed over- and under-expressed genes separately, a one-sided t test was conducted per each gene per each profiling study (See Methods). As with the proteomic approach, comparisons between benign and clinically localized prostate cancer (**Figure 7.3A**) and localized disease and metastatic disease (**Figure 7.3B**) were made. If an mRNA transcript was significantly over-expressed in a particular study it was coded red, under-expressed transcripts were coded blue, and white was used for unchanged transcripts.

**Figure 7.3** presents an integrative analysis of proteomic data with gene expression meta-data in prostate cancer progression. An mRNA transcript alteration was considered "concordant" with a proteomic alteration if a majority of the microarray profiling studies (at least 50%) showed the same qualitative differential (increased, decreased, or unchanged) as the high-throughput immunoblot approach. According to these criteria, 289 (60.1%) out of 481 mRNA transcripts were concordant with protein levels in clinically localized prostate cancer relative to benign prostate tissue. Similarly, 291 (59.1%) out of 492 mRNA transcripts were concordant with protein levels in metastatic prostate cancer relative to clinically localized disease. Out of the 156 proteomic alterations identified between metastatic and localized prostate cancer, 50 were concordant with mRNA transcript and 90 were discordant with mRNA transcript while the remaining alterations did not have mRNA measurements to map to (**Figure 7.3B-C**). Thus, similar to studies done in yeast (31,32), bacteria (33), and cell lines (34), there was only weak concordance between protein and mRNA levels in prostate cancer progression.

To further explore the poor concordance we observed between protein and meta-data from transcriptomic analyses, we profiled the pooled samples as well as the individual samples that comprised the pools on Affymetrix HG-U133 plus 2 microarrays. The same integrative analysis was carried out to examine the concordant relationship between the protein alterations observed in the pooled tissues by immunoblotting and transcript alterations observed in the corresponding pooled and individual tissues. The individual samples were included in order to calculate statistical significance for transcript alterations. Similar or even lower concordance was observed between protein and transcript (61.0% concordance in clinically localized prostate cancer relative to benign prostate tissue, and 48.2% for metastatic prostate cancer relative to clinically localized disease, **Figure 7.4A**).

We also investigated the protein and mRNA concordance in individual samples. We focused on the 86 proteins identified as outliers in the larger high-throughput screen. The immunoblot intensities were semi-quantitated and correlation coefficients were calculated for each protein (see Methods). We found that a total 55 out of 86 proteins were observed to a have a positive correlation with mRNA, which led to 64.0% concordance between proteins and transcripts (**Figure 7.4B**). On sub classification, we observed a concordance of 54.7% and 66.3% in case of localized prostate cancer relative to benign prostate tissues and the metastatic disease relative to localized prostate cancer respectively.

This proteomic screen identified proteins that are altered from benign prostate to clinically localized prostate cancer and a distinct set of alterations between clinically localized disease to metastatic disease. Since we are interested in the transition from clinically localized to metastatic disease we next focused on this comparison. As the metastatic tissues analyzed in this study are androgen-independent (35), and by contrast the clinically localized tumors are generally androgen-dependent, we evaluated whether there was an enrichment of androgen-regulated proteomic alterations discovered by our screening. Androgen regulated genes (ARGs) are essential for the normal development of the prostate as well as the pathogenesis of prostate cancer (36-38). Pertinent to this analysis, Velasco *et al*. developed a meta-analysis of ARGs which represents a cross-comparison of 4 gene expression (39-42) and 2 SAGE datasets (43,44). ARGs were then defined as a union of these 6 datasets, all of which represented functional induction of mRNA transcript by androgen *in vitro*. Interestingly, 27 out of the 150 protein alterations (exclusive of post-translational modifications) we identified as being differential between metastatic and clinically localized disease, were designated as androgen-regulated by the Velasco *et al* (42) ARG compendium. To demonstrate that this finding is statistically significant, we selected random sets of 150 genes from the Yu *et al*. (18) or the Glinsky *et al*. (45) prostate cancer profiling studies and found that the chance of selecting 27 ARGs was minimal (*Ps* < 0.001 for both of the Yu *et al.* and Glinsky *et al*. data). Thus, androgen-regulated proteins are significantly enriched in the differential comparison between androgen-dependent and independent prostate cancer.

While examining concordant proteomic alterations, interestingly, we found that EZH2, a Polycomb group protein that we and others have previously characterized as being over-expressed in aggressive prostate and breast cancer (23,46) was one of the 50 proteins identified as being concordantly over-expressed in metastatic tissues at the mRNA and protein level (**Figure 7.3B-C**). As EZH2 was a member of this 50 gene concordant signature, we hypothesized that proteomic alterations that distinguish metastatic prostate cancer from clinically localized disease may serve as a multiplex signature of prostate cancer progression when applied to clinically localized disease (i.e., “more aggressive” genes would be expressed in progressive prostate cancer). While antibodies have yet to be developed to test all of these proteomic alterations in situ by immunohistochemistry, we postulated that mRNA transcript levels could be used instead due to their concordance with protein levels in this signature. To test this hypothesis we selected prostate cancer gene expression datasets that monitored over 85% of the genes in the concordant genomic/proteomic signature, included biochemical recurrence information (time to PSA recurrence), as well as reported on at least 50 clinically localized specimens. According to Dobbin et al. (47), the number of samples required for developing prognostic markers was approximately 51 or above for a general human gene expression dataset with the variance of a gene over samples as 0.5, type I error as 0.001, and type II error as 0.05. Thus we chose n=50 as our minimal sample size requirement in this analysis.

The prostate cancer gene expression datasets that fulfilled these criteria were carried out by Yu *et al.* (18) and Glinsky *et al.*(45), both of which represent Affymetrix oligonucleotide datasets and each of which measured at least 44 out of the 50 genes in the concordant signature. Although the Singh *et al.* and LaPointe et *al.* studies reported over 50 samples in their studies, the number of samples for which we have available follow-up information was less than 30 (29 and 20 samples for the LaPointe and Singh dataset, respectively). In addition, the average follow-up time for the samples in LaPointe study was only 10.7 months. Thus we excluded both datasets in the analysis. We then chose to build our prediction models with the Yu *et al.* data set and test the performance on the Glinsky *et al.* data set. Utilizing an approach described earlier (48), unsupervised hierarchical clustering in the space of this 44-gene concordant signature resulted in two main clusters of individuals in the Yu *et al.* study (**Figure 7.5A**). Kaplan-Meier (KM) survival analysis of the clusters indicated that the two groups of individuals are significantly different based on time to recurrence status (*P* = 0.035, **Figure 7.5A**). Notably, when we use the 90 discordant genes (mRNA transcripts that are not qualitatively concordant with protein levels) we found that these signatures did not generate a clinical outcome distinction (P= 0.238). Moreover, by permutation test, we also observed that random sets of 44 genes did not generate such prognostic distinctions (See Methods), indicating that our concordant signature was not likely due to chance. To assess the validity of this concordant signature, we utilized the Glinsky *et al.* study as an independent test set (**Figure 7.5B**). Each of the samples in the Glinsky dataset were classified as high- or low-risk based on a *k*-nearest neighbor (*k*-NN) model developed using the Yu *et al*. study as a training set (*k*=3). Based on the class predictions derived from the concordant signature, KM survival analysis revealed a significant difference in survival based on the risk stratification (P = 0.001, **Figure 7.5B**). As expected, this was not the case with the discordant signature when applied to the Glinsky *et al*. sample set (P= 0.556). A similar result was observed when a predictive model built on the Glinsky et al. data was applied to the Yu *et al*. data ($P$ < 0.001 and $P$ = 0.02 for the Glinsky et al. and Yu *et al*. data, respectively). We then carried out multivariate Cox proportional-hazards regression analysis of the risk of recurrence on the Glinsky *et al*. validation set. **Table 7.1** shows that the concordant signature predicted recurrence independently of the other clinical parameters such as surgical margin status, Gleason sum, and pre-operative PSA. With an overall hazard ratio of 3.66 (95% CI: 1.36-7.02, $P$<0.001), it was by far the strongest predictor of prostate cancer recurrence in the model.

Next, we sought to refine the concordant signature of prostate cancer progression by reducing the number of genes required. By using the Yu *et al*. study as a training set, the 44 concordant genes were ranked by a univariate cox model. The same clustering procedure was employed to identify two clusters based on the top number of genes ranging from a minimum of 5 to a maximum of 44. Based on this iterative analysis, we identified 9 genes that demarcated two main clusters that differed most significantly by KM survival analysis (**Figure 7.5A**, Methods). The Glinsky *et al*. study was again used as an independent validation set confirming that the 9-gene concordant signature identified two groups of individuals which differed significantly based on recurrence (**Figure 7.5B**). Together, this integrative analysis suggests that mRNA transcripts that correlate with protein levels in metastatic prostate cancer can be used as gene predictors of progression in clinically localized disease.

Next, we sought to explore the generality of the concordant progression signature in other solid tumors. We identified four tumor profiling datasets from the Oncomine compendium (49) that fulfilled the same criteria that we used in the prostate cancer analyses (see above). In 95 primary breast adenocarcinomas (50), tumors bearing the 50-gene concordant progression signature were more likely to progress to metastasis than those lacking this signature ($P$ = 0.0025). We observed a similar result in 80 primary breast infiltrating ductal carcinomas (51) ($P$ = 0.002, **Figure 7.5C**). Moreover, this result was also observed in a series of 84 primary lung adenocarcinomas (52) ( $P$ = 0.03; **Fig 5C**) and 56 gliomas (53) ($P$ =0.01; **Figure 7.5C**). Furthermore, we used two common gene expression prediction models (diagonal linear discriminant analysis and *k*-nearest neighbor analysis) and conducted direct comparisons of the performances of the progression signature and the “study-specific” signature in each individual study where such a specific signature was available (see **Table 7.2**). The result indicated that the progression signature was able to retrieve similar or even superior prediction performance in most of the studies, especially when employing the *k*-nearest neighbor prediction model. This is remarkable as this signature was derived exclusively from prostate samples but had utility not only in prostate cancer datasets but also in breast cancer, lung cancer, and glioma datasets. Again, this suggests that there is likely biology inherent in the integrated predictor. Of note, we found that the smaller 9-gene model was only effective in discriminating prognostic classes in the Freije et al glioma study (P=0.016) but not in the other solid tumor data sets. This suggests that the 9-gene model may be relatively specific for prostate cancer while the 50-gene model has more universal applicability. Taken together, our observations suggest that the progression proteomic/genomic signature identified by the integrative analysis of metastatic prostate cancer may have utility in the prognostication of clinically localized solid tumors in general. Biologically, this suggests that aggressive tumors of different tissue origin begin to share the molecular machinery of a de-differentiated state.

While these proteomic alterations have potential to serve as a multiplex biomarker of cancer aggressiveness, they may also shed light into the biology of neoplastic progression. As proteins, rather than RNA transcripts, are the primary effectors of the cell, they play the central and most distal role in the functional pathways to cancer. Interestingly, EZH2, which we previously have shown to have a role in prostate cancer progression (23), is a member of this concordant genomic/proteomic signature, suggesting that other members of this signature may have utility as biomarkers as well as could have a role in the biology of progression. For example, this screen identified

Aurora-A kinase (STK15) as being overexpessed in metastatic prostate cancer as well as being a member of the 50-gene concordant signature. This serine-threonine kinase has been shown to be amplified in a number of human cancers (54,55), play a key role in G2/M cell cycle progression (56), and inhibit p53 (57), among other functions. Another candidate cancer regulatory molecule in the 50-gene concordant signature was KRIP1 (KAP-1), which is known to repress transcription via binding the methyltransferase SETDB1 (58).

In this study, we initially used a pooling strategy to perform high-throughput immunoblot analysis. While it would be more ideal to involve replicate protein measurements across multiple prostate tissues and then make comparisons to mRNA, the difficulty in monitoring thousands of antibodies on many individual samples and the cost of running multiple samples across thousands of antibodies required us to adopt the pooling approach. Further, analyses of concordance with mRNA expression on individual samples that comprised the pool confirmed the general feasibility of this strategy. We also noticed that there were recognized problems with annotations for microarrays. A recent study (59) reported that up to 50% Affymetrix probes do not have a matching-sequence in the Reference Sequence database (Refseq), questioning the reliability of such probes. As this study represents an initial foray in the area of integrative analyses, we used basic gene identifier-based matching for cross-platform annotations. Another potential limitation in the present study is that some immunoblots exhibit reactivity at multiple sizes potentially representing multiple protein isoforms. Thus, measuring the protein intensity for one 'expected' band may not be adequate for determining a correlation with transcripts. However, most of the reported changes here are the result of alterations in the reported or predicted molecular weight isoform. In future studies, we will investigate the various isoforms and proteolytically cleaved products.

Taken together we provide a general framework for the integrative analysis of proteomic and transcriptomic data from human tumors (**Figure 7.6**). Proteomic profiling of prostate cancer progression identified over one hundred altered proteins in the transition from clinically localized to metastatic disease (a significant fraction of which were androgen regulated). While this approach was useful to integrate high-throughput immunoblot data, the general paradigm can also be applied to mass spectrometry or protein microarray based technologies as they mature in the future. Differential proteins were then mapped to mRNA transcript levels to assess mRNA/protein concordance levels in a human disease state. Importantly, gene expression alterations that matched protein alterations qualitatively could be used as predictors of prostate cancer progression in clinically confined disease. Together, this would suggest that clinically aggressive prostate cancer bears a “signature” set of genes/proteins that is characteristic of metastatic disease. The observation that the concordant proteomic/genomic signature can be applied to other solid tumors suggests commonalities in the undifferentiated state of advanced tumors.

**Methods**

High-throughput Immunoblot Analysis. Tissues utilized were from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program, which are both part of University of Michigan Prostate Cancer Specialized Program of Research Excellence (S.P.O.R.E.) Tissue Core. Institutional Review Board approval was obtained to procure and analyze the tissues used in this study. To develop the tissue extract pools the following frozen tissue blocks were identified: 5 each of benign prostate tissues, clinically localized prostate cancer, and hormone-refractory metastatic tissues (35). Based on examination of the frozen sections of each tissue block, specimens were grossly dissected maintaining at least 90% of the tissue of interest. Total proteins were extracted from each tissue by homogenizing samples in boiling lysis buffer. One hundred micrograms of protein from each tissue extract pool was boiled in sample buffer and subjected to 4-15% preparative SDS-PAGE and transferred to PVDF and probed with different antibodies. To supplement the number of proteins analyzed, the same extracts were analyzed using two commercial service providers, BD biosciences and Kinexus. Validation immunoblots for selected proteins in different functional classes were carried out using 4-15% linear gradient SDS-PAGE for protein separation. The signal intensities were semi-quantitated using Scion Image software.

Microarray Analysis. Total RNA from the individual and pooled samples were analyzed on Affymetrix U133 2.0 Plus arrays by the University of Michigan Comprehensive Cancer Center Affymetrix Core. The amount and integrity of RNA was analyzed by spectrophotometry and the Agilent Bioanalyzer (Agilent Technologies). Biotin-labled cRNA synthesis, hybridization, washing, staining and scanning were done following the manufacturer’s protocols (Affymetrix). All RNA samples and arrays met standard quality control metrics.

Tissue Microarray Analysis (TMA). A prostate cancer progression TMA composed of benign prostate tissue, clinically localized prostate cancer, and hormone refractory metastatic prostate cancer was developed. These cases came from well fixed radical prostatectomy specimens as described previously (24). A total of 216 tissue samples were collected from 51 patients. Protein expression was determined using a validated scoring method (10,23,24) where staining was evaluated for intensity and the percentage of cells staining positive. Benign epithelial glands and prostate cancer cells were scored for staining intensity on a 4 tiered system ranging from negative to strong expression. Hierarchical clustering on samples and proteins was carried out after data normalization. Measurements for duplicated samples in the same patient were averaged and each measurement was divided by the global mean of the entire dataset and then base 2 log-transformed.

Integrative Molecular Analysis. To map the antibodies and their respective protein targets, we retrieved the official gene names from the NCBI Locuslink for our antibody/protein lists. To complement protein levels, transcriptome data was assembled from 8 publicly available prostate cancer gene expression datasets (10-14,16-18) and each probe was mapped to Unigene Build #173. Expression values from multiple clones or probe sets mapping to the same Unigene Cluster ID were averaged. Each gene in each study was normalized across samples so that the mean equaled zero and the standard deviation equaled to 1. Missing data was imputed by the *k-nearest neighbors (k=5) imputation* approach (60).

The eight prostate cancer profiling studies were included in the analysis of clinically localized prostate cancer relative to benign prostate tissue, while only 4 studies were included in the analysis of metastatic prostate cancer vs. localized prostate cancer due to the availability of metastatic samples in those studies. Genes that were only found in one-fourth of studies or less were excluded, leading to 483 genes involved in the former analysis and 494 involved in the latter analysis. A one-sided permutation t-test was conducted per gene per study using the multtest package in R 2.0. A gene was considered differentially expressed if its p-value was less than 0.05 without adjustment for multiple testing. An mRNA transcript alteration was considered "concordant" with a proteomic alteration if a majority of the microarray profiling studies (at least 50%) showed the same qualitative differential (increased, decreased, or unchanged) as the high-throughput immunoblot approach. The gene/proteins were then assigned to concordant and discordant groups based on this criterion.

Integrative Genomic and Proteomic Analysis of Individual Prostate Cancer Samples. We carried out profiling of mRNA expression analysis in 13 of the 14 individual samples used for the individual protein measurements (one was excluded due to an insufficient amount of tissue). We examined the concordance between proteins and transcripts for individual samples, focusing on the 86 proteins identified as outliers in the larger high-throughput screen. The immunoblot intensities were semi-quantitated using Scion Image software and the Spearman's rank correlation was calculated for each protein. An mRNA transcript alteration was considered “concordant” with a proteomic alteration if a positive correlation was found.

Clinical Outcomes Analysis. Six different cancer profiling studies (18,45,50-53) were used for evaluation of prognostic value of these concordant genes. Average linkage hierarchical clustering using an uncentered correlation similarity metric was used to identify two main clusters of clinically localized prostate cancer samples based on the 44 concordant mRNA transcripts that were qualitatively concordant with protein expression in the Yu *et al*. (18) study (only 44 out of 50 of the concordant signature were assessed on these arrays). Kaplan-Meier survival analysis of cluster-defined subgroups was then conducted and the log-rank test was used to calculate the statistical significance of difference between the two subgroups (SPSS 11.5). High-/low- risk labels were then assigned to each group. A permutation test was performed to evaluate the significance of this “progression” concordant signature. We selected 1000 random sets of 44 genes from the Yu *et al*. data set and then used these gene sets to carry out 1000 independent clusterings of the primary prostate cancer samples, and subjected each grouping to Kaplan-Meier survival analysis.

To validate the prognostic association of the 44-gene concordant signature, an independent (clinically localized) prostate cancer gene expression dataset from Glinsky *et al*. (45) was used. The Yu *et al.* clustering functioned as the “training set” to define high-/low-risk groups. Each “test” sample of the Glinsky *et al*. study was classified into one of the two groups based on *k*-nearest neighbor classification (*k*=3). Kaplan-Meier survival curves were plotted for the two groupings. This “progression” signature was then refined by reducing the number of genes involved. By using Yu *et al*. study as a training set, the concordant genes were ranked by univariate Cox model. Again, the clustering procedure was used to identify two clusters based on the top number of genes (ranging from 5 to 44). The Glinsky *et al*. study was then used as a validation set to verify performance of the refined signature by *k*-nearest neighbors (*k*=3) prediction analysis. The generality of this “progression” signature was evaluated by using other solid tumor datasets. The signature was applied to two breast cancer (51)$^{-}$(50), one lung cancer (52) and one glioma (53) gene expression study. Hierarchical clustering was used to identify two main clusters for patients in each study and Kaplan-Meier survival analysis was conducted to evaluate the statistical significance of differences between survival curves.

Multivariable Analysis. We used a Cox proportional-hazards regression model to carry out the multivariate analysis. The dichotomized values of the concordant “progression” signature, preoperative PSA, Gleason sum score from prostatectomy specimens, preoperative clinical stage, age, and status of surgical margins were included as covariates. The calculation was performed with the R 2.0 statistical package.

**Table 7.1** Multivariable Cox proportional-hazards analysis of the risk of recurrence as a first event on the Glinsky et al. validation set.

**Table 7.1**. Multivariable proportional-hazards analysis of the risk of recurrence as a first event on the Glinsky et. al. validation Set

| **Variable** | **Hazard Ratio (95% CI)** | **P Value** |
|---|---|---|
| High-Risk signature (vs. low-risk signature ) | 3.66 (1.77 – 7.59) | <0.001 |
| PSA | 1.04 (1.00 – 1.09) | 0.043 |
| Gleason Sum Score | | |
| Score >7 (vs. score <=7) | 1.73 (0.79 – 3.76) | 0.17 |
| Tumor Stage | | |
| Stage T2 (vs. stage T1) | 0.85 (0.42 – 1.75) | 0.67 |
| Age | 1.06 (1.00 – 1.13) | 0.06 |
| Surgical Margins | | |
| Positive (vs. negative) | 2.18 (0.92 – 5.18) | 0.08 |

**Table 7.2** Comparisons of the performances of the progression signature and study-specific signatures in individual study cohort.

| **Signature** | **Cohort** | | | |
|---|---|---|---|---|
| | Glinsky et al.<br>(Prostate) [2] | Van’t Veer et al.<br>(Breast) | Huang et al.<br>(Breast) | Freije et al.<br>(Glioma) |
| # of genes in a signature [1] | 14 | 70 | 164 | 595 |
| **Diagonal linear discriminant analysis** | | | | |
| Progression signature | 73% (19 of 26) | 79% (15 of 19) | 73% (38 of 52) | NA |
| Study-specific signature | 73% (19 of 26) | 79% (15 of 19) | 83% (43 of 52) | NA |
| ***k*-nearest neighbor analysis (*k*=3)** | | | | |
| Progression signature | 77% (20 of 26) | 79% (15 of 19) | 69% (36 of 52) | NA |
| Study-specific signature | 57% (15 of 26) | 68% (13 of 19) | 77% (40 of 52) | NA |
| **Kaplan-Meier survival analysis (log rank test)** [3] | | | | |
| Progression signature | NA | NA | NA | $P = 3.7x\ 10^{-5}$ |
| Study-specific signature | NA | NA | NA | $P = 2x10^{-4}$ |

1. The number of genes in a signature reported in the original study. When the signature was applied to other datasets, genes were cross-referenced by UniGene cluster IDs and values of multiple reporters mapping to the same gene were averaged.
2. For Glinsky et al. study, we randomly assigned two-thirds samples into a training set and used the rest of samples as a validation set due to that the entire data was used as a validation set in the original publication; For van’t veer et al. and Freije et al studies, we used same validation sets as described in the original publications. For Huang et al. study, we followed the same strategy as described in the study and used a leave-one-out cross validation in order to make a fair comparison. All of accuracies reported here were calculated based on the validation sets.
3. A log rank test, same as described in the study was used to evaluate the difference of two distinct patient groups derived from two main clusters of hierarchical clustering performed on the validation set.

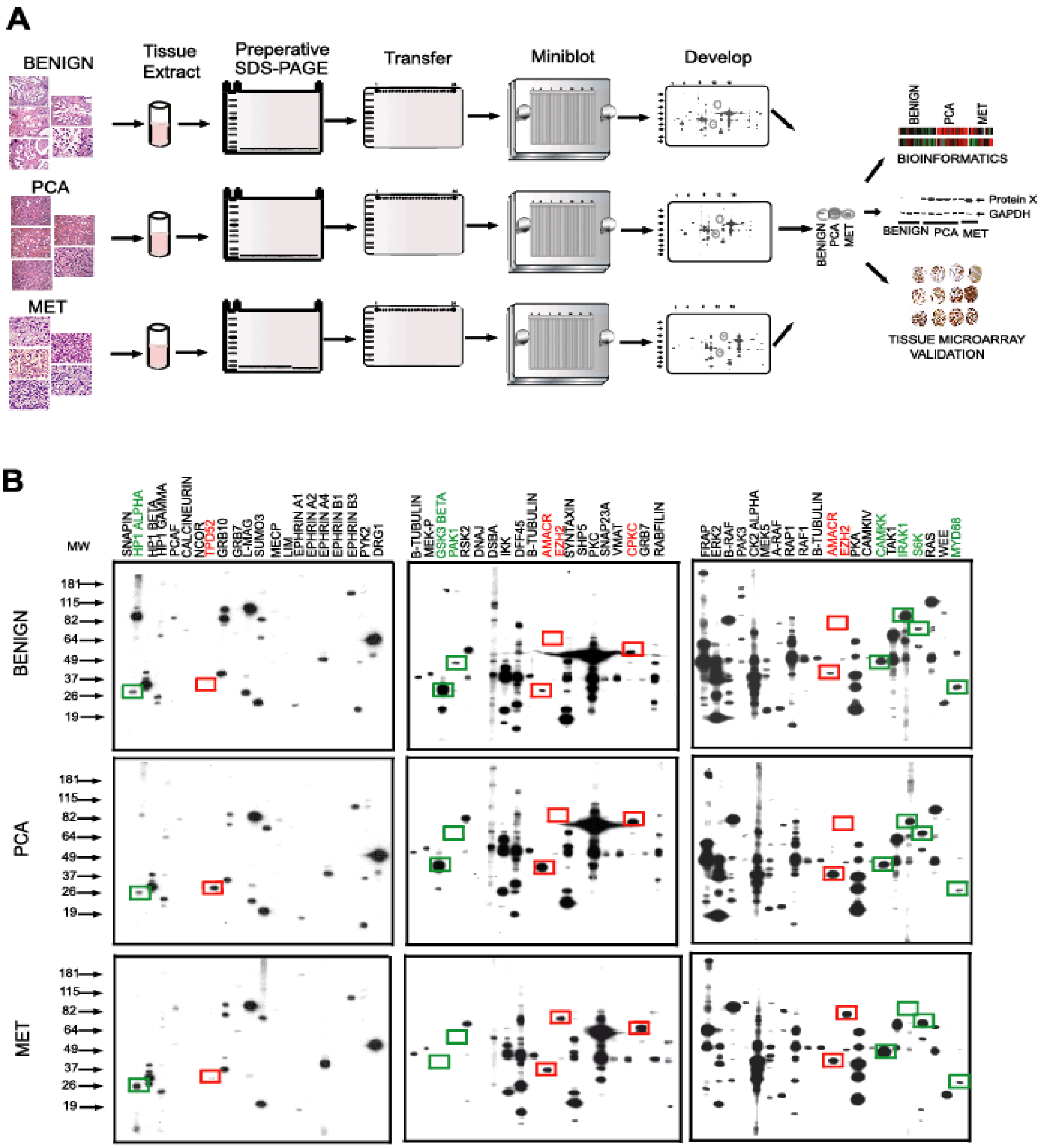


**Figure 7.1** High-throughput immunoblot analysis to define proteomic alterations in prostate cancer progression. **A,** A flowchart of the general methodology employed to profile proteomic alterations in tissue extracts. Pooled tissue extracts (n=5 each) from clinically localized prostate cancer, hormone-refractory metastatic prostate cancer, and benign prostate tissues were separated on preparative SDS-PAGE gels and transferred to PVDF membranes. The membranes were incubated with commercial antibodies using a miniblotter system. PCA, clinically localized prostate cancer. MET, metastatic prostate cancer. **B,** Representative high-throughput immunoblots performed for pooled benign, clinically localized prostate cancer and metastatic prostate cancer tissues. Each lane represents analysis of an individual protein. Three representative blots are displayed for each tissue extract. Selected proteins altered in prostate cancer progression are highlighted. MW, molecular weight.

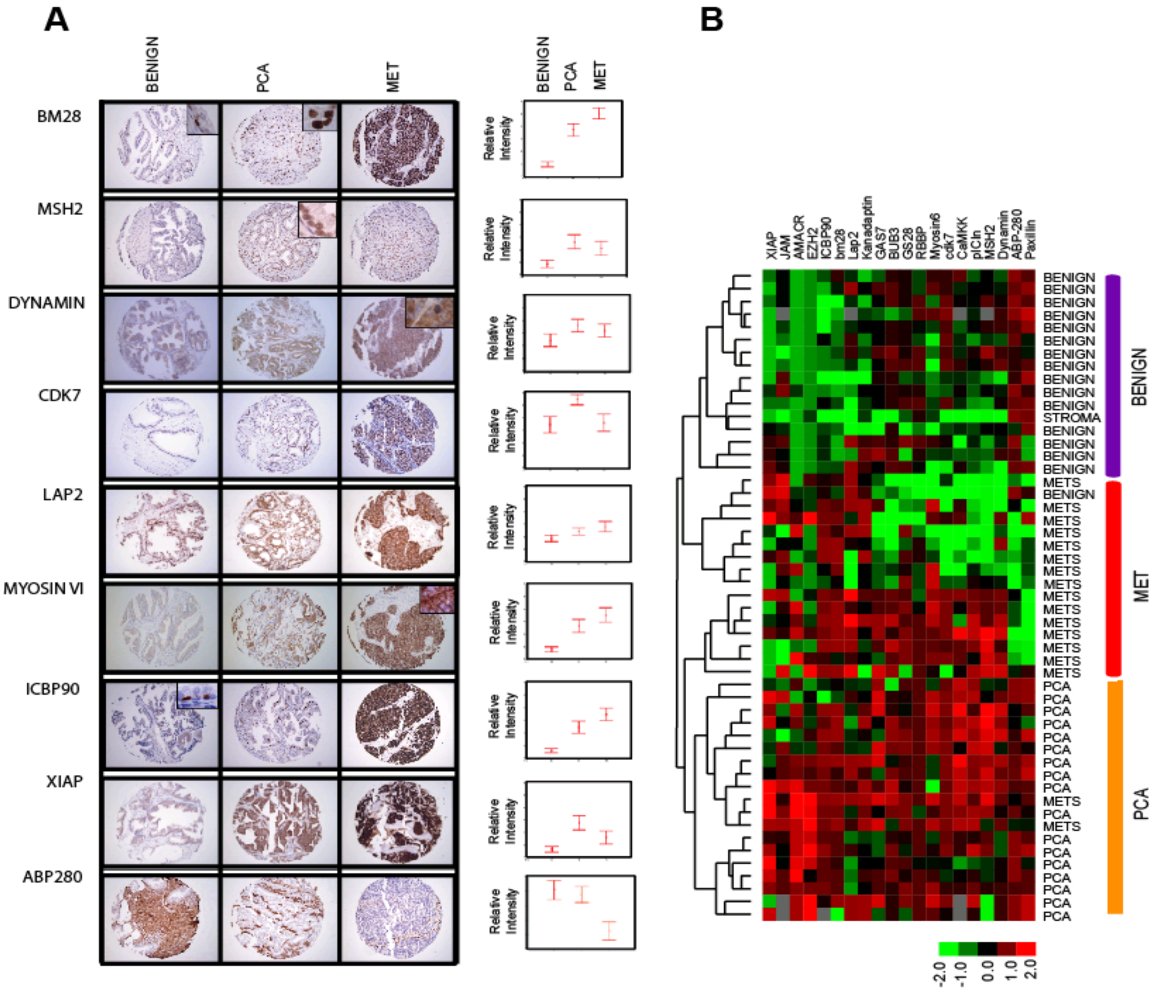


**Figure 7.2** Tissue microarray analyses of protein markers deregulated in prostate cancer progression. **A**. Selected images of tissue microarray elements representing immunohistochemical analysis of proteins altered in prostate cancer progression. Relative levels of proteins as assessed by blinded pathology analysis of tissue microarrays (n=216 specimens) is provided to the right. **B,** Cluster analysis of twenty proteins dysregulated in prostate cancer progression evaluated for in situ protein levels by tissue microarrays. Unsupervised hierarchical clustering of protein levels (columns) and samples (rows) was performed and a heatmap generated. Red color represents high protein levels while black refers to intermediate levels and green represents low or absent protein levels.

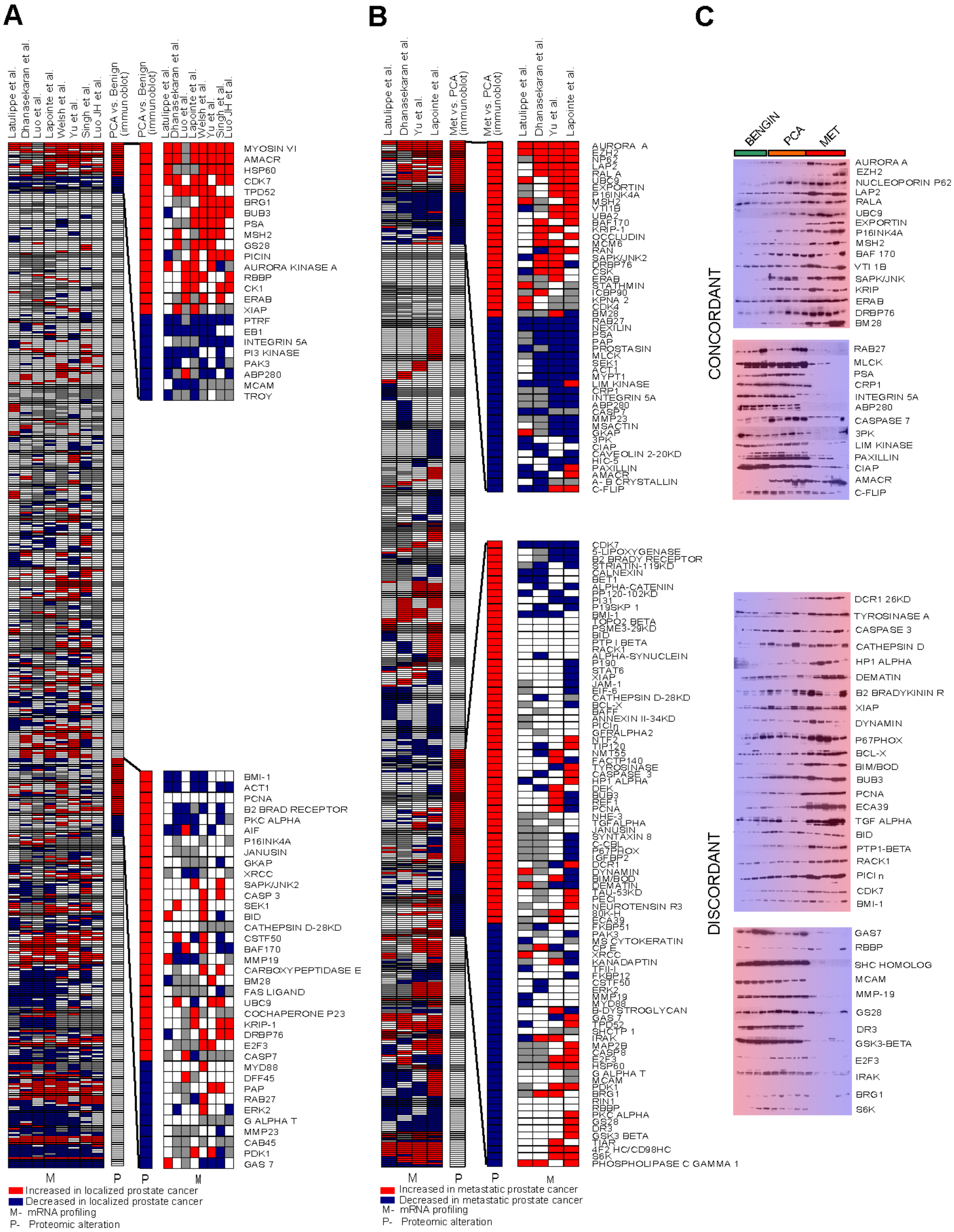


**Figure 7.3** Integrative analysis of proteomic and transcriptomic meta-data in prostate cancer progression. **A,** Color map of integrative analysis relating protein alterations to gene expression in clinically localized prostate cancer relative to benign prostate tissue. For gene expression meta-analysis (transcript analysis), the first author of each prostate cancer gene expression study is indicated in columns while individual genes are

represented as rows. Red color indicates significantly increased expression at the *P*=0.05 threshold level for prostate cancer relative to benign tissue, while blue indicates down-regulation at the same threshold, and white indicates unchanged expression. Protein levels (protein) in pooled clinically localized prostate cancer extracts (as described in **Figure 7.1**), were visually qualified by high-throughput immunoblot analysis as over-expressed (red), under-expressed (blue), or unchanged (white) and mapped to the corresponding mRNA transcript. Proteins which were not expressed (or corresponding antibodies that did not produce an immunoreactive band of the correct molecular weight) or the corresponding mRNA transcript level was not present in over one fourth of the profiling studies were excluded from the integrative analysis. Proteomic alterations in prostate cancer that were concordant or discordant with the meta-analysis of gene expression were expanded to the right. **B**, As in **A** except the integrative analysis was carried out between metastatic prostate cancer relative to clinically localized prostate cancer. **C**, Conventional immunoblot validation of selected proteins differentially expressed between metastatic prostate cancer and clinically localized prostate cancer. Individual tissue extracts from 3-4 benign, 5 clinically localized prostate cancer, and 5 metastatic prostate cancer samples are shown.

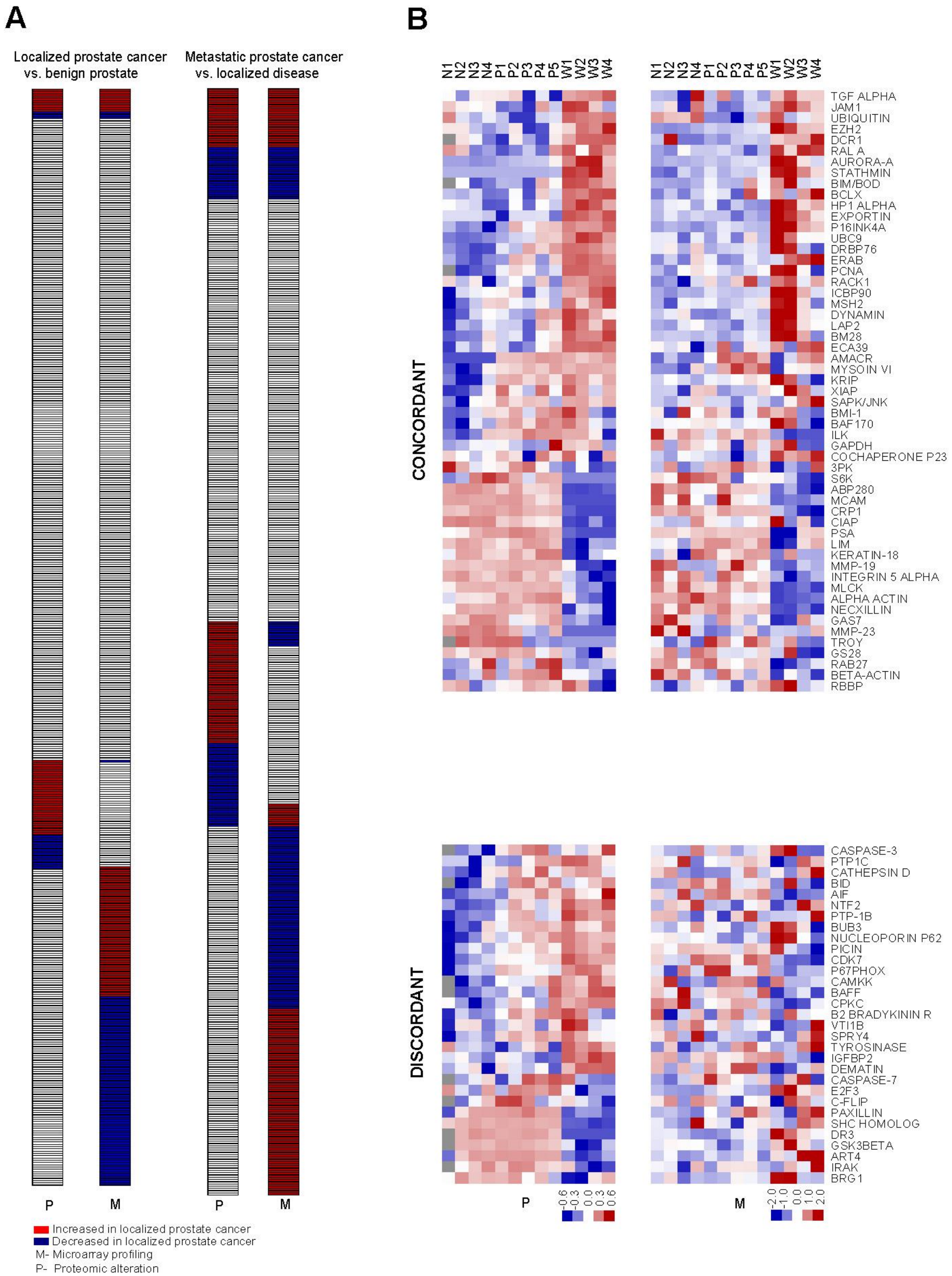


**Figure 7.4** Integrative genomic and proteomic analysis of pooled and individual prostate tissue extracts. **A.** Color maps of integrative analyses relating protein alterations observed in pooled tissues by immunoblotting and transcript alterations observed in the pooled and

individual tissues by gene expression analyses. Please refer to **Figure 7.3A** for color scheme. **B,** Color maps depicting integrative genomic and proteomic analysis of individual prostate tissue samples. Proteins in each tissue extract were assessed based on intensities derived from conventional immunoblot analysis. We focused on the 86 proteins identified as outliers in the larger high-throughput screen. Transcriptomic profiles from the same samples were derived from Affymetrix microarrays. The immunoblot intensities were semi-quantitated and a correlation was calculated for each protein. Concordance was defined based on positive correlation between proteins and transcripts (See Methods).

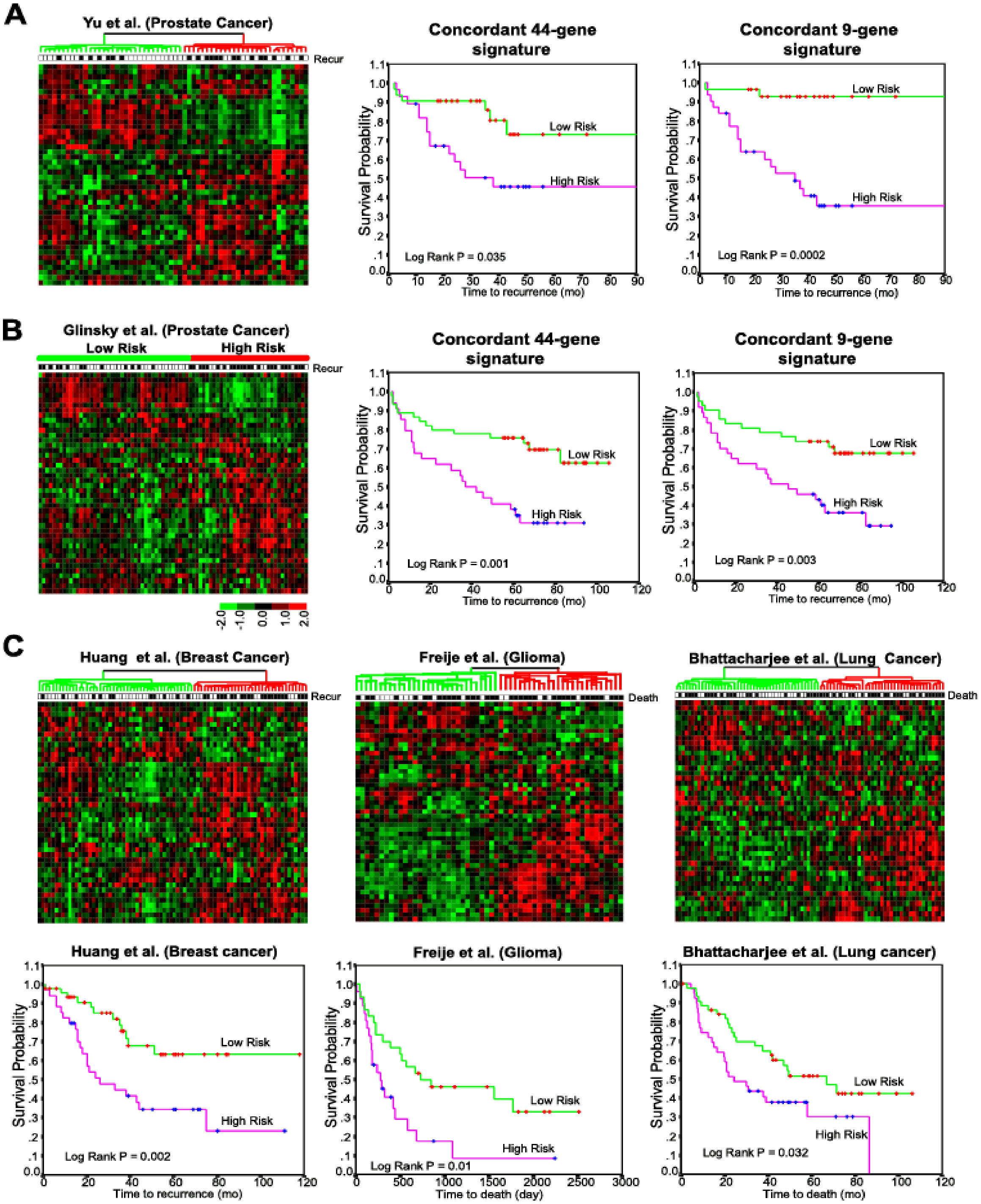


**Figure 7.5** Proteomic alterations in metastatic prostate cancer identify gene predictors of cancer aggressiveness. **A**, A concordant 44- (out of 50) gene predictor was developed based on proteomic alterations that were concordant with gene expression (**Figure 7.3B**) and subsequently evaluated for prognostic utility on a prostate cancer gene expression dataset (Yu *et al.*). Hierarchical clustering of the tumor samples (columns) and genes (rows) is provided (left panel). Red indicates high relative levels of gene expression while green represents low relative levels of gene expression. Horizontal bars above the heat

maps indicate the recurrence status of each patient (black box, biochemical or tumor recurrence; white box, recurrence-free). Patients were categorized into two major clusters defined by the 44-gene signature. The prediction model was further refined to a 9-gene signature. Kaplan-Meier survival analysis based on the groups defined by the 44-gene concordant cluster (middle panel) and the 9-gene concordant cluster (right panel). **B,** The concordant 44-gene predictor and the refined concordant 9-gene predictor were evaluated in an independent prostate cancer profiling dataset. Each sample was assigned to a low-risk or high-risk group by *k*-nearest neighbor classification using cluster-defined low-/high-risk groups of the Yu *et al.* as a training dataset (left panel; see methods). Kaplan-Meier plot of the predicted high-/low-risk groups in the space of the concordant 44 genes (middle panel) or the concordant 9 genes (right panel). **C,** Same as **A**, except the concordant predictor was evaluated in other solid tumors. Huang *et al*. (51) breast adenocarcinoma (left panel), Freije *et al*. glioma (53) (middle panel), and Bhattacharjee *et al*. (52) lung adenocarcinoma (right,).

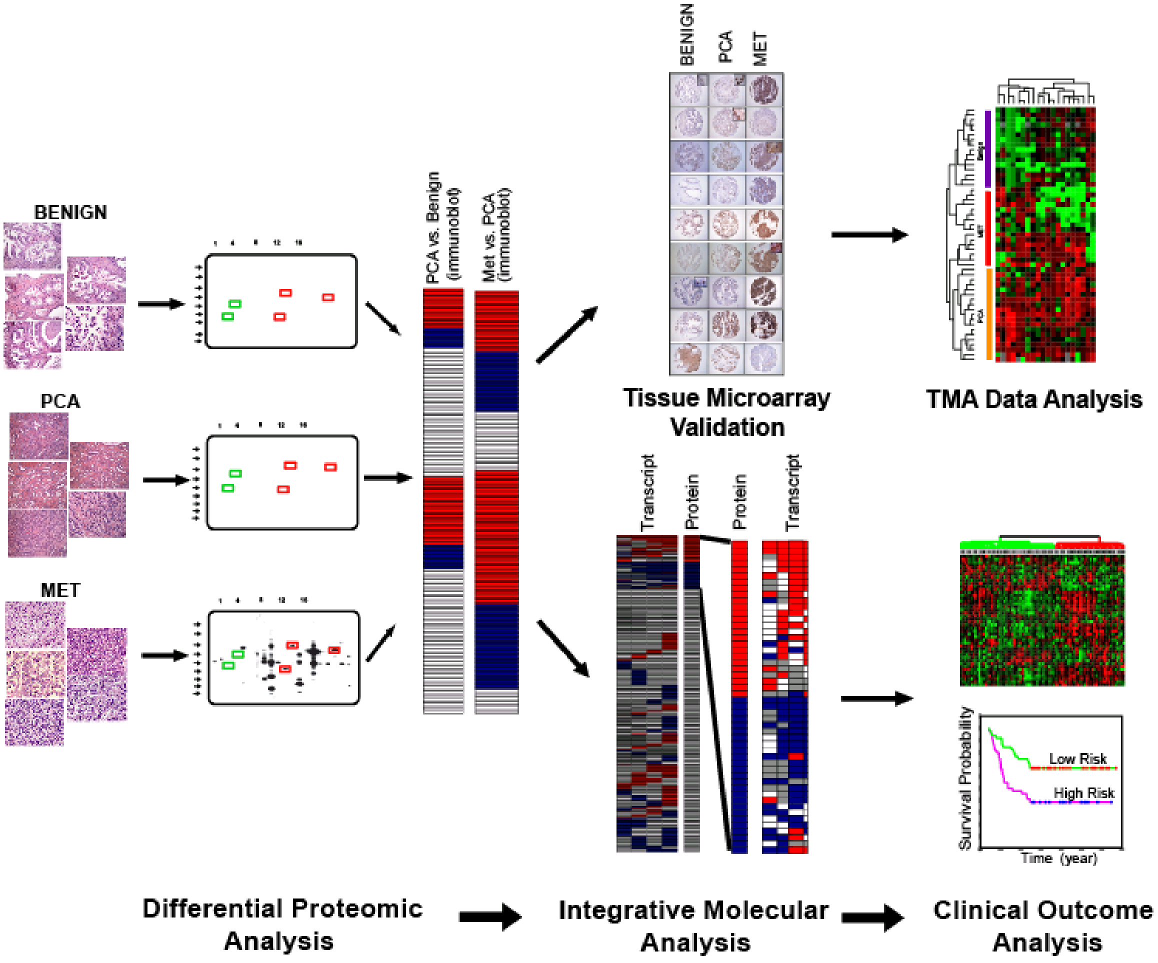


**Figure 7.6** Integrative molecular analysis of cancer to identify gene predictors of clinical outcome. Proteomic profiles comparing metastatic prostate cancer to clinically localized prostate cancer were used to identify a composite gene predictor of clinical outcome in localized disease. This integrated proteomic-transcriptomic signature represents a prostate cancer progression signature and can be extended to other solid tumors.

# REFERENCES

1. Chan, JM, Jou, RM, and Carroll, PR (2004). The relative impact and future burden of prostate cancer in the United States. *J Urol* **172**, S13-16; discussion S17.

2. Linton, KD, and Hamdy, FC (2003). Early diagnosis and surgical management of prostate cancer. *Cancer Treat Rev* **29**, 151-160.

3. Johansson, JE, Holmberg, L, Johansson, S, Bergstrom, R, and Adami, HO (1997). Fifteen-year survival in prostate cancer. A prospective, population-based study in Sweden. *Jama* **277**, 467-471.

4. Albertsen, PC, Hanley, JA, Gleason, DF, and Barry, MJ (1998). Competing risk analysis of men aged 55 to 74 years at diagnosis managed conservatively for clinically localized prostate cancer. *Jama* **280**, 975-980.

5. Kumar-Sinha, C, and Chinnaiyan, AM (2003). Molecular markers to identify patients at risk for recurrence after primary treatment for prostate cancer. *Urology* **62 Suppl 1**, 19-35.

6. Hood, L, Heath, JR, Phelps, ME, and Lin, B (2004). Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640-643.

7. Grubb, RL, Calvert, VS, Wulkuhle, JD, et al. (2003). Signal pathway profiling of prostate cancer using reverse phase protein arrays. *Proteomics* **3**, 2142-2146.

8. Petricoin, EF, 3rd, Ornstein, DK, Paweletz, CP, et al. (2002). Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* **94**, 1576-1578.

9. Paweletz, CP, Charboneau, L, Bichsel, VE, et al. (2001). Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981-1989.

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

11. Lapointe, J, Li, C, Higgins, JP, et al. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* **101**, 811-816.

12. LaTulippe, E, Satagopan, J, Smith, A, et al. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res* **62**, 4499-4506.

13. Luo, J, Duggan, DJ, Chen, Y, et al. (2001). Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res* **61**, 4683-4688.

14. Luo, JH, Yu, YP, Cieply, K, et al. (2002). Gene expression analysis of prostate cancers. *Mol Carcinog* **33**, 25-35.

15. Magee, JA, Araki, T, Patil, S, et al. (2001). Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Res* **61**, 5692-5696.

16. Singh, D, Febbo, PG, Ross, K, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203-209.

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

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

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

20. Hedenfalk, I, Duggan, D, Chen, Y, et al. (2001). Gene-expression profiles in hereditary breast cancer. *N Engl J Med* **344**, 539-548.

21. Perou, CM, Sorlie, T, Eisen, MB, et al. (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747-752.

22. Alizadeh, AA, Eisen, MB, Davis, RE, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

23. Varambally, S, Dhanasekaran, SM, Zhou, M, et al. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature* **419**, 624-629.

24. Rubin, MA, Zhou, M, Dhanasekaran, SM, et al. (2002). alpha-Methylacyl coenzyme A racemase as a tissue biomarker for prostate cancer. *Jama* **287**, 1662-1670.

25. Luo, J, Zha, S, Gage, WR, et al. (2002). Alpha-methylacyl-CoA racemase: a new molecular marker for prostate cancer. *Cancer Res* **62**, 2220-2226.

26. Jiang, Z, Woda, BA, Rock, KL, et al. (2001). P504S: a new molecular marker for the detection of prostate carcinoma. *Am J Surg Pathol* **25**, 1397-1404.

27. Kononen, J, Bubendorf, L, Kallioniemi, A, et al. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* **4**, 844-847.

28. Nielsen, TO, Hsu, FD, O'Connell, JX, et al. (2003). Tissue microarray validation of epidermal growth factor receptor and SALL2 in synovial sarcoma with comparison to tumors of similar histology. *Am J Pathol* **163**, 1449-1456.

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

30. Krajewska, M, Krajewski, S, Banares, S, et al. (2003). Elevated expression of inhibitor of apoptosis proteins in prostate cancer. *Clin Cancer Res* **9**, 4914-4925.

31. Griffin, TJ, Gygi, SP, Ideker, T, et al. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol Cell Proteomics* **1**, 323-333.

32. Washburn, MP, Koller, A, Oshiro, G, et al. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **100**, 3107-3112.

33. Baliga, NS, Pan, M, Goo, YA, et al. (2002). Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach. *Proc Natl Acad Sci U S A* **99**, 14913-14918.

34. Tian, Q, Stepaniants, SB, Mao, M, et al. (2004). Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics* **3**, 960-969.

35. Shah, RB, Mehra, R, Chinnaiyan, AM, et al. (2004). Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program. *Cancer Res* **64**, 9209-9216.

36. Mooradian, AD, Morley, JE, and Korenman, SG (1987). Biological actions of androgens. *Endocr Rev* **8**, 1-28.

37. Culig, Z, Hobisch, A, Hittmair, A, et al. (1998). Expression, structure, and function of androgen receptor in advanced prostatic carcinoma. *Prostate* **35**, 63-70.

38. Koivisto, P, Kolmer, M, Visakorpi, T, and Kallioniemi, OP (1998). Androgen receptor gene and hormonal therapy failure of prostate cancer. *Am J Pathol* **152**, 1-9.

39. DePrimo, SE, Diehn, M, Nelson, JB, et al. (2002). Transcriptional programs activated by exposure of human prostate cancer cells to androgen. *Genome Biol* **3**, RESEARCH0032.

40. Nelson, PS, Clegg, N, Arnold, H, et al. (2002). The program of androgen-responsive genes in neoplastic prostate epithelium. *Proc Natl Acad Sci U S A* **99**, 11890-11895.

41. Segawa, T, Nau, ME, Xu, LL, et al. (2002). Androgen-induced expression of endoplasmic reticulum (ER) stress response genes in prostate cancer cells. *Oncogene* **21**, 8749-8758.

42. Velasco, AM, Gillis, KA, Li, Y, et al. (2004). Identification and validation of novel androgen-regulated genes in prostate cancer. *Endocrinology* **145**, 3913-3924.

43. Waghray, A, Feroze, F, Schober, MS, et al. (2001). Identification of androgen-regulated genes in the prostate cancer cell line LNCaP by serial analysis of gene expression and proteomic analysis. *Proteomics* **1**, 1327-1338.

44. Xu, LL, Su, YP, Labiche, R, et al. (2001). Quantitative expression profile of androgen-regulated genes in prostate cancer cells and identification of prostate-specific genes. *Int J Cancer* **92**, 322-328.

45. Glinsky, GV, Glinskii, AB, Stephenson, AJ, Hoffman, RM, and Gerald, WL (2004). Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest* **113**, 913-923.

46. Kleer, CG, Cao, Q, Varambally, S, et al. (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A* **100**, 11606-11611.

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

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

49. Rhodes, DR, Yu, J, Shanker, K, et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6.

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

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

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

53. Freije, WA, Castro-Vargas, FE, Fang, Z, et al. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* **64**, 6503-6510.

54. Jeng, YM, Peng, SY, Lin, CY, and Hsu, HC (2004). Overexpression and amplification of Aurora-A in hepatocellular carcinoma. *Clin Cancer Res* **10**, 2065-2071.

55. Neben, K, Korshunov, A, Benner, A, et al. (2004). Microarray-based screening for molecular markers in medulloblastoma revealed STK15 as independent predictor for survival. *Cancer Res* **64**, 3103-3111.

56. Hirota, T, Kunitoku, N, Sasayama, T, et al. (2003). Aurora-A and an interacting activator, the LIM protein Ajuba, are required for mitotic commitment in human cells. *Cell* **114**, 585-598.

57. Katayama, H, Sasai, K, Kawai, H, et al. (2004). Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53. *Nat Genet* **36**, 55-62.

58. Schultz, DC, Ayyanathan, K, Negorev, D, Maul, GG, and Rauscher, FJ, 3rd (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev* **16**, 919-932.

59. Mecham, BH, Klus, GT, Strovel, J, et al. (2004). Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* **32**, e74.

60. Troyanskaya, O, Cantor, M, Sherlock, G, et al. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525.

# PART 5: CONCLUSION

## CHAPTER 8

### Conclusion

Molecular cancer classification, the classification of tissue or other specimens for diagnostic, prognostic, and predictive purposes on the basis of multiple gene expression, has been demonstrated a promising technology for optimizing the management of patients with cancer. In this dissertation, I have developed (1) an estrogen-regulated gene signature that can robustly predict cancer outcome in human breast cancer; (2) efficient yet comprehensible molecular classifiers using genetic programming for cancer classification; (3) non-invasive diagnostic tools for early detection of prostate cancer based on either patient serum or patient urine profiling; (4) a system approach to model metastatic progression in prostate cancer. These results support that high-throughput microarray profiling and resulted candidate biomarkers, if used properly and thoughtfully, are capable of developing more accurate diagnostic or prognostic tests for human cancer in clinic, supplementary to traditional histopathological methods.

Breast cancer is the most common cancer among women in the US, accounting for nearly 1 of every 3 cancers diagnosed. Despite of current advance in breast cancer research, accurate prognosis for breast cancer patients has been a more daunting task. In this thesis work (**Chapter 3**), we analyzed gene expression profiles of breast cancer cells in vitro and in vivo in order to uncover a molecular signature which may serve as a better indicator of cancer outcome. We set out to mine estrogen signaling pathway to identify estrogen-regulated genes as estrogen plays an essential role in breast cancer progression. We focused on *in vitro* estrogen-regulated genes and further selected a subset that is associated with patient outcome *in vivo* in human breast tumors. The final 73-gene signature developed by leave-one-out cross validation successfully predicts clinical outcome in over ten patient cohorts. Besides correctly assigning most ER- tumors in each dataset into high-risk group, this signature is able to stratify the ER+ samples into prognostic subtypes, suggesting that it may better reflect tumor aggressiveness than ER status alone. Most importantly, the signature provides additional prognostic information beyond standard clinical factors and yields overall best performance against previously reported breast cancer outcome predictors. This signature may be thus valuable in selection of high-risk patients for adjuvant therapy as well as in sparing some hormone-sensitive patients from aggressive therapy.

One important facet of clinical tests is cost-effectiveness, which makes the expression profiling of a large number of genes simultaneously less attractive in clinical trials. Thus, developing accurate yet simple classifiers using a handful of genes is in high demand. In this dissertation, we evaluated the capability of one evolutionary algorithm, Genetic Programming (GP), in building molecular classifiers using a practical set of genes (**Chapter 4**). We tested it on one Small Round Blue Cell Tumors (SRBCTs), one lung adenocarcinoma and five prostate cancer datasets. We have found that GP repetitively uses a small set of highly discriminative feature genes to produce classifiers, which often comprise five or less genes and are able to predict samples correctly in independent datasets. As GP utilizes the quantitative information among genes to connect genes each other, the classifiers generated by GP are usually simple and human-readable. In addition, comparing to other conventional classification methods, GP yields better or similar classification performance. Thus, given these unique characteristics of GP, it stands out as a good algorithm of choice for application of DNA microarray profiling to clinic.

Contrast to breast cancer in women, prostate cancer is the most common form of cancer affecting men in the Western world. Current common screening test for prostate cancer is to use prostate specific antigen (PSA). While PSA testing has led to a dramatic increase in the detection of prostate cancer, it has substantial false positive rate, supporting that additional cancer biomarkers or signatures may be required to ameliorate the accuracy of prostate cancer diagnosis. As cancer patients produce autoantibodies against antigens in their tumors and prostate cells can be detected in the urine of patients with prostate cancer, serum/urine based diagnostic tests have the advantage of being non-invasive. In **Chapter 5**, we developed a phase-display protein microarrays to analyze serum samples from 119 patients with prostate cancer and 138 controls. By profiling global humoral immune response of these samples, we discovered 22 phage peptides grouped as a predictor that yielded 88.2% specificity and 81.6% sensitivity in discriminating between the group with prostate cancer and the control group of a validation set and outperformed PSA testing. This work demonstrates the feasibility of multiplex humoral immune response as the basis for a screening test for prostate cancer although further extension and confirmation in community-based screening cohorts is needed.

In **Chapter 6**, we combine prostate cancer "outlier" genes and known prostate cancer biomarkers to develop a multiplexed qPCR based urine test for detection of prostate cancer. Previously identified "outlier" genes (*ERG*, *TMPRSS2:ERG* and *SPINK1*) by our lab and known prostate cancer biomarkers such as *PCA3*, *AMACR*, *GOLPH2* are assessed in sedimented urine using qPCR. We analyzed 234 patient samples and found that a multiplexed model including *PCA3*, *GOLPH2*, *SPINK1* and *TMPRSS2:ERG* yielded an area under roc (AUC) of 0.76, significantly outperforming serum PSA or *PCA3* alone in detecting prostate cancer (AUC 0.57 for serum PSA, and 0.66 for *PCA3*). While urine-based testing for *PCA3* expression has already been documented in large screening programs, we demonstrate that this multiplexed qRT-PCR based assay can further improve the diagnostic accuracy of prostate cancer.

Finally, with the explosion of gene expression data and the advent of high-throughput proteomic profiling, interrogative efforts in both oncoproteomics and the cancer transcriptomics ushered in a 'systems' era that necessitates integrated approaches to analysis. **Chapter 7** delineates an integrative model for culling a molecular signature of metastatic progression in prostate cancer from proteomic and transcriptomic analyses. Proteomic profiling of prostate cancer progression identified over one hundred altered proteins in the transition from clinically localized to metastatic disease. These differential proteins were then mapped to mRNA transcript levels in multiple expression studies to assess mRNA/protein concordance levels in prostate cancer, leading to discovery of a 50-gene signature of prostate cancer progression. While this approach is used to integrate high-throughput immunoblot data in this thesis work, the general framework of integrating multiple-source data can be extended to other proteomic platforms such as quantitative mass spectrometry, or protein microarray based technologies as they mature in the future. More critically, the discovered 50-gene signature not only predicts clinical outcome in localized prostate cancer, but also can be extrapolated to other solid tumor types including primary tumors of the breast, lung, and gliomas, suggesting common molecular machinery in poorly differentiated aggressive neoplasms. This is a powerful validation of the integrative model showing clinical import. The marriage of such a model with those in early detection has the potential to manifest in overt survival benefit for cancer patients.

# APPENDIX

Supplementary Information for the individual chapters is available online at the following addresses:

CHAPTER 4

http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1854845&blobname=neo0904_0292SD1.pdf

CHAPTER 5

http://content.nejm.org/cgi/content/full/353/12/1224/DC1

CHAPTER 7

http://www.cancercell.org/cgi/content/full/8/5/393/DC1/

Multiple individuals contributed to the work presented in these chapters and resulting manuscripts. Contributions of individuals for each chapter are as follows:

CHAPTER 3

Jianjun Yu, Marc Lippman, and Arul Chinnaiyan conceived the experiments represented in this chapter. Jianjun Yu, Jindan Yu and Arul Chinnaiyan wrote the manuscript. Jianjun Yu performed all in silico analyses. James Rae, Kevin Cordero, and Michael Johnson performed gene expression profiling of breast cancer cell lines. Debashis Ghosh provided biostatistical support.

## CHAPTER 4

Jianjun Yu, William Worzel and Arul Chinnaiyan conceived the experiments. Jianjun Yu, Jindan Yu and Arul Chinnaiyan wrote the manuscript. Jianjun Yu, with technical assistance of Arpit Almal, performed the data analysis. William Worzel and Arpit Almal provided hardware and software support. Debashis Ghosh provided biostatistical support and Saravana Dhanasekaran provided gene expression data.

## CHAPTER 5

Xiaoju Wang, Jianjun Yu, and Arul Chinnaiyan conceived the experiments and wrote the manuscript represented in this chapter. Xiaoju Wang, with assistance of Arun Sreekumar, designed the phage-display microarray system and performed sample profiling. Jianjun Yu and Ronglai Shen, with biostatistical support of Debashis Ghosh, performed all data analysis. Sooryanarayana Varambally performed immunoblot validation and immunofluorescence staining. Mark Rubin and Rohit Mehra performed immunohistochemical analysis of tissue microarrays. John Wei, Kenneth Pienta and Philip Kantoff obtained serum samples and tissue information.

## CHAPTER 6

Bharathi Laxman, David Morris, Jianjun Yu, Scott Tomlins and Arul Chinnaiyan conceived the experiments and wrote the manuscript. David Morris obtained the urine samples and Bharathi Laxman performed qPCR experiments. Jianjun Yu developed the multiplex model and performed survival analysis of tissue microarrays. Daniel Rhodes performed outlier analysis. Rohit Mehra, and Anders S. Bjartell performed immunohistochemical analysis of tissue microarrays.

CHAPTER 7

Sooryanarayana Varambally, Jianjun Yu and Arul Chinnaiyan conceived the project and wrote the manuscript. Jianjun Yu performed in silico analyses and developed the integrative model. Sooryanarayana Varambally and Bharathi Laxman performed high-throughput immunoblot. Uma Chandran, Federico Monzon, and Michael Becich provided a set of gene expression data. John Wei, Rajal Shah, Kenneth Pienta obtained tissue samples. Rohit Mehra, and Mark Rubin evaluated immunohistochemistry on tissue microarrays. Daniel Rhodes performed pathway analysis. Debashis Ghosh provided biostatistical support.