UNIVERSITY OF CALIFORNIA, SAN DIEGO

PATHWAY-BASED MODELING AND DIAGNOSIS OF CANCER DEVELOPMENT AND PROGRESSION

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Han-Yu Chuang

Committee in charge:

Professor Trey Ideker, Chair
Professor Thomas Kipps, Co-chair
Professor Steve Briggs
Professor Sanjoy Dasgupta
Professor Jean Wang

2010

The dissertation of Han-Yu Chuang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-chair

_____
Chair

University of California, San Diego
2010

DEDICATION

To my grandparents

I have always been missing you.

To my parents

This is for you.

To Irene

I couldn't have done this without you.

To Julie and Phoeny

My cutest life savers.

EPIGRAPH

Information is not knowledge.

Albert Einstein (Nobel Prize for Physics in 1921. 1879-1955)

When you think about it, what other choice is there but to hope? We have two options, medically and emotionally: give up, or Fight Like Hell.

Lance Armstrong (American Cyclist, b.1971)

TABLE OF CONTENTS

Signature Page ..... **iii**

Dedication ..... **iv**

Epigraph ..... **v**

Table of Contents ..... **vi**

List of Figures ..... **ix**

List of Tables ..... **xii**

Acknowledgements ..... **xiii**

Vita ..... **xvii**

Abstract of The Dissertation ..... **xix**

1. INTRODUCTION **1**

1.1 Gene expression analyses in cancer systems biology 7

1.2 Pathway-based expression analyses 10

1.3 Protein interaction networks 12

2. PINNACLE **16**

2.1 Scoring subnetworks 17

2.2 Searching for significant subnetworks 18

2.3 Software availability 19

2.4 Acknowledgement 19

3. NETWORK-BASED CLASSIFICATION OF BREAST CANCER METASTASIS **24**

3.1 Background and significance 24

3.2 Overview of subnetwork marker identification 27

3.3 Subnetwork markers correspond to the hallmarks of cancer 30

3.4 Subnetwork markers have increased reproducibility across datasets 31

3.5 Subnetwork markers increase the classification accuracy of metastasis 32

3.6 Subnetwork markers are informative of non-discriminative disease genes.....36

3.7 Acknowledgement.....38

4. NETWORK-BASED ANALYSIS OF CHRONIC LYMPHOCYTIC LEUKEMIA IDENTIFIES PATHWAYS THAT CONTRIBUTE TO DISEASE EVOLUTION.....51

4.1 Background and significance 51

4.2 Gene expression profiling of peripheral blood from CLL patients.....54

4.3 IGHV mutation status cannot reliably predict treatment-free survival from sample collection 55

4.4 Transcriptional activity converges between patients of different IGHV status as disease advances 56

4.5 Previous microarray studies yield gene markers of equivalent or less predictive power as IGHV status 57

4.6 Protein networks stratify CLL patients into different risk groups 58

4.7 Predicting the timing of therapy from the date of sample collection 59

4.8 Convergence of dynamic cll subnetwork transcriptome with disease progression..61

4.9 Scoring, searching, and pruning subnetworks.....63

4.10 Prognosis evaluation 64

4.11 Real time PCR for serial gene expression 65

4.12 Protein expression analysis using flow cytometry and immuno-blotting 66

4.13 Discussion 66

4.14 Acknowledgement.....70

5. INFERRING PATHWAY ACTIVITY TOWARD PRECISE DISEASE CLASSIFICATION.....84

5.1 Background and significance 85

5.2 Datasets 87

5.3 Condition-responsive gene identification and pathway activity inference 88

5.4 Previous gene-set ranking approaches and other pathway-based classification methods 89

5.5 Marker robustness evaluation.....90

5.6 Classification evaluation 91

5.7 Pathway markers amplify signals over multiple weak gene markers 93

5.8 Pathway markers increase the classification accuracy.....94

5.9 Pathway markers and their corgs provide biologically informative models for lung cancer prognosis 96

5.10 Acknowledgement 97

6. DISSECTING DISEASE PROGRESSION OF CHRONIC LYMPHOCYTIC LEUKEMIA USING AN INTEGRATED QUANTITATIVE PROTEOMIC AND GENOMIC ANALYSIS 110

6.1 Background and significance 111

6.2 MS-based shotgun proteomics with isobaric tag for relative and absolute quantification 114

6.3 pSAM − Significance Analysis of Mass spectrum based protein quantification ..116

6.4 Experimental Design 117

6.5 pSAM removes selection biases on protein abundance or size 120

6.6 Protein markers selected by pSAM are more functionally correlated and coherent with gene expression changes 121

6.7 HIP1R and CD74 are promising novel protein markers of CLL progression risk. 122

6.8 Differential protein expression is coherent with protein interactions 124

6.9 Integrated strategies for targeted proteomic 125

6.10 Acknowledgement 126

7. CONCLUSION 136

7.1 Pathway-based molecular diagnosis 136

7.2 Protein biomarker identification 140

REFERENCES 143

LIST OF FIGURES

FIGURE 1.1. OVERVIEW OF THE TWO HIGH-THROUGHPUT TECHNIQUES FOR PROTEIN-PROTEIN INTERACTIONS 14

FIGURE 1.2. GRAPHICAL USER INTERFACE OF CYTOSCAPE.....15

FIGURE 2.1. IDENTIFICATION OF PROTEIN SUBNETWORK MARKERS AND PROGNOSIS OF DISEASE SEVERITY 21

FIGURE 2.2. SCHEMATIC OVERVIEW OF SUBNETWORK IDENTIFICATION....22

FIGURE 2.3. COMBINED ACTIVITY CAPTURES TWO GENES OF HETEROGENEOUS EXPRESSION.....23

FIGURE 3.1. SCHEMATIC OVERVIEW OF SUBNETWORK IDENTIFICATION....39

FIGURE 3.2. MUTUAL INFORMATION BETWEEN ACTIVITIES AND CLASS LABELS 40

FIGURE 3.3. SUBNETWORK ENRICHED FOR THE HALLMARKS OF CANCER .41

FIGURE 3.4. MARKER REPRODUCIBILITY AND METASTASIS PREDICTION PERFORMANCE.....42

FIGURE 3.5. CLASSIFICATION PERFORMANCE BY USING SVM.....43

FIGURE 3.6. SENSITIVITY AND SPECIFICITY OF CLASSIFIERS USING SUBNETWORK MARKERS OR SINGLE-GENE MARKERS IN FIGURE 3.4B.....44

FIGURE 3.7. SUBNETWORK MARKERS CONTAINING HER-2/NEU (ERBB2), MYC, OR CYCLIN D1 (CCND1) 45

FIGURE 3.8. DETECTION OF 60 KNOWN DISEASE GENES IN BREAST CANCER 46

FIGURE 3.9. DETECTION OF 71 GENES WITH SOMATIC MUTATIONS ASSOCIATED WITH BREAST CANCER IN SJOBLOM et al. 47

FIGURE 4.1. DISPARITY BETWEEN THE DATE OF DIAGNOSIS (DX) AND THE DATE OF TUMOR SAMPLE COLLECTION (SC) FOR PATIENT STRATIFICATION 71

FIGURE 4.2. SURVIVAL ANALYSIS ON DX→TX OF THE UCSD PATIENTS WITH REGARD TO THE TWO RISK GROUPS DEFINED BY IGHV MUTATION STATUS 72

FIGURE 4.3. SCHEMATIC OVERVIEW OF SUBNETWORK IDENTIFICATION AND DEFINITION OF RISK GROUPS 73

FIGURE 4.4. EXAMPLE SUBNETWORKS OF CLL DISEASE PROGRESSION ENRICHED FOR THE HALLMARKS OF CANCER 74
FIGURE 4.5. SUBNETWORK SIGNATURES OF CLL DISEASE PROGRESSION .. 76
FIGURE 4.6. PROGNOSIS OF NEW PATIENTS 77
FIGURE 4.7. SURVIVAL ANALYSIS ON SC→TX OF THE EUROPEAN COHORT WITH REGARD TO THE TWO RISK GROUPS DEFINED BY IGHV MUTATION STATUS 78
FIGURE 4.8. SERIAL EXPRESSION OF EXAMPLE SUBNETWORK GENES AND THE SUBNETWORK SIGNATURE ALONG DISEASE PROGRESSION 79
FIGURE 4.9. SERIAL PROTEIN EXPRESSION OF EXAMPLE SUBNETWORK GENES DURING DISEASE PROGRESSION 81
FIGURE 4.10. CANCER GENE ENRICHMENT IN EACH MARKER SET 82
FIGURE 4.11. PREDOMINANT CELLULAR FUNCTIONS ASSOCIATED WITH THE 38 SUBNETWORKS 82
FIGURE 5.1. A SCHEMATIC DIAGRAM OF KEY GENE IDENTIFICATION AND ACTIVITY INFERENCE 98
FIGURE 5.2. MARKER REPRODUCIBILITY OF PATHWAY-BASED AND GENE-BASED SELECTION 99
FIGURE 5.3. DISCRIMINATIVE POWER OF PATHWAY AND GENE MARKERS IN THE BREAST AND LUNG CANCER DATASETS 100
FIGURE 5.4. CLASSIFICATION ACCURACY (A) WITHIN- AND (B) ACROSS-DATASETS 101
FIGURE 5.5. CLASSIFICATION ACCURACY WITHIN- AND ACROSS-DATASETS USING DIFFERENT CLASSIFIER 102
FIGURE 5.6. CLASSIFICATION ACCURACY WITHIN- AND ACROSS-DATASETS USING SEQUENTIAL SELECTION (SEQ) OR FORWARD SELECTION (FWD) 103
FIGURE 5.7. PATHWAY ACTIVITY OF THE TOP FREQUENTLY-USED MARKERS IN THE TWO BREAST CANCER DATASETS 104
FIGURE 5.8. PATHWAY ACTIVITY OF THE TOP FREQUENTLY-USED MARKERS IN THE TWO LUNG CANCER DATASETS 105
FIGURE 5.9. DISTRIBUTION OF NUMBERS OF CORGS IN TOP 10% PATHWAYS 106

FIGURE 6.1. A SCHEMATIC DIAGRAM OF A MS/MS EXPERIMENT USING ITRAQ AND PSAM 127
FIGURE 6.2. EXPERIMENTAL DESIGNS OF MS/MS EXPERIMENTS 128
FIGURE 6.3. PROTEIN SELECTION BIAS WHEN USING SIMPLE RATIOS OR PSAM 129
FIGURE 6.4. COMPARISONS OF TOP PROTEINS SELECTED BY MS1 AND MS2 QUANTIFICATION 130
FIGURE 6.5. DIFFERENTIALLY EXPRESSED PROTEINS BETWEEN THE THREE CLASSES OF MATURE B-CELLS 131
FIGURE 6.6. HIP1R PROTEIN EXPRESSION IN NEWLY DIAGNOSED PATIENTS 132
FIGURE 6.7. CD74 PROTEIN EXPRESSION IN NEWLY DIAGNOSED PATIENTS 133
FIGURE 6.8. CORRELATION BETWEEN PROTEIN DIFFERENTIAL EXPRESSION AND PHYSICAL INTERACTIONS 134
FIGURE 6.9. STRATEGIES FOR TARGETED PROTEOMICS 135

LIST OF TABLES

TABLE 3.1. CLASSIFICATION ACCURACIES OF THE PREVIOUS GENE SIGNATURES IN THEIR ORIGINAL STUDIES 48

TABLE 3.2. LIST OF 60 BREAST CANCER SUSCEPTIBILITY GENES 49

TABLE 5.1. THE SEVEN DATA SETS USED IN METHOD EVALUATION..... 107

TABLE 5.2. FREQUENTLY SELECTED PATHWAY MARKERS FOR BREAST CANCER PROGNOSIS 108

TABLE 5.3. FREQUENTLY SELECTED PATHWAY MARKERS FOR LUNG CANCER PROGNOSIS 109

ACKNOWLEDGEMENT

Flying over the great Pacific Ocean, life in the graduate school on another continent has been quite a journey for me. Luckily, I am not an island. With the help and company from many others, this journey has been joyful and fruitful. I would like to, first and foremost, acknowledge the support of my advisor, Dr. Trey Ideker. None of this work would have been possible without his guidance and inspiration through countless discussions and many long nights. Dr. Ideker always brought in boundless energy and brilliant thinking to stimulate fresh and insightful views on research problems. His enthusiasm and dedication to science encouraged me to tackle the obstacles in research once more.

I would also like to thank my co-advisor, Dr. Thomas Kipps, without whom I would not have been able to think anywhere near a cancer researcher. Even in a day full of intense clinical duty, Dr. Kipps would always find out time and energy to discuss additional angles and possibilities in my research. His lucid reasoning and work ethic have proved to be invaluable.

I also want to thank all of the members of my committee for the time and support they have provided during various committee and individual meetings: Dr. Steve Briggs, Dr. Sanjoy Dasgupta, and Dr. Jean Wang. I would especially like to thank Dr. Steve Briggs for his contribution. His valuable insight and sense of humor encouraged me to explore unknowns, not only in research but also in my career development.

I am also very thankful to my many co-authors for their valuable scientific work in my dissertation. More than being essential to the development of algorithmic and

statistical techniques vital tothe study, Dr. Eunjung Lee has also been my best friend during my graduate school and sharing with me both excitement and disappointment by the research puzzles. Dr. Laura Rassenti is another best friend who provided voluminous amounts of data and biological insight to ground the study in chronic lymphocytic leukemia. I am also grateful to Dr. Zhouxin Shen for his expertise in mass spectrometry, Michelle Salcedo for her help in protein expression experiments and Kate Licon for her help in gene expression measurements. Although not a co-author, Samad Lotia implemented my algorithms into software packages. Outside my dissertation, I have also worked with various bright collaborators and own them a great debt for broaden my knowledge in science: Dr. Kiyoung Lee, Dr. Weizhou Zhang, Dr. Arnon Kater, Dr. Liguang Chen, and Dr. Suping Zhang and Matan Hofree.

I would also like to thank all my colleagues in the Ideker lab and the Kipps lab for suggestion and help on my research during weekly lab meetings and daily bench work. In particular, I want to thank Stephanie Mirkin and Carolina Bump for their help on coordinating discussion sessions on my research project. I also want to thank Merill Gersten, Dr. Ryan Kelley, Dr. Silpa Suthram, Dr. Sourav Bandyopadhyay, Dwight Kuo, Menzies Chen, Gregory Hannum, Rohith Srivas, Rob DeConde, and Dr. Bing Cui for their friendship.

I am grateful to all my friends outside the labs, back in Taiwan or here in USA, for all the happy times they gave me. Specially, I would like to thank Bo-Juen Chen for always be there for my insaneness and complains. I also want to thank Dr. Pei-Jen Lee and Yin Wang for their help during the very last moment in my graduate study.

Lastly, I want to thank all my family for their endless support. I was born to be lucky because of the surrounding love. My grandparent, Yu-Fu and Yue-E, inspired me to be a scientist. My grandfather bought me my first comic book which was an intriguing fiction in medicine. My parents, Fu-Jui and Pao-Huei, taught me the appreciation for education and knowledge and support me to pursue and live all my dreams. Finally, I thank my partner Irene for always standing by me and believing in me more than I can ever do. Oh, also for my cute dogs, Julie and Phoeny, I have to thank them for eating my face whenever I was stressed out. A man without knowledge is not a free man. I am blessed that I am one step toward being free now.

Chapter 2, in part, quotes from the materials published in Chuang, HY, Lee, E, Liu, YT, Lee, D, and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol.* 3:140 (2007). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a re-editing of the materials published in Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol.* 3:140 (2007). The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a re-editing of the materials submitted for publication. Chuang, H.Y., Rassenti, L., Salcedo, M., Licon, K., Ideker, T., Kipps, T. Network-based analysis of Chronic Lymphocytic Leukemia identifies pathways that contribute to disease evolution, **submitted**. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a re-editing of the materials published in Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T., Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Compu Biol.* 4(11): e1000217 (2008). The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a re-editing of the materials currently being prepared for submission for publication. Chuang, H.Y., Shen, Z., Rassenti, L., Ideker, T., Kipps, T., Briggs, S. Dissecting disease progression of Chronic Lymphocytic Leukemia using an integrated quantitative proteomic and genomic analysis, **in preparation**. The dissertation author was the primary investigator and author of this paper.

VITA

2001	Bachelor of Science	National Taiwan University
2003	Master of Science	National Taiwan University
2010	Doctor of Philosophy	University of California, San Diego

PUBLICATIONS

Chuang, H. -Y., Shen, Z., Ressenti, L., Ideker, T., Briggs, S., and Kipps, T., Dissecting disease progression of Chronic Lymphocytic Leukemia using an integrated quantitative proteomic and genomic approach, **in preparation**.

Chuang, H. -Y., Hofree, M., Ideker, T., A decade in Systems Biology, **submitted**.

Chuang, H. -Y., Ressenti, L., Ideker, T. and Kipps, T., Network-based analysis of genes expressed in Chronic Lymphocytic Leukemia identifies pathways that contribute to disease evolution, **submitted**.

Lee, P., **Chuang, H. -Y.***, Shen, Z. and Briggs, S., Proteome-wide analysis reveals *Sall4* to be a major regulator of self-renewal in embryonic stem cells, **in preparation**. **co-first authorship*

Lee, K., Byun, K.H., **Chuang, H. -Y.**, Mak, C., Paek, S.H., Lee, B., and Ideker, T., Proteome-wide prediction of protein translocation identifies *PSPN* complex as essential for glioma progression, **submitted**.

Lee, K., Sung, M.K., **Chuang, H. -Y.**, Huh, W.K., and Ideker, T., Dynamic protein translocation under stressful conditions, **in preparation**.

Zhan, S., Chen, L., Cui, B., **Chuang, H. -Y.**, Yu, J., Tang, L. and Kipps, T., Broad Expression In Human Neoplasia Of ROR1 - An Oncofetal Receptor For *Wnt5a* Involved In Tumor Growth and Progression, **submitted**.

Zhang, W., Kater, A. P., Widhopf, G. F., **Chuang, H. -Y.**, Enzler, T., Danelle, J., Herschman, H., Kipps, T., and Karin, M., B Cell Activating Factor And *c-MYC* Regulate Progression Of B Cell Chronic Lymphocytic Leukemia, **submitted**.

Enzler, T., Kater, A. P., Zhang, W., Widhopf, G. F., **Chuang, H. -Y.**, Lee, J., Avery, E., Croce, C. M., Karin, M. and Kipps, T. (2009) *BAFF* Promotes *TCL1*-induced B-cell leukemogenesis by enhancing neoplastic cell survival, *Blood*, **114(20)**: 4469-4476.

Lee, A., **Chuang, H. -Y.***, Ideker, T. and Lee, D. (2008) Inferring pathway activity toward precise disease classification, *PLoS Comp. Biol.*, **4(11)**: e1000217. **co-first authorship*

Lee, K., **Chuang, H. -Y.**, Beyer, A., Sung, M.K., Huh, W.K., Lee, B., and Ideker, T. (2008) Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species, *Nucleic Acids Research*, **36(20)**: e136.

Chuang, H. -Y., Lee, A., Liu, Y. -T., Lee, D., and Ideker, T. (2007) Network-based classification of breast cancer metastasis, *Nature Mol. Syst. Biol.* **3**:140. *Highlighted by Nature "News and Views" and other scientific news media.*

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., **Chuang, H. -Y.**, et al., (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucl. Acids Res.* **33**: 5691-5702.

Chuang, H. -Y., Liu, Y. -T., Rassenti, L. Z., Huynh L., Carson, D. and Kipps, T. J. (2005) Expression of lymphocyte activation gene 3 (*LAG-3*/*CD223*) by chronic lymphocytic leukemia B cells, *Blood (ASH Annual Meeting Abstracts)*, **106**: 2952.

Liu, Y. -T., Rassenti, L. Z., Shen, Z., **Chuang, H. -Y.**, Briggs, S. P., Kipps, T. J., and Carson D. (2005) Differential expression profile of the proteome and transcriptome in aggressive and indolent chronic lymphocytic leukemia, *Blood (ASH Annual Meeting Abstracts)*, **106**: 2101.

ABSTRACT OF THE DISSERTATION

PATHWAY-BASED MODELING AND DIAGNOSIS OF CANCER DEVELOPMENT AND PROGRESSION

by

Han-Yu Chuang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego 2010

Professor Trey Ideker, Chair
Professor Thomas Kipps, Co-chair

The advent of whole-genome expression profiling technology has made it possible to identify transcriptional dysregulation that contribute to or result from disease mechanisms and can also serve as biomarkers for disease. However, expression-alone classification can be challenging in complex diseases due to factors such as genetic

heterogeneity across patients or noise in mRNA levels. Moreover, it remains unclear how these marker genes interrelate within a larger functional network.

We propose a novel approach to integrate gene expression with protein interactions to dissect cancer development and outcome. The new prognostic markers are not individual genes or proteins, but as sets of coherently expressed genes whose products interact within a larger human protein interaction network. In breast cancer, we show that this integrated strategy predict the risk of metastasis potential more accurately than previous approaches based only on gene expression. More than being more reproducible and robust, our network markers also give molecular models for how the cancer susceptibility genes might be associated with cancer metastasis.

We next apply this network-based analysis to develop a new system for accurately stratifying patients of chronic lymphocytic leukemia (CLL) at different risk levels of disease progression. The network markers represent an array of disease pathways whose expression converge over time among patients regardless their initial risk levels, implicating novel understanding for cancer evolution and for the development of treatment strategies. Besides incorporating protein interaction network into gene expression analyses, we also identify condition-responsive genes within canonical pathways to infer dysfunctional pathway activation. Contrast to methods based on static pathway knowledge, our dynamic pathway markers lead to better clinical performance for various cancers, including leukemia, prostate cancer, breast cancer and lung cancer.

Another way to address the difficulties seen in gene expression studies is to obtain direct measurement of protein levels and states by quantitative mass spectrometry.

We develop a method to select protein markers based on the change in expression relative to the standard deviation of repeated measurements across experimental replicates. In CLL disease progression, our protein markers are shown to be involved in the same pathways and more prognostic of newly diagnosed patients. We further discuss strategies for targeted proteomic profiling with the guidance of protein interaction networks.

1. INTRODUCTION

Mapping the molecular mechanisms that drive neoplasm transformation and progression is critical to our understanding and treatment of cancer. Over the past couple decades, enormous progress has revealed cancer to be a disease involving dynamic changes in the whole genome and disruptions of various cellular processes[1]. Cancer cells of the same phenotype can exhibit a great range of genetic variability. This complex and heterogeneous nature implies a manifestation of alterations more than only on genes or cells. Instead, several lines of evidence indicate that distinct cellular pathways and microenvironmental factors collectively dictate malignant growth.

In glioblastoma multiforme (GBM), for instance, nearly all the tumor samples in the cancer genome atlas GBM project[2] harbored at least one genetic event in the ***RTK***/***RAS***/***PI3K*** oncogenic signaling pathway. However, individual tumors exhibit diverse mechanisms for alteration of the pathway – mutation or homozygous deletion of the suppressor genes ***PTEN*** and ***NF1***, amplification or mutation of the upstream receptor genes ***EGFR***, ***ERBB2***, ***PDGFRA*** and ***MET***, or mutation of the core genes ***RAS*** and ***PI3K***. Studies in other cancer types[3, 4] also suggested that different combinations of genetic alterations can incapacitate each trait that cancer cells must acquire toward their route to malignant transformation. Furthermore, several oncogenic lesions in leukemia[5] had been shown to work cooperatively to drive the cell to tumorigenesis. The interactions among molecular alterations that can give rise to cancer provoke a number of questions. Can the large and diverse collection of cancer-associated genes be tied to the operations of a small group of regulatory pathways? What types of cellular regulatory

pathways within a target cell must be disrupted in order to drive neoplasms? What ways can the distinct regulatory pathways cooperate to direct the pathologic transition? Is there a common set of dysregulated cellular pathways contributing to the disparate neoplasms arising in the human body? To understand the structure of such mechanisms, it is helpful to dissect malignant transformation and progression based on a holistic view of biological systems.

Until recently, cancer research has been necessarily conducted in a reductionist manner to focus on a specific gene or signaling pathway, limited primarily by the lack of technology and tools needed to interrogate at a higher level for interplay across these single components. Enabled by the exponential development in high-throughput technologies within the past decade, Systems Biology provides a framework for assembling models of biological pathways from systematic measurements[6, 7]. The extensive genomic, proteomic and other global cell measurements include, among many others, genomic sequences[8, 9], gene expression and genotypic profiles[10], DNA-binding profiles from chromatin immunoprecipitation[11], and protein abundance from mass spectrometry[12]. It is impossible to study a biological system as a whole without these systematic data. On the other hand, it is also impossible to perform hypothesis-driven science on genome-wide measurements without the advance in computational analyses of the vast amount of data[13]. Together, these developments in both experimental and computational methods have afforded a profound opportunity to characterize the differences between cancer cells and their normal counterpart through integrative systems approaches[5, 14-20].

In addition to deciphering disease mechanisms, another crucial question in cancer research is to develop tools to diagnose cancer more accurately. More than 100 distinct types of cancer have been documented, and subtypes of tumors can be found within specific organs[1]. Precise classification and prognosis is critical for effective treatment plans, given that patients with the same type of cancer frequently respond very differently to the same treatment. The advent of DNA microarrays has surged gene expression profiling to become the method of choice for identifying diagnostic biomarkers able to diagnose the severity of disease and predict future disease outcomes. Markers are selected by scoring each individual gene for how well its expression pattern can discriminate between different classes of disease or between cases and controls. The disease status of new patients is predicted using classifiers tuned to the expression levels of the markers. Recent applications of gene expression profiling on molecular characterization and diagnosis of cancer is concisely summarized in **Section 1.1**.

Despite their promise, gene-expression-based diagnostics continue to face serious challenges due to their questionable accuracy for predicting patient outcomes in some diseases[21, 22]. Moreover, it is a common phenomenon that for the same disease different research groups each identified sets of gene markers but fail to observe an acceptable overlap between these gene signatures. For example in breast cancer, two large-scale expression studies by van't Veer *et al.*[15] and Wang *et al.*[23] each identified a set of ~70 gene markers that were of equivalent prognosis power, rivaling the performance of established criteria. Strangely, however, these marker sets shared only three genes in common. Furthermore, it is usually hard to explain functional relationships between those marker genes.

Problems are thought to arise due to at least two factors: cellular heterogeneity within tissues and genetic heterogeneity across patients. The impact of cellular heterogeneity depends on the nature of the disease: for some diseases, such as B-cell lymphoma, the diseased cell population is very well defined such that it is possible to harvest a relatively pure cell population yielding a distinct expression signature. In other diseases, such as breast cancer, it has been very difficult to cleanly separate tumor from normal cells, such that the resulting expression profile represents an average signal diluted over a mixed cell population.

In contrast to cellular heterogeneity, genetic heterogeneity refers to the fact that the same genes may not be dysregulated in each patient. For instance, patient A may have gene A dysregulated, patient B may have gene B dysregulated, patient C may have gene C dysregulated, and so on. Given this disparity across patients who nevertheless may have the same clinical outcomes (e.g., aggressive cancer), classification algorithms have trouble because there is no single marker that is indicative of the status of all (or even most) patients.

To address these problems and improve on gene-expression-based diagnostics, we and several groups are beginning to integrate patient expression profiles with system-wide maps of the pathways in the cell[24-40]. The rationale for including pathway information is that it provides an overarching layer of organization which can tie seemingly disparate expression responses together into a common pattern. For instance, although any gene A, B, or C may indicate an aggressive form of disease, if we are given the knowledge that the protein products of genes A, B, and C form a coherent module—

e.g., subunits of a common protein complex, successive enzymes in a metabolic pathway, or successive steps in a signal transduction cascade—this allows us to formulate new biomarker functions that take all of these proteins into account. Some approaches based on known pathway knowledge are introduced in **Section 1.2**.

Intuitively, one's ability to classify disease states should be improved by introducing relevant pathway information. However, a remaining hurdle to pathway-based analysis is that the majority of human genes have not yet been assigned to a definitive pathway. The recent availability of large protein networks provides one means to at least partially address these challenges. Using protein–protein interaction networks derived from literature, the yeast two-hybrid system, or mass spectrometry, a number of approaches have been demonstrated for extracting relevant subnetworks based on coherent expression patterns of their genes[41, 42] or on conservation of subnetworks across multiple species[43]. Each subnetwork is suggestive of a distinct functional pathway or complex, yielding many known and novel pathway hypotheses in organisms for which sufficient protein interaction data have been measured. Large protein networks have only recently become available for human[44-47], enabling new opportunities for elucidating pathways involved in major diseases and pathologies[24]. A brief review on the advances of high-throughput technologies in large-scale discovery of protein interactions within a cell is given in **Section 1.3**.

In this dissertation, we first pursue a protein-network-based approach for identifying cancer pathway markers within gene expression profiles, which can be used to identify genetic alterations, to assess progression risk and predict the treatment need in

unknown samples. The markers in question are not encoded as individual genes or proteins, but as subnetworks of interacting proteins within a larger human protein–protein interaction network. In **Chapter 2**, we describe a novel algorithm which uses a genome-wide protein interaction network as a systematic framework for the identification of protein complexes or signaling cascades playing a role in cancer formation and further progression. Beyond charting molecular mechanisms underlying disease, we discuss how the identified molecular maps can be used to develop better diagnostic tools when different subtypes of cancer are known in advance in **Chapter 3** or to identify patient subgroups of different risk profiles in **Chapter 4**. In **Chapter 3**, we identify protein network markers of breast cancer metastasis and show that such a network-centric method has several advantages over previous analyses of differential expression on identification of susceptibility genes and prediction of metastatic likelihood in unknown samples. We then integrate the patient survival data into the network-based approach to identify clinically relevant cancer subtypes in chronic lymphocytic leukemia (CLL) which has a very heterogeneous clinical course. In **Chapter 4**, we show that the identified CLL subtypes and corresponding network markers can reliably predict the relative risk for disease progression. From the profiles of longitudinal tumor samples, we find convergence in expression of these networks over time, regardless of the initial risk category at diagnosis. This suggests that degenerate pathways may converge into common pathways that govern disease progression. Besides extracting pathway information from unbiased protein interaction networks, we present another method for pathway marker selection that incorporates static literature-curated pathways in a condition-specific manner. In **Chapter 5**, we demonstrate that our dynamic-pathway

based approach outperforms previous analyses of differential expression in classifying samples across seven different cancer datasets, including lung cancer, prostate cancer, breast cancer and acute leukemia. In all, we demonstrate that effectively incorporating pathway information into expression-based disease diagnosis can provide better discriminative and more biologically defensible models.

Another reason for the discrepancy seen in gene expression studies is that mRNA measurements are often noisy and do not necessarily correlate with the activity of the corresponding protein. A solution to this difficulty is to obtain direct measurements of protein levels and states. The revolutionized generation of large-scale proteomic tools that are based on chromatography and mass spectrometry enable the identification of proteins and simultaneous measurement of their abundances, and can also identify when they have complex secondary modifications. In **Chapter 6**, we present a method for quantitative measurements of protein expression using mass spectrometry.

Conclusion on the findings and implication of both pathway-based diagnostics and protein expression profiling is given in **Chapter 7**. We also discuss possible improvements on the proposed methods as well as potential directions for method extension and applications.

1.1 Gene expression analyses in cancer systems biology

With the help of high-throughput technologies, we now have vast amounts of data providing a global view of molecular events contributing to or associated with oncogenesis. In the past decade, DNA microarrays, in particular, have made significant contribution to cancer research by generating global quantitative profiles of gene

expression in cancer in hundreds of large-scale experiments. By comparing to the expression profiles from corresponding normal tissues, most tumors show a unique and recognizable expression pattern, i.e., one set of differentially expressed genes per tumor type. Besides the "cancer vs. normal" studies, another common experimental design is to compare cancer samples based on their degree of progression, as determined by histological grade, invasiveness, or metastatic potential. Known types and subtypes of cancer have been readily distinguished by their gene-expression patterns, and also new molecular subtypes of cancer have been discovered. For many types of cancer, such "gene signatures" make it possible to develop expression-based classifications to diagnose the severity of disease and predict future disease outcomes.

An increasing number of diagnostic markers of various disease states, outcomes, or responses to treatment have been identified through analysis of such genome-wide expression profiles (reviewed by Chung *et al.*[48], Asyali *et al.*[49], Quackenbush *et al.*[50], and Cheang *et al.*[51]). Marker genes are selected by scoring how well their expression levels can discriminate between different classes of disease. This method has achieved >90% accuracy for some leukemias such as acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)[14], even outperforming the conventional clinical risk factors.

As many different tumors have been profiled for system-specific studies, scientists started to examine whether any patterns of gene expression are common among diverse tumor types. One way is to seek for genes showing a statistically significant difference in expression level between tumor profiles which have the feature of interest

and tumor profiles which do not have that feature[52, 53]. For example, Ramaswamy *et al*.[53] analyzed a set of expression profiles from metastatic adenocarcinoma of distinct origin and compared this set with a set of primary adenocarcinomas representing the same spectrum of tumor types. Another idea is to investigate overlaps of the individual gene signatures from different profiles. Rhodes *et al*.[54] assessed the intersection of multiple "cancer vs. normal" gene signatures from a diverse collection of microarray datasets to identify the genes that play a critical role in the neoplastic phenotype. Tomlins *et al*.[55] defined several gene signatures related to prostate cancer progression by comparing any two stages. They linked each pair-wise comparison to another microarray studies for prostate cancer if the two gene signatures shared significant proportion of differentially expressed genes. Lamb *et al*.[16] extend the similar pattern-matching idea to characterize the molecular signatures arising from specific pharmacological interventions in the cell. A tumor profile would be associated with a drug if its gene signature has a significant fraction of genes overlapped with the signature of the drug-response expression profile.

In addition to the development of molecular diagnosis of cancer, these gene expression data have also been integrated with other types of information for the identification of oncogenes, tumor-suppressor genes, and even entire oncogenic pathways. A highly recurrent gene fusion event, for instance, was identified in prostate cancer from gene expression profiles using an 'outlier' analysis approach[17]. Copy-number data and gene expression profiles were successfully used in the identification of specific chromosomal amplifications in breast cancer[19]. Another example is to use reference signatures of specific activated pathways to characterize tumors and establish drug sensitivity[20]. Besides gene expression levels, DNA microarrays can also be used to

measure genomic variation across patient tumors. Genome-wide SNP profiling and array-based comparative genomic hybridization were applied to the identification of germ-line and somatic lesions in several cancers, including leukemia[5] and breast cancer[18].

1.2 Pathway-based expression analyses

One of the main difficulties in gene expression profiling studies is to interpret the relationship between identified differentially expressed genes and the phenotype of interest. Several approaches have been proposed to use external functional information for interpreting gene signatures[26, 56]. Gene Ontology[57] (GO database) is the most often used source of functional annotations. A gene signature is examined against each of the predefined sets of genes representing different functions, to determine whether any set is overrepresented in the gene signature compared with all the genes in the expression study.

The above simply-counting approach is reasonable but has some shortcomings, regarding its statistical significance, pointed out in several studies[29, 30, 37]. Moreover, such posteriori analyses may miss the pathways which involves moderate effects that are not capture by the differentially expressed genes in the signature. Furthermore, it has been shown that differentially expressed genes selected from a small number of samples can be highly variable[58, 59]. However, the sufficient number of samples can be over thousands[60]. Alternative approaches[30, 37] are to consider the distribution of pathway genes in the entire gene list generated by ranking genes according to their evidence for differential expression. The first innovative method Gene Set Enrichment Analysis (GSEA) has demonstrated that some coordinately dysfunctional processes could only be

uncovered when functionally related gene sets were examined in a priori fashion[61]. Another study[62] in diffuse large B-cell lymphoma (DLBL) used GSEA to show that the three subtypes of DLBL can be characterized by distinct biological processes.

Segal *et al*.[63] analyzed hundreds of gene sets in the context of a compendium of diverse cancer profiles, in order to address the commonalities and variations between different types of tumor. They compiled gene sets from GO database and co-expressed clusters in microarray studies. This analysis revealed that some gene sets were shared across many cancer types, whereas others were specific to cancer types or subtypes. The resulted module map suggests several hypotheses for the biological processes underlying a specific cancer types. Careful interpretation and validation of such functional linkage will be required to fully appreciate the value of the approach.

In addition to explaining gene expression differences between phenotypes, the pathway information can be used in predicting new expression profiles of unknown disease states. Some of these approaches represent pathway activity with a function summarizing the expression values of member genes[25, 28, 34, 35]; other approaches estimate probabilities of pathway activation based on the consistency of changes in gene expression[40, 64, 65]. Others have engineered normal cells to activate pre-selected oncogenic pathways, in order to determine gene signatures that can distinguish tumor characteristics[20, 66]. For example, Bild *et al*.[20] over-expressed a panel of oncogenes, one at a time, in primary cultures of human mammary epithelial cells. The goal was to link each oncogene with a distinct set of dysregulated genes. Given these links, they showed

that the expression profile of a new tumor sample could be analyzed in order to identify which oncogenes had been activated.

1.3 Protein interaction networks

Proteins regulate and mediate many of the processes in the cell. In most cases, they act in concert with other proteins as part of pathways or larger molecular assemblies called complexes. Systematic discovery of protein interactions can help us to understand all the possible reactions or catalytic steps underlying cellular behavior. In the past several decades, these physical associations were mainly discovered by small-scale methods such as co-immunoprecipitation and FRET microscopy[67]. Only a small number of protein interactions could be revealed in one experiment. Recently, high-throughput techniques like yeast two-hybrid (Y2H)[68] and tandem affinity purification coupled with mass spectrometry (TAP-MS)[69] accumulate our knowledge of protein interactions at the level of the whole proteome, resulting in the generation of a large number of protein interactions (see **Figure 1.1** and a recent review by Cusick *et al.*[70]).

In TAP-MS studies proteins are used as bait in a co-immunoprecipitation assay and the pulled down proteins are separated and identified using mass spectrometry. Y2H is a technique which is based on the functional reconstitution of an intact transcription factor that activates reporter gene expression. Both Y2H and TAP-MS have been used to generate large interaction networks for different species. Presently, genome-scale protein-protein interaction networks are available for the bacteria: *H. pylori*[71] and E. coli[72]; for the model eukaryotes: *S. cerevisiae*[69, 73-75], *C. elegans*[76], and *D. melanogaste*[77]; and finally for human[46, 47]. Recently, a combination of experimental methods have also

been proved useful in determining protein interactions in a condition- specific manner on a large scale. Bouwmeester *et al.*[78] used an integrated approach of proteomic pathway analysis via tandem affinity purification and loss functional analysis with RNA interference to identify new tumor necrosis factor pathway interactions. The tandem affinity purification method is a sensitive and selective method for reconstructing interaction maps of particular signal transduction pathways that may have therapeutic significance.

Usually in a protein interaction network, nodes denote proteins and there exists a link between two nodes if the corresponding proteins interact with each other. Large-scale protein networks can be visualized by many software tools, such as Cytoscape[79], NAViGaTOR[80], and VisANT[81]. Cytoscape is selected to be used in this thesis because of the wealth of publicly available plugins for many types of integration, visualization, and query of biological networks and other types of functional genomic data (**Figure 1.2**). It combines the ability to view and manipulate genome-sized graphs of cellular pathways and an extensible architecture such that new search tools can be added dynamically.

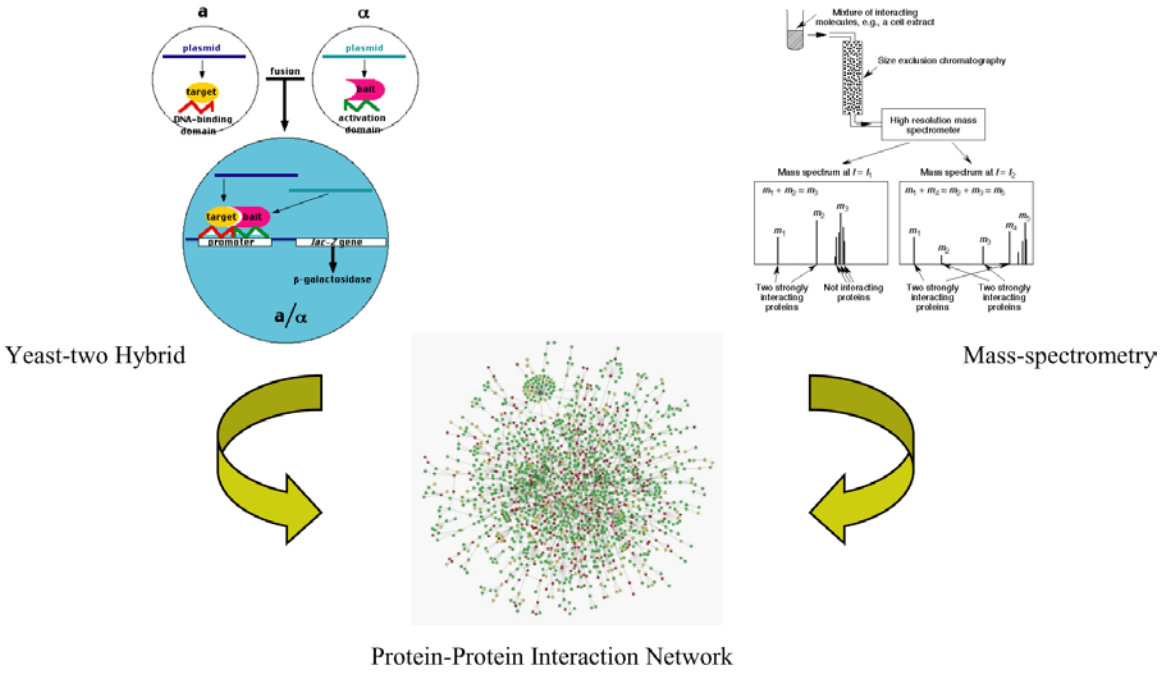


Figure 1.1. Overview of the two high-throughput techniques for protein-protein interactions

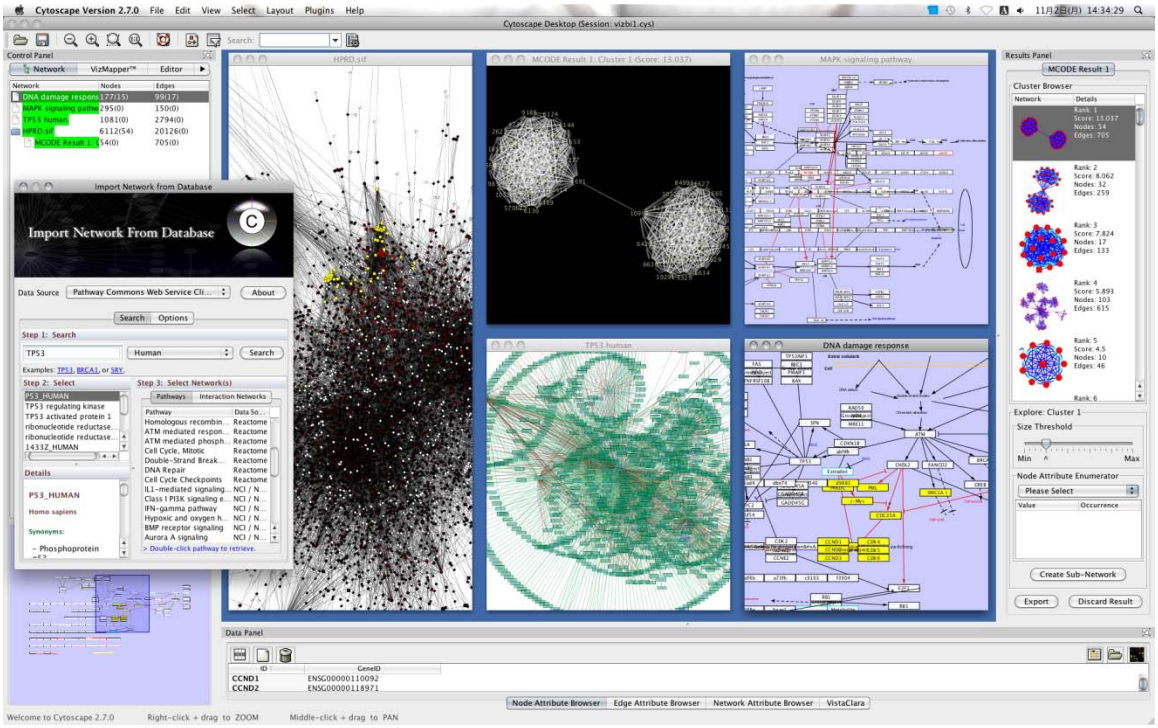


Figure 1.2. Graphical user interface of Cytoscape.

Each window showcases a different analysis or visualization of protein interaction networks and integrated data.

2. PINNACLE

− Protein InteractioN Network Aided Classification Engine

To provide a systematic and integrative framework for incorporating data of cancer disease mechanisms and outputting predictions of cancer outcome, we develop a new method to identify dysfunctional pathways or protein complexes contributing to disease phenotypes from patients' gene expression profiles with a view of protein interaction networks (**Figure 2.1**). To obtain a proteome-wide human protein interaction network, one can assemble a pooled data set of more than 50,000 interactions among more than 10,00 proteins for both protein-protein interactions and protein-DNA binding, integrated from yeast two-hybrid experiments[46, 47], predicted interactions via orthology and co-citation[45], and curation of the literature[44, 82-86]. To integrate the expression and network datasets, we overlay the expression values of each gene on its corresponding protein in the network and searched for subnetworks whose activities across the tumor samples are highly discriminative of tumor phenotypes or associated with patient survival. This process involved several scoring and search steps, as illustrated in **Figure 2.2** and described further in below.

Briefly, a candidate subnetwork was first scored to assess its activity in each sample, defined by averaging its normalized gene expression values. This step yielded an activity score per subnetwork per sample. Second, the predictive potential of a candidate subnetwork was computed based on the statistic of choice between its activity score and the clinical variable of interest. Significantly predictive subnetworks were identified by comparing their predictive potentials to those of random networks. The selected

subnetworks provide an array of small models for the underlying disease mechanisms. Meanwhile, the corresponding activity scores can be used to identify distinct signatures associated with different risk groups of patients and further enable the development of advanced prognostics for newly diagnosed patients (**Figure 2.1**).

2.1 Scoring subnetworks

A subnetwork is defined as a gene set that induces a single connected component in the protein-protein interaction network. Given a particular subnetwork *M*, let *a* represent its vector of activity scores over the tumor samples, and let *c* represent the corresponding vector of clinical variables of interest (for examples, different disease status or survival times). To derive *a*, expression values g_{ij} are normalized to z-transformed scores z_{ij} which for each gene *i* have mean μ=0 and standard deviation σ=1 over all samples *j* (**Figure 2.2**). The normalized expression value z_{ij} can be interpreted as the fold change of g_i in sample *j* compared to a virtual basal expression of the gene. The individual z_{ij} of each member gene in the subnetwork are averaged into a combined z-score which is designated the activity a_i. By summarizing the fold changes of each member in the subnetwork for a sample *j*, we hope to transform the heterogeneity of gene expression to a robust signal at pathway activity (**Figure 2.3**).

Many types of statistic could be used to score the relationship between *a* and *c*. If *c* is a discrete variable, for example two distinct tumor subtypes or drug-responsive versus non-responsive, one can use discriminative statistics such as mutual information[87], *t*-score[88] or Wilcoxon signed-rank score[89], to quantify how different the activity score *a* of a subnetwork are between the two disease status of *c* over all patients. If *c* is a

continuous variable, for example the time length of relapse from primary tumor removal, one can evaluate the association between *a* and *c* by correlation metrics such as Pearson's product-moment coefficient[90], Spearman's rank coefficient[91] or proportional hazards models[92]. In this thesis, we demonstrate the usage of three statistics, mutual information, Cox proportional hazards, and *t*-score, as the predictive score function *S*(*M*) to access the prognostic power of a subnetwork *M* on the clinical variable of interest *c* in **Chapters 3**, **4** and **5**, respectively.

2.2 Searching for significant subnetworks

Given the predictive score function *S*, a greedy search is performed to identify subnetworks within the protein interaction network for which the scores are locally maximal. Candidate subnetworks are seeded with a single protein and iteratively expanded. At each iteration, the search considers addition of a protein from the neighbors of proteins in the current subnetwork. An addition that yields the maximal score increase is adopted. After each addition, the search considers deletion of each protein from the current subnetwork (except those proteins essential to subnetwork connectivity), and deletions that yield higher score are accepted. The search stops when no addition increases the score over a specified improvement rate *r*. The parameter *r* may be chosen by users to avoid over-fitting to the expression data used.

To assess the significance of the identified subnetworks, three tests of significance are performed. For the first test, a global test, we perform the same search procedure over 100 random trials in which the expression vectors of individual genes are randomly permuted on the network. Such permutation disrupts the correlation between expression

and interaction. The *S* score of each real subnetwork is indexed on the "global" null distribution of all *S* scores of random subnetworks. The second test, a local test, indexes each *S* score on a "local" null distribution, estimated from the scores of 100 random subnetworks initialized from the same seed protein as the real subnetwork. Third, we test whether the *S* score with the true disease state is stronger than that obtained with random assignments of groups to patients[30]. For the random model, these assignments are permuted in 20,000 trials, yielding a null distribution of *S* scores for each trial; the real *S* score of each subnetwork is indexed on this null distribution. Significant subnetworks are selected which satisfy all three tests with desired $p1$, $p2$ and $p3$ (user-defined), according to the three different null distributions of *S*.

2.3 Software availability

The network-based method is implemented as a Cytoscape plugin named PinnacleZ and can be downloaded at http://chianti.ucsd.edu/cyto_web/plugins/index.php. The current version supports two types of predictive score function $S(M)$ for searching subnetworks discriminative of disease phenotypes, mutual information and *t*-score. Source codes can also be reached at http://chianti.ucsd.edu/svn/csplugins/trunk/ucsd/slotia/pinnaclez/src/pinnaclez/ for further extension. The PinnacleZ plugin has been proved to be useful in identification of susceptibility subnetworks in diseases other than cancer[93].

2.4 Acknoledgement

Chapter 2, in part, quotes from the materials published in Chuang, HY, Lee, E, Liu, YT, Lee, D, and Ideker, T. Network-based classification of breast cancer metastasis.

Mol Syst Biol. 3:140 (2007). The dissertation author was the primary investigator and author of this paper.

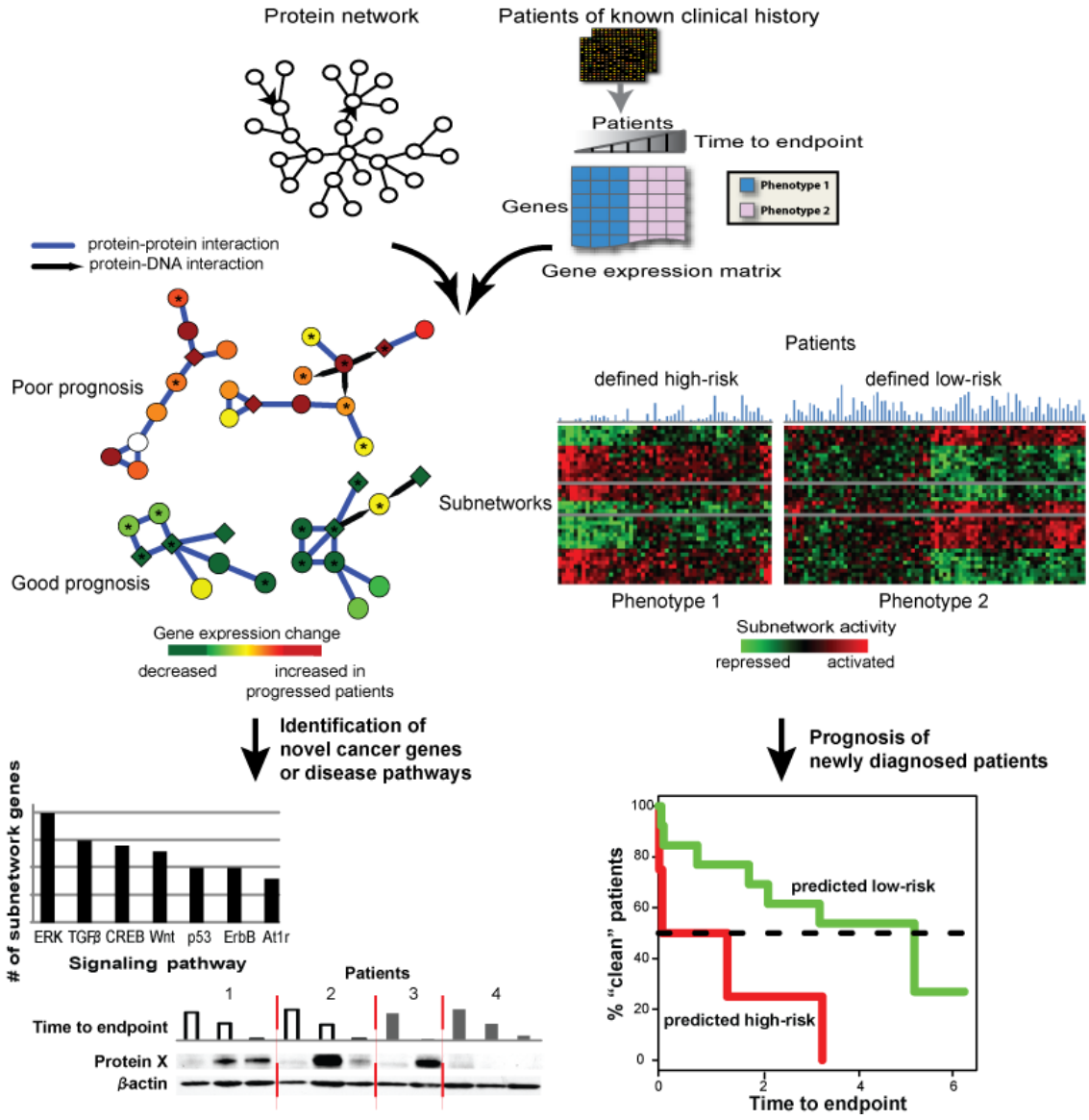


Figure 2.1. Identification of protein subnetwork markers and prognosis of disease severity.

Gene expression profiles from patient samples of known clinical history are overlaid onto a protein interaction network. Iterative exploration of all possible protein subnetworks generates a set of subnetwork markers whose activity scores across patient samples are statistically significantly associated with survival times or discriminative between disease states. The selected protein subnetworks provide potential insights into the molecular mechanisms involved in disease progression, allowing ones to discover novel disease genes or pathways. The corresponding subnetwork activity scores identify distinct activity signatures associated with different risk groups that are then used to develop prognostics for newly diagnosed patients.

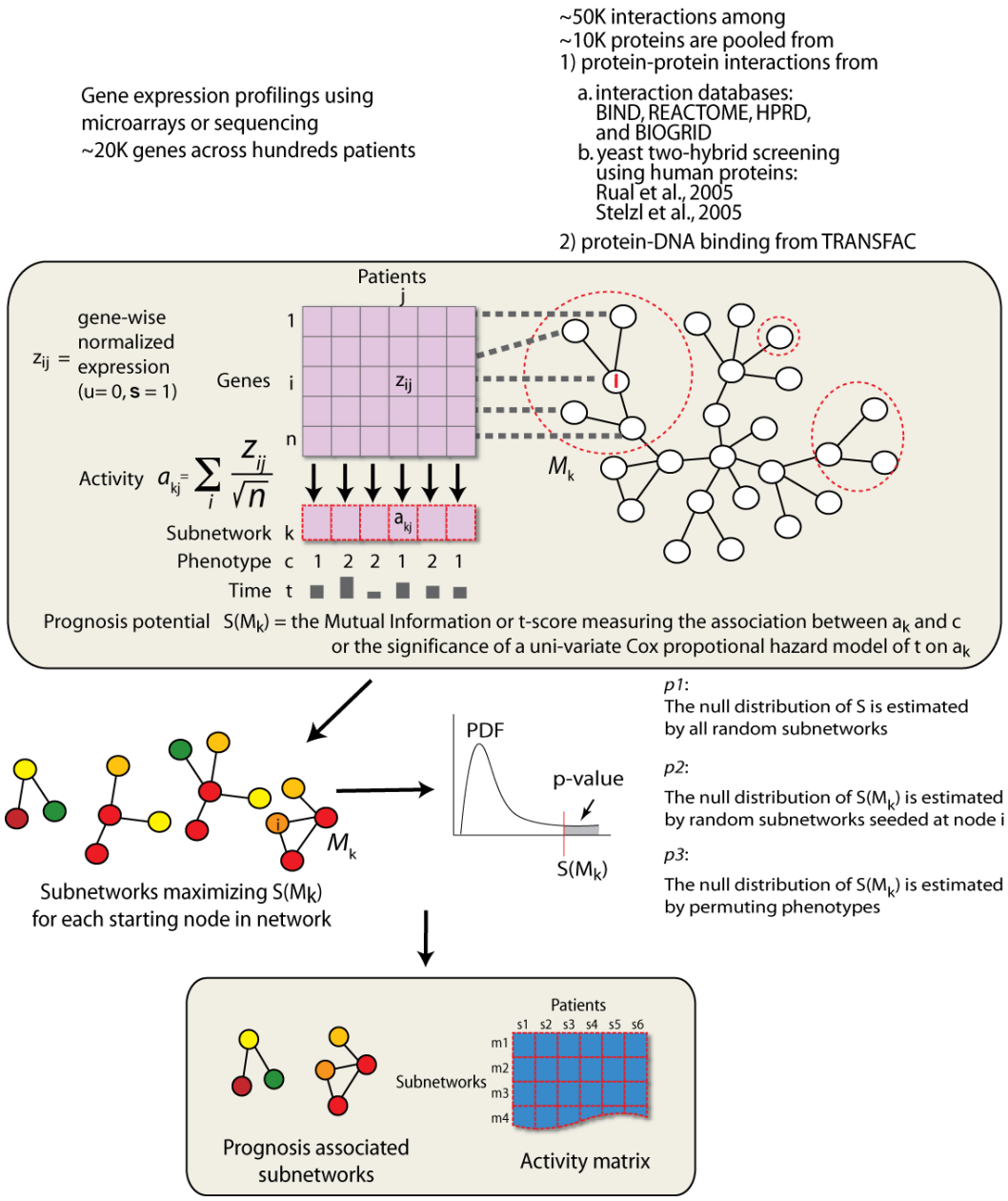


Figure 2.2. Schematic overview of subnetwork identification.

Protein interaction networks are used to assign sets of genes to discrete subnetworks. Gene expression profiles of tissue samples drawn from each type of cancer (i.e., metastatic or non-metastatic) are transformed into a "subnetwork activity matrix". For a given subnetwork M_k in the interaction network, the activity is a combined z-score derived from the expression of its individual genes. After overlaying the expression vector of each gene on its corresponding protein in the interaction network, subnetworks with discriminative activities are found via a greedy search. Significant subnetworks are selected based on null distributions estimated from permuted subnetworks (see **Section 2.2**). Subnetworks are then used to identify disease genes, and the subnetwork activity matrix is used to train a classifier for prognosis of newly diagnosed patients.

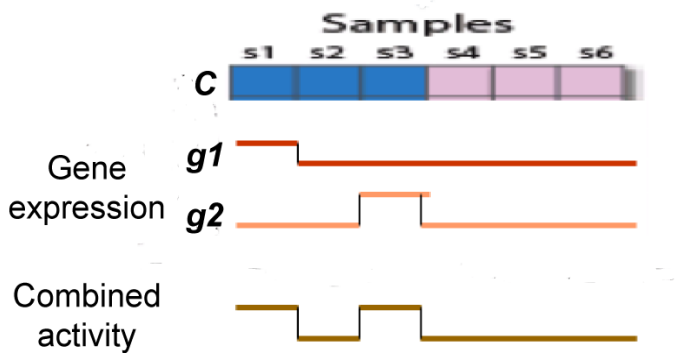


Figure 2.3. Combined activity captures two genes of heterogeneous expression.

3. NETWORK-BASED CLASSIFICATION OF BREAST CANCER METASTASIS

Mapping the pathways that give rise to metastasis is one of the key challenges of breast cancer research. Recently, several large-scale studies have shed light on this problem through analysis of gene expression profiles to identify markers correlated with metastasis. Here, we apply a protein-network-based approach that identifies markers not as individual genes but as subnetworks extracted from protein interaction databases. The resulting subnetworks provide novel hypotheses for pathways involved in tumor progression. Although genes with known breast cancer mutations are typically not detected through analysis of differential expression, they play a central role in the protein network by interconnecting many differentially-expressed genes. We find that the subnetwork markers are more reproducible than individual marker genes selected without network information, and that they achieve higher accuracy in the classification of metastatic versus non-metastatic tumors.

3.1 Background and significance

Distant metastases are the main cause of death among breast cancer patients[94]. Clinical and pathological risk factors, such as patient age, tumor size, and steroid receptor status, are commonly used to assess the likelihood of metastasis development. When metastasis is likely, aggressive adjuvant therapy can be prescribed which has led to significant decreases in breast cancer mortality rates[94]. However, for the majority of patients with intermediate-risk breast cancer, the traditional factors are not strongly predictive[23]. Accordingly, approximately 70% to 80% of lymph-node-negative patients

may undergo adjuvant chemotherapy that is in fact unnecessary[15]. Moreover, it is believed that many of the current risk factors are likely to be secondary manifestations rather than primary mechanisms of disease. An ongoing challenge is to identify new prognostic markers that are more directly related to disease and that can more accurately predict the risk of metastasis in individual patients.

In recent years, an increasing number of disease markers have been identified through analysis of genome-wide expression profiles[14, 53, 95, 96]. Marker sets are selected by scoring each individual gene for how well its expression pattern can discriminate between different classes of disease. In breast cancer, two large-scale expression studies by van't Veer *et al.*[15] and Wang *et al.*[23] each identified a set of ~70 gene markers that were 60-70% accurate for prediction of metastasis, rivaling the performance of clinical criteria[15, 23]. Strangely, however, these marker sets shared only three genes in common, with the first set of markers predicting metastasis less successfully when scoring patients from the second study, and vice versa[60]. One possible explanation for the different marker sets is that changes in expression of the relatively few genes governing metastatic potential may be subtle compared to those of the downstream effectors which may vary considerably from patient to patient[17, 21, 97].

Due to these types of difficulties, many groups have hypothesized that a more effective means of marker identification may be to combine gene expression measurements over groups of genes that fall within common pathways. Several approaches have been proposed to score known pathways by the coherency of expression changes among their member genes[26, 27, 29-31, 37, 98]. Known pathways are drawn from

sources such as the Gene Ontology[57] and KEGG[99] databases. Recently, pathway-based analysis has been extended to perform classification of expression profiles and applied to discriminate irradiated from non-irradiated yeast cells[100]. However, a remaining hurdle to pathway-based analysis is that the majority of human genes have not yet been assigned to a definitive pathway.

The recent availability of large protein networks provides one means to at least partially address these challenges. Using protein-protein interaction networks derived from literature, the yeast two-hybrid system, or mass spectrometry (reviewed in Mendelsohn *et al.*[101]), a number of approaches have been demonstrated for extracting relevant subnetworks based on coherent expression patterns of their genes[41, 42, 102] or on conservation of subnetworks across multiple species[43]. Each subnetwork is suggestive of a distinct functional pathway or complex, yielding many known and novel pathway hypotheses in organisms for which sufficient protein interaction data have been measured. Large protein networks have only recently become available for human[44-47, 103], enabling new opportunities for elucidating pathways involved in major diseases and pathologies[24].

Here, we pursue a protein-network-based approach for identifying markers of metastasis within gene expression profiles, which can be used to identify genetic alterations and to predict the likelihood of metastasis in unknown samples. The markers in question are not encoded as individual genes or proteins but as subnetworks of interacting proteins within a larger human protein-protein interaction network. We find that the network-based method has several advantages over previous analyses of

differential expression. First, the resulting subnetworks provide models of the molecular mechanisms underlying metastasis. Second, although genes with known breast cancer mutations are typically not detected through analysis of differential expression, such as P53, KRAS, HRAS, HER-2/neu and PIK3CA, they play a central role in the protein network by interconnecting many expression-responsive genes. Third, the identified subnetworks are significantly more reproducible between different breast cancer cohorts than individual marker genes selected without network information. Finally, network-based classification achieves higher accuracy in prediction, as ascertained by selecting markers from one dataset and applying them to a second independent validation dataset.

3.2 Overview of subnetwork marker identification

We applied a protein-network-based approach to analyze the expression profiles of the two cohorts of breast cancer patients previously reported by van de Vijver *et al.*[104] and Wang *et al.*[23]. Both sets of expression profiles had been obtained from primary breast tumors but hybridized to different microarray platforms (Agilent oligonucleotide Hu25K microarrays and Affymetrix HG-U133a GeneChips, respectively). We restricted our analysis to the 8,141 genes present in both datasets. For 78 patients in van de Vijver *et al.*[104] and 106 in Wang *et al.*[23], metastasis had been detected during follow-up visits within five years of surgery. Profiles for these patients were assigned to the class "Metastatic," while profiles for the remaining 217 and 180 patients were labeled "Non-metastatic." To obtain a corresponding human protein-protein interaction network, we assembled a pooled data set consisting of 57,235 interactions among 11,203 proteins,

integrated from yeast two-hybrid experiments[46, 47], predicted interactions via orthology and co-citation[45], and curation of the literature[44, 82, 84].

To integrate the expression and network data sets, we overlaid the expression values of each gene on its corresponding protein in the network and searched for subnetworks whose activities across the patients were highly discriminative of metastasis. This process involved several scoring and search steps, as illustrated in **Figure 3.1**. Briefly, a candidate subnetwork was first scored to assess its activity in each patient, defined by averaging its normalized gene expression values. This step yielded 295 and 286 activity scores per subnetwork, corresponding to the numbers of breast cancer patients in the two datasets, respectively. Second, the discriminative potential of a candidate subnetwork was computed based on the mutual information between its activity score and the Metastatic/Non-metastatic disease status over all patients (see below). Significantly discriminative subnetworks were identified by comparing their discriminative potentials to those of random networks.

Chapter 2 provides the general overview of the method framework but there are couple details specific to this breast cancer study. Given a particular subnetwork *M*, let *a* represent its vector of activity scores over the tumor samples, and let *c* represent the corresponding vector of class labels (metastatic or non-metastatic). In this study, we define the discriminative score *S*(*M*) as MI(*a′*,*c*), the mutual information MI between *a′*, a discretized form of *a*, and *c*:

$$S(\mathrm{M}) = MI(\mathrm{a}',\mathrm{c}) = \sum_{x \in \mathrm{a}'} \sum_{y \in \mathrm{c}} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

where x and y enumerate values of a and c respectively, $p(x, y)$ is the joint probability density function (pdf) of a' and c, and $p(x)$ and $p(y)$ are the marginal pdfs of a' and c. To derive a' from a, activity levels are discretized into $\lfloor \log 2(\#\text{ of samples}) + 1 \rfloor = 9$ equally-spaced bins[105]. A rationale for using MI in cancer classification is to capture potential heterogeneity of expression in cancer patients[17] i.e., differences not only in the mean but in the variance of expression. For examples of the computation of MI see **Figure 3.2**. The particular gene set maximizing $S(M)$ is regarded as optimal for classification. When assessing the significance of the identified subnetworks, for the random models we use gamma-distribution[106] to estimate the null distribution of MI in the statistical tests. Sgnificant subnetworks are selected which satisfy all three tests with $p1<0.05$, $p2<0.05$ and $p3<0.00005$, according to the three different null distributions of S.

When we search the subnetworks in the vast interaction network, at each iteration, the search considers addition of a protein from the neighbors of proteins in the current subnetwork and within a specified network distance d from the seed. Given that the median distance between any two proteins in the human protein-protein interaction network is five (i.e., the network diameter is 10), we set $d=2$ to provide a sufficient number of neighbors while keeping the search local. The improvement ratio r is chosen as 0.05 to avoid over-fitting to the expression data used. The majority of searches terminate due to the constraint on r; increasing the value of d has only marginal effect on the results.

3.3 Subnetwork markers correspond to the hallmarks of cancer

A total of 149 and 243 discriminative subnetworks were identified in the van de Vijver *et al.*[104] and Wang *et al.*[23] data sets (consisting of 618 and 906 genes, respectively, and based on a panel of three separate tests for statistical significance—see **Section 2.2**). A compendium including all of these subnetworks is available online via the CellCircuits database[107] (www.cellcircuits.org), which provides each subnetwork in both graphical (GIF) and machine-readable (SIF) formats. Each significant subnetwork may be viewed as a putative marker for breast cancer metastasis that is not based on a single gene but rather on the aggregate behavior of genes connected in a functional network. This feature is a significant departure from conventional expression-alone analysis, which does not provide functional insight into the identified markers.

In all, 47.3% (van de Vijver *et al.*[104]) and 65.4% (Wang *et al.*[23]) of the discriminative subnetworks were enriched for proteins functioning in a common Biological Process as annotated by the Gene Ontology database[57] (hypergeometric test with a False Discovery Rate of 5%). To test whether this functional enrichment might be solely due to network topology, we extracted 1000 random subnetworks of the same size as the identified discriminative subnetworks but without regard to the expression profiles. In the two sets of random subnetworks, $25.4 \pm 0.6\%$ and $26.5 \pm 0.1\%$ (mean ± stdev) were enriched for proteins with a common Biological Process. Our higher rate suggests that integrating protein networks with cancer expression profiles is able to identify proteins coordinately functioning in pathways. Among the discriminative subnetworks, 66 identified from van de Vijver *et al.*[104] and 153 identified from Wang *et al.*[23]

corresponded to signaling of cell growth and survival, cell proliferation and replication, apoptosis, cell and tissue remodeling, circulation and coagulation, or metabolism (see **Figure 3.3** for some example subnetworks; see CellCircuits database for all functional annotations). Together, these processes contribute to the major events that have been implicated in the progression of cancer[1]. Many extracellular matrix and inflammatory proteins related to tumor aggression, such as matrix metallopeptidase 9 (***MMP9*** in **Figure 3.3d**) and interleukins (**Figure 3.3h**), were also included in the identified subnetworks. Approximately 88% of the 149 subnetworks identified from van de Vijver *et al*.[104] had higher activity levels in metastatic breast tumors than in non-metastatic ones, whereas the 243 subnetworks identified from Wang *et al*.[23] were split roughly equally in their direction of activity change (124 vs. 119).

3.4 Subnetwork markers have increased reproducibility across datasets

Next, we examined the agreement between markers identified from the two breast cancer cohorts using our network-based approach. As shown in **Figure 3.4a**, the subnetwork markers were significantly more reproducible between data sets than were individual marker genes selected without network information (12.7% versus 1.3%). In terms of biological function, extracellular signal-regulated kinase 1 (***MAPK3***) was reproducible as a central node in subnetworks identified from both datasets (**Figure 3.4c** versus **Figure 3.4d**). **Figure 3.4e** and **Figure 3.4f** illustrate two other subnetworks that were discriminative in both datasets, although there was less consistency in the expression levels of genes comprising these subnetworks. For instance, ***PKMYT1*** is significantly differentially expressed in van de Vijver *et al*.[104] but not Wang *et al*.[23]

(**Figure 3.4e**; diamond versus circle), while ***CD44*** is significantly differentially expressed in Wang *et al*.[23] but not van de Vijver *et al*.[104] (**Figure 3.4f**). However, by aggregating the expression ratios of these genes with their network neighbors, the subnetworks containing these genes are found to be significant in both datasets.

One concern is that the increased overlap between subnetwork markers might be expected, given that the number of all possible subnetworks is smaller than the number of gene sets (selected irrespective of the network). However, the observed overlap between subnetworks was also significantly greater than that achieved among 1000 same-size sets of connected subnetworks chosen at random (p<0.002). Another question is why, even using subnetworks, the percentage overlap is not larger. One reason may be the difference in clinical design of the two datasets. While all of patients in Wang *et al*.[23] had lymph-node-negative breast cancer, approximately half of the patients in van de Vijver *et al*.[104] were lymph-node-positive and underwent adjuvant therapy before expression profiling. Another explanation may be the difference in microarray platforms or the incompleteness of the protein-protein interaction network, which covered only ~40% of the gene expression levels measured in either study.

3.5 Subnetwork markers increase the classification accuracy of metastasis

We next tested the predictive performance of subnetwork markers during classification of a new expression profile as Metastatic or Non-metastatic. To use the subnetworks for classification, the expression levels of the genes in each subnetwork were averaged to compute a subnetwork activity score, in the same way the activity score was computed in identifying the subnetwork markers originally (see above). These

activity scores were then used as feature values by a classifier based on logistic regression.

Logistic regression models[108] are trained on the subnetwork activity matrix (significant subnetworks versus patient samples). Subnetwork markers are selected using the whole first dataset (van de Vijver *et al.*[104]) and then tested on the second dataset (Wang *et al.*[23]; or vice versa). To measure unbiased classification performance, the patient samples in the second dataset are divided into five subsets of equal size: three subsets are used as the training set to build the classifier using markers from the first dataset; one subset is used as the validation set; and one subset is used as the test set. The p-value of discriminative power to classify training samples ($p3$) is used to rank subnetwork markers, after which the logistic regression model is built by adding markers sequentially in increasing order of *p*-value. The number of markers used in the classifier is optimized by evaluating its Area Under ROC Curve (AUC, see Swets *et al.*[109] for details) on the validation set. The final classification performance is reported as the AUC on the test set using the optimized classifier. Each of the five patient subsets in the second dataset is evaluated in turn as the test set, with the other four sets providing training and validation. The averaged AUC values among the five test sets are reported as a final classification performance.

At a fixed sensitivity of 90%, the subnetwork markers achieved 70.1% (van de Vijver *et al.*[104]) and 72.2% (Wang *et al.*[23]) accuracy, measured as the percentage of correct classifications using the technique of five-fold cross validation within each dataset. This accuracy compares favorably with those reported in the original studies[23,

[104] (62% and 63%; see **Table 3.1**). In this five-fold cross validation, one fifth of the samples were designated as "test" data and withheld during classifier training (in which the relative weights of each subnetwork feature are determined). However, the subnetwork features themselves were identified using all microarray samples prior to classification, which introduces possible circularity into the validation procedure.

To achieve an unbiased evaluation of subnetwork performance, we further tested the subnetwork markers selected from one cohort of breast cancer patients as predictors of metastasis on the other cohort. This same cross-dataset analysis was also run using individual marker genes according to the conventional method (controlled for size by providing the classifier with the set of 618 or 906 top discriminative genes in van de Vijver *et al.*[104] or Wang *et al.*[23], respectively, which is the same number of genes covered by the subnetwork markers). Similar to the procedure for the subnetwork markers, five-fold cross-validation was performed on one dataset using the genes selected from the other dataset.

At 90% sensitivity, the subnetwork markers from van de Vijver *et al.*[104] achieved 48.8% accuracy in classifying samples in Wang *et al.*[23]; 55.8% accuracy for the reciprocal test. The single-gene markers achieved 45.3% and 41.5% accuracies, respectively. Although all marker sets have decreased performance in predicting metastasis in an independent dataset, the accuracies remain significantly higher than random guesses (31.2% and 39.7%, respectively). To show that the better performance was not dependent on the chosen classification algorithm, we evaluated the markers by Support Vector Machines[110] (SVM) which led to the same trends (**Figure 3.5**).

To capture performance over the entire range of sensitivity/specificity values, we also analyzed the classifiers using the AUC metric (Area Under ROC Curve). As shown in **Figure 3.4b** and **Figure 3.6**, the subnetwork markers significantly outperformed the single-gene markers in both datasets. Subnetwork classification performance was also higher than classifiers built on random subnetworks (p = 0.046 and 0.012 against 1000 sets of same-sized random subnetworks on van de Vijver *et al.*[104] and Wang *et al.*[23], respectively); strangely, performance of the conventional classifiers was not (p = 0.124 and 0.174, respectively).

Finally, we compared the classification performance of the subnetwork markers with markers based on predefined groups of functionally-related genes (**Figure 3.4b**). These included 1,446 sets of functionally-related genes extracted from the Gene Ontology Database[57] and 522 from the Molecular Signatures Database[111] (v1.0). Neither of these functionally-related groupings performed as well as either the subnetwork markers or individual genes. This finding might indicate that some of the functional groupings relevant to breast cancer metastasis have not yet been curated in the current pathway databases.

Beyond achieving better performance, the discriminative subnetworks lend insight into the biological basis for why samples are classified as metastatic or non-metastatic. For instance, a single cell-cycle-related subnetwork was identified from Wang *et al.*[23] that could be used to predict the metastatic outcome of ~60% of patients in van de Vijver *et al.*[104] (**Figure 3.4g**). Thioredoxin (***TXN***), which was not differentially-expressed, mediated interconnections among many cell mobility and DNA replication

proteins that were differentially expressed in Wang *et al.*[23], forming subnetworks that were informative for metastasis in van de Vijver *et al.*[104] (see **Figure 3.4h** for the ***TXN*** core motif shared in multiple subnetworks). Conversely, several subnetworks identified from van de Vijver *et al.*[104], such as the ***RAD54L***-related proteasome (**Figure 3.4i**) and a ***Ras***-related subnetwork (***RAB1A*** and ***RAB11A***; **Figure 3.4j**), were predictive for patients in Wang *et al.*[23].

3.6 Subnetwork markers are informative of non-discriminative disease genes

Unlike conventional expression clustering or classification methods, network-based analyses can implicate proteins with low discriminative potential (e.g., those that are not differentially-expressed) if such proteins participate in a subnetwork whose overall activity is discriminative. Such proteins can arise within a significant subnetwork if they are essential for maintaining its integrity, i.e., they are required to interconnect many higher-scoring proteins. This property is important for discovery of disease-causing genes, because the phenotypic changes most indicative of breast cancer metastasis need not be regulated at the level of expression[112].

Overall, 85.9% and 96.7% of the significant subnetworks contained at least one protein that was not significantly differentially expressed in metastasis ($p > 0.05$ from a two-tailed *t*-test). Many well-established prognostic markers of breast cancer disease outcome, such as ***HER-2***/neu (***ERBB2***), ***Myc***, and ***cyclin D1***, were not present in gene signatures from conventional expression-alone analysis[15] but played a central role in the discriminative subnetworks by interconnecting many expression-responsive genes (see **Figures 3.3c** and **3.3j** for examples and **Figure 3.7** for all). Other examples are the

SMAD family and the phosphoinositide-3-kinase catalytic subunit (***PIK3CA***) (**Figures 3.3e-f**, **3.3i** and **3.3k**): Changes in ***SMAD*** phosphorylation have been linked to breast cancer metastasis[113], and somatic mutations in ***PIK3CA*** are associated with constitutive up-regulation of kinase activity in ~30% of breast cancers[114, 115].

To evaluate the power of a network-based method to uncover disease genes, we assembled a list of 60 breast cancer susceptibility genes that had been reported as such in previous literature and were also represented in our expression datasets[116-118] (the complete list is provided in **Table 3.2**). We found that 32 out of 149 discriminative subnetworks from van de Vijver *et al.*[104] and 27 out of 243 from Wang *et al.*[23] contained at least one known cancer susceptibility gene (7 and 5 subnetworks, respectively, contained two or more known susceptibility genes). Some notable examples are ***RAD51*** and ***TP53*** shown in **Figure 3.3a**; ***ESR1*** and ***TP53*** in **Figure 3.3b**; ***ERBB2*** in **Figure 3.3c**; ***BRCA1*** in **Figure 3.3f**; ***ESR1***, ***BRCA1*** and ***CYP1A1*** in **Figure 3.3g**; ***PIK3CA*** and ***HRAS*** in **Figure 3.3i**; ***GSTT1*** in **Figure 3.3j** and ***KRAS*** and ***PIK3CA*** in **Figure 3.3k**.

We compared these levels of enrichment to a conventional expression-alone analysis which did not incorporate information on pathway structure. As shown in **Figures 3.8a** and **3.8b**, subnetworks were significantly enriched with cancer susceptibility genes, in contrast to genes identified by a conventional analysis. Disease genes that can only be detected using network information include ***TP53***, ***KRAS***, ***HRAS***, ***ERBB2*** and ***PIK3CA***.

Finally, we also examined the enrichment of the discriminative subnetworks for a recently-published list of 122 genes with somatic mutations associated with breast

cancer[119] (71 of these were represented in the expression datasets we examined). Genes in this list were determined by DNA sequencing to have mutations in at least one of eleven breast cancer cell lines, with no cancer cell line having more than six mutant genes in common with any other cancer. A total of 11 mutations mapped to proteins found in the discriminative subnetworks (see **Figures 3.8c-e** for examples). Although still higher than the conventional method in van de Vijver *et al.*[104] (**Figure 3.9a**), this enrichment was not significant by either approach (p = 0.434 for subnetwork markers and 0.914 for single-gene markers). One explanation could be that the cancer cell lines capture a different disease state than that found in the population of patients surveyed by microarray profiling. Only two genes (***p53*** and ***BRCA1***) reported in the sequencing study were linked with breast cancer in OMIM[118], perhaps because the newly-discovered mutations are rare or not genetically transmissible.

3.7 Acknoledgement

Chapter 3, in full, is a re-editing of the materials published in Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., and Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol.* 3:140 (2007). The dissertation author was the primary investigator and author of this paper.

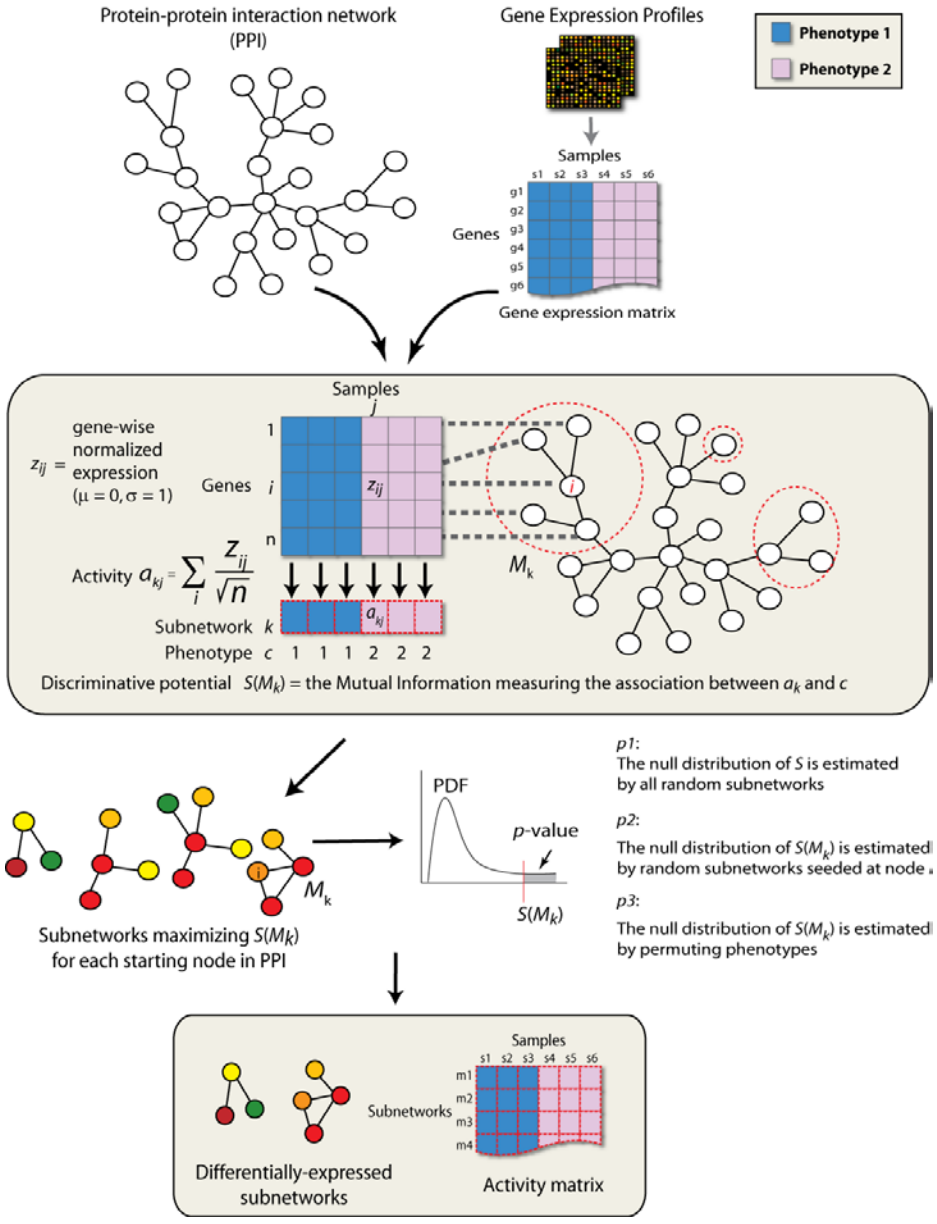


Figure 3.1. Schematic overview of subnetwork identification.

Protein-protein interaction networks are used to assign sets of genes to discrete subnetworks. Gene expression profiles of tissue samples drawn from each type of cancer (i.e., metastatic or non-metastatic) are transformed into a "subnetwork activity matrix". For a given subnetwork M_k in the interaction network, the activity is a combined z-score derived from the expression of its individual genes. After overlaying the expression vector of each gene on its corresponding protein in the interaction network, subnetworks with discriminative activities are found via a greedy search. Significant subnetworks are selected based on null distributions estimated from permuted subnetworks (see Methods). Subnetworks are then used to identify disease genes, and the subnetwork activity matrix is also used to train a classifier.

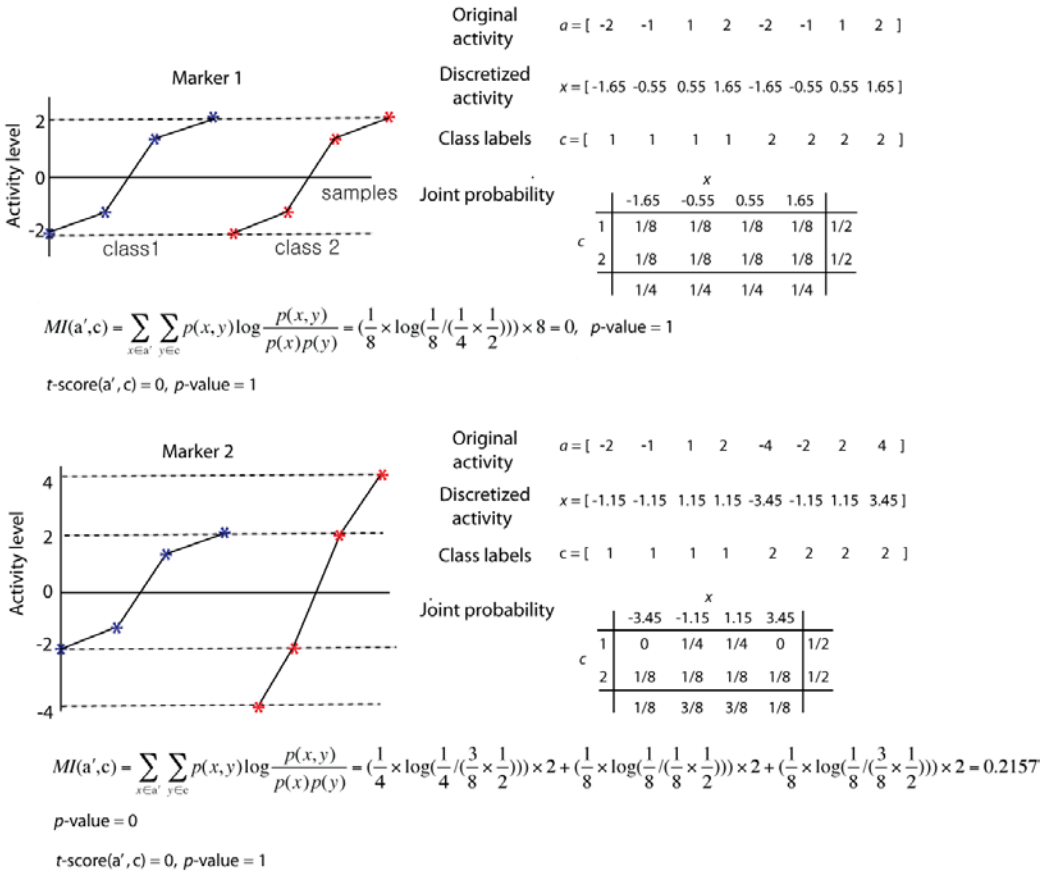


Figure 3.2. Mutual Information between activities and class labels.

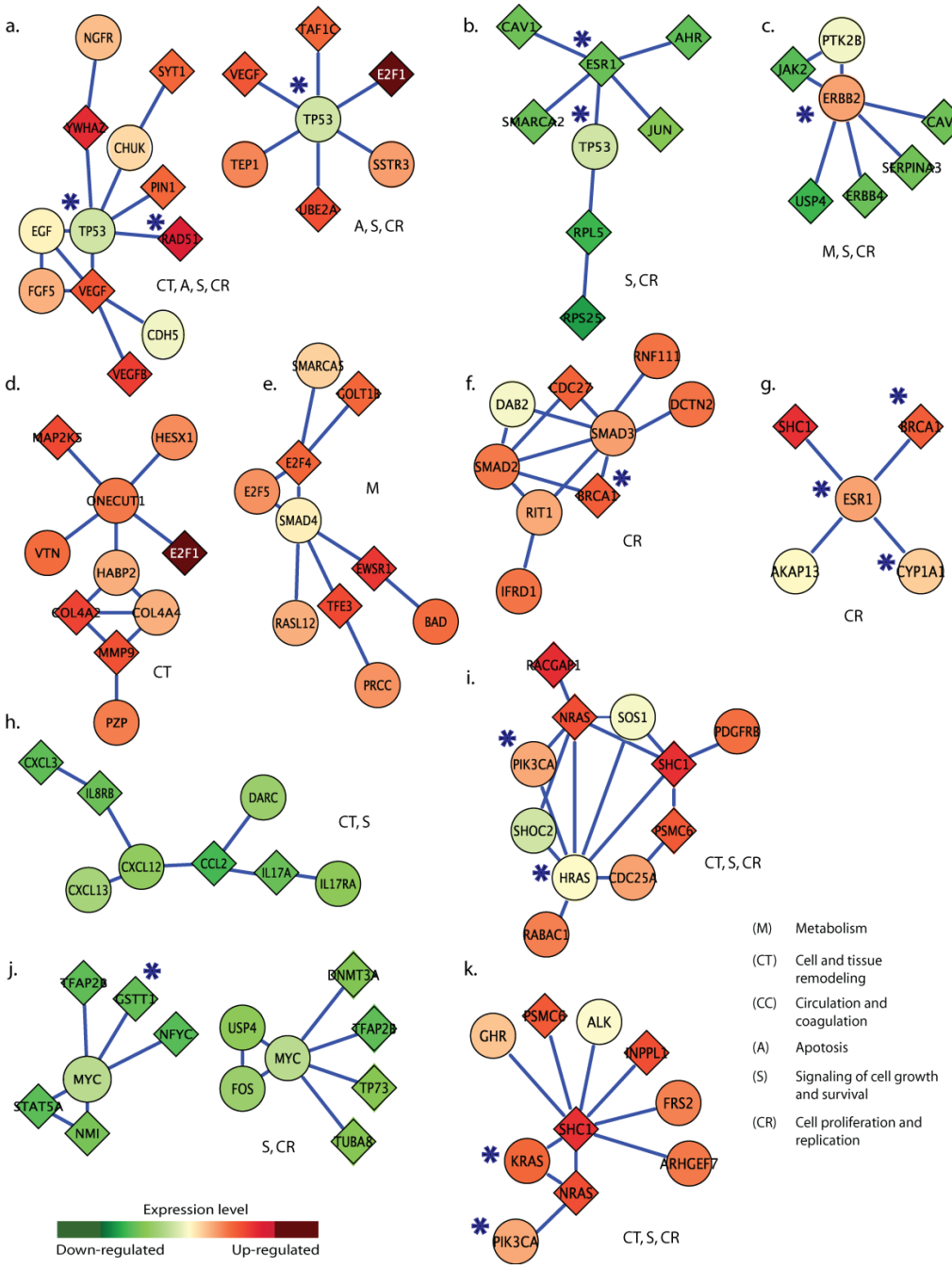


Figure 3.3. Subnetwork enriched for the hallmarks of cancer.

Example discriminative subnetworks from van de Vijver *et al.*[104] are shown in **(a-e)**, while those from Wang *et al.*[23] are shown in **(f-k)**. Nodes and links represent human proteins and protein interactions, respectively. The color of each node scales with the change in expression of the corresponding gene for metastatic versus non-metastatic cancer. The shape of each node indicates whether its gene is significantly differentially-expressed (diamond; $p < 0.05$ from a two-tailed *t*-test) or not (circle). The predominant cellular functions are indicated next to each module. Known breast cancer susceptibility genes are marked by a blue asterisk.

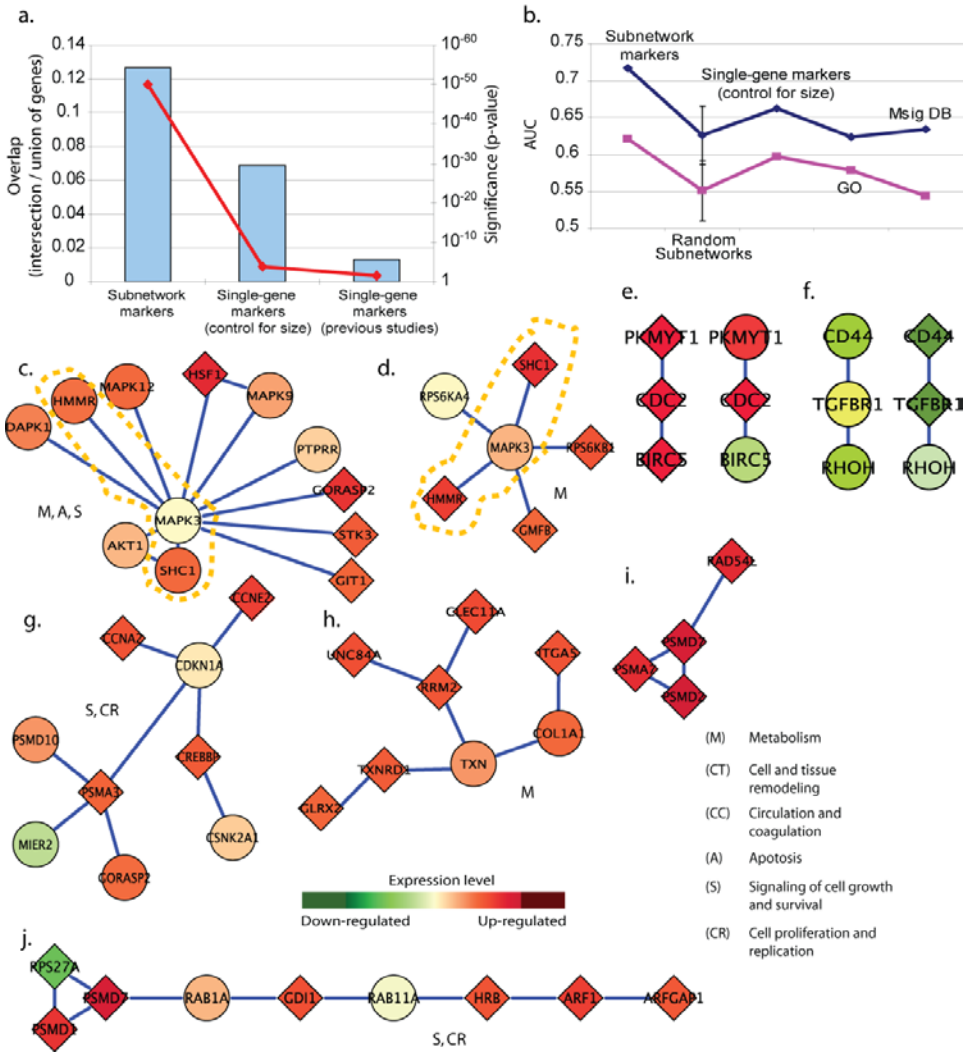


Figure 3.4. Marker reproducibility and metastasis prediction performance.

(**a**) Agreement in markers selected from the van de Vijver *et al.*[104] data set versus those selected from Wang *et al.*[23]. Blue bars chart the magnitude of overlap on the left axis; the red line charts the hypergeometric *p*-values of overlap on the right axis. The first "single-gene" analysis was performed by using the same number of top discriminative genes as the number of genes covered by subnetwork markers. The second "single-gene" analysis was performed by using the same number of top discriminative genes as those in the gene signatures published in van de Vijver *et al.*[104] and Wang *et al.*[23]. (**b**) Area Under Curve (AUC) classification performance of subnetworks, individual genes, or modules from GO or MSigDB. The blue line charts the performance of markers selected based on the Wang *et al.*[23] dataset and tested on the van de Vijver *et al.*[104] dataset; the pink line represents the reciprocal test. The performance of the 1000 random subnetworks is denoted by its mean±stdev. (**c-d**) ***Erk1*** (***MAPK3***) subnetworks in van de Vijver *et al.*[104] and Wang *et al.*[23] (**e-f**) Example network motifs shared between subnetworks selected from the two cohorts. The left-hand side motif is from van de Vijver *et al.*[104] and the right-hand side is from Wang *et al.*[23] (**g-h**) Examples of highly predictive subnetwork markers from Wang *et al.*[23] (**i-j**) Examples of highly predictive subnetwork markers from van de Vijver *et al.*[104]

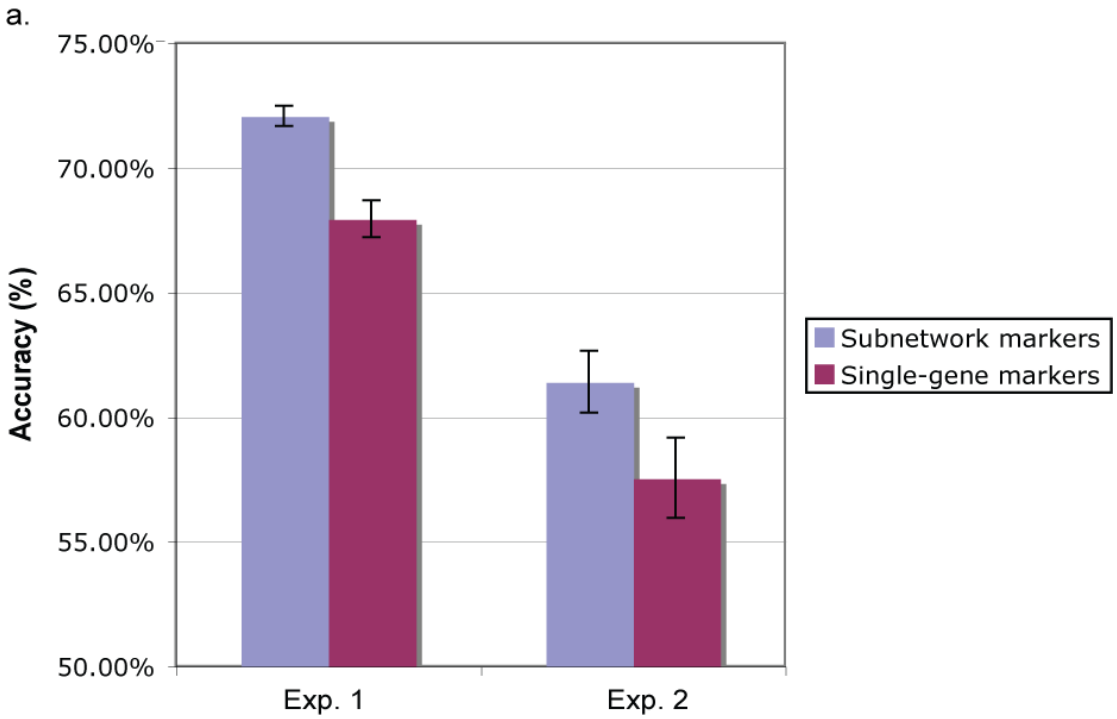


Figure 3.5. Classification performance by using SVM.

Exp. 1 shows the averaged accuracy of markers selected based on the Wang *et al.*[23] dataset and tested on the van de Vijver *et al.*[104] dataset in a 5-fold CV; Exp. 2 represents the reciprocal test (mean±stderr). LIBSVM was trained in the similar way as the logistic regression classifiers in **Figure 3.4b**. The RGF kernel was used with the default parameter setting except the choice of the cost parameter for generalization. The cost parameter was tuned to optimize the accuracy on the validation set.

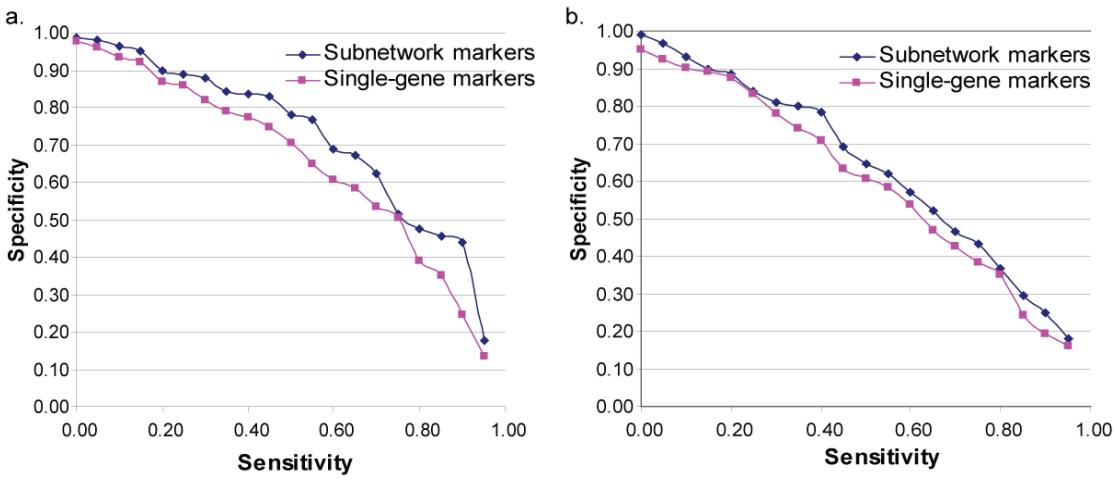


Figure 3.6. Sensitivity and specificity of classifiers using subnetwork markers or single-gene markers in Figure 3.4b.

The corresponding specificity at fixed sensitivity of classifiers using markers based on the Wang *et al.*[23] dataset and tested on the van de Vijver *et al.*[104] dataset **(a)**; **(b)** represents the reciprocal test.

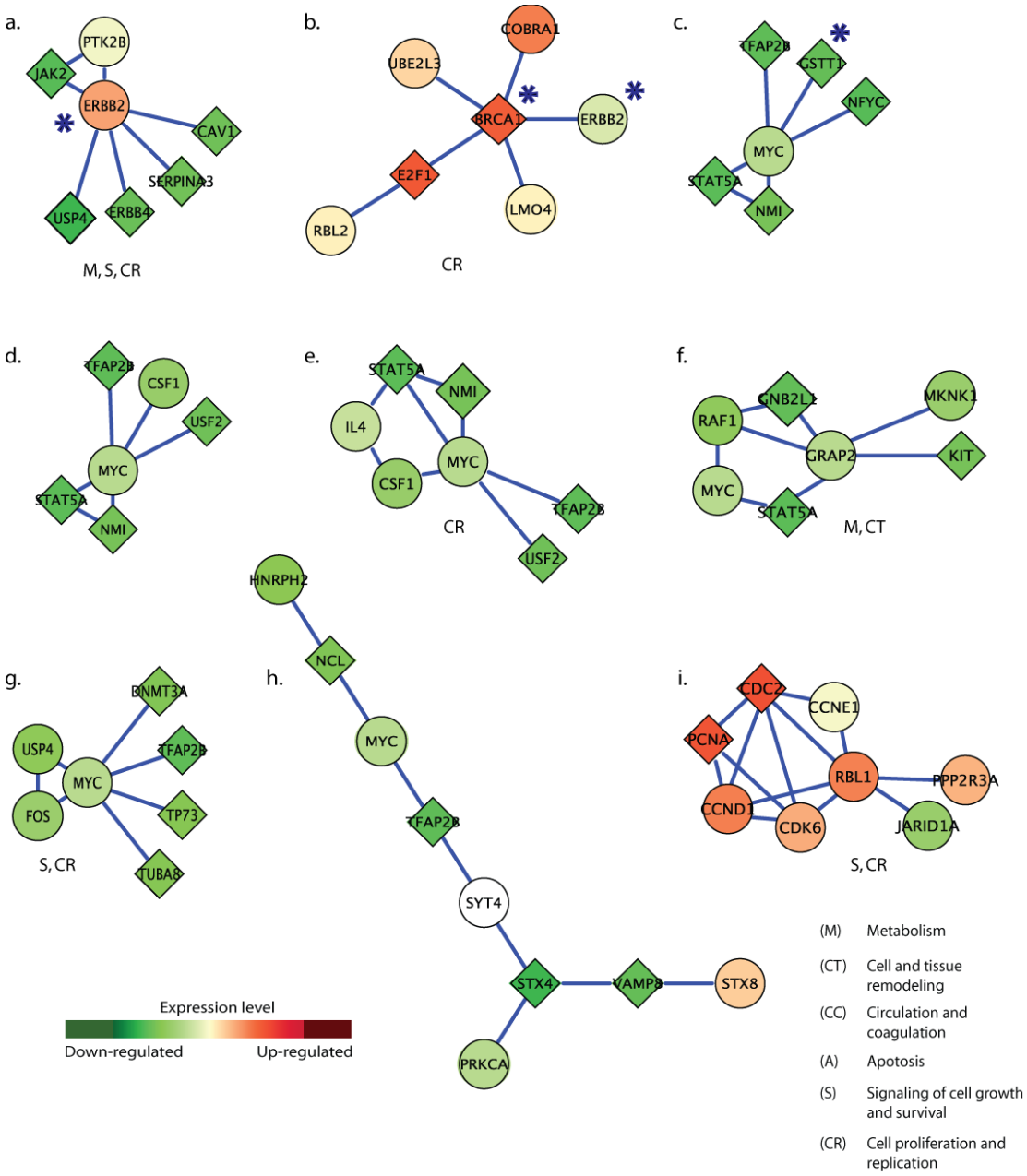


Figure 3.7. Subnetwork markers containing *HER*-2/neu (*ERBB2*), *Myc*, or cyclin D1 (*CCND1*).

Nodes and links represent human proteins and protein interactions, respectively. The colour of each node scales with the change in expression of the corresponding gene for metastatic versus non-metastatic cancer. The shape of each node indicates whether its gene is differentially-expressed: a diamond is significantly differentially-expressed ($p < 0.05$ from a two-tailed *t*-test) while a circle is not. The predominant cellular functions are indicated next to each module. Known breast cancer susceptibility genes are marked by a blue asterisk.

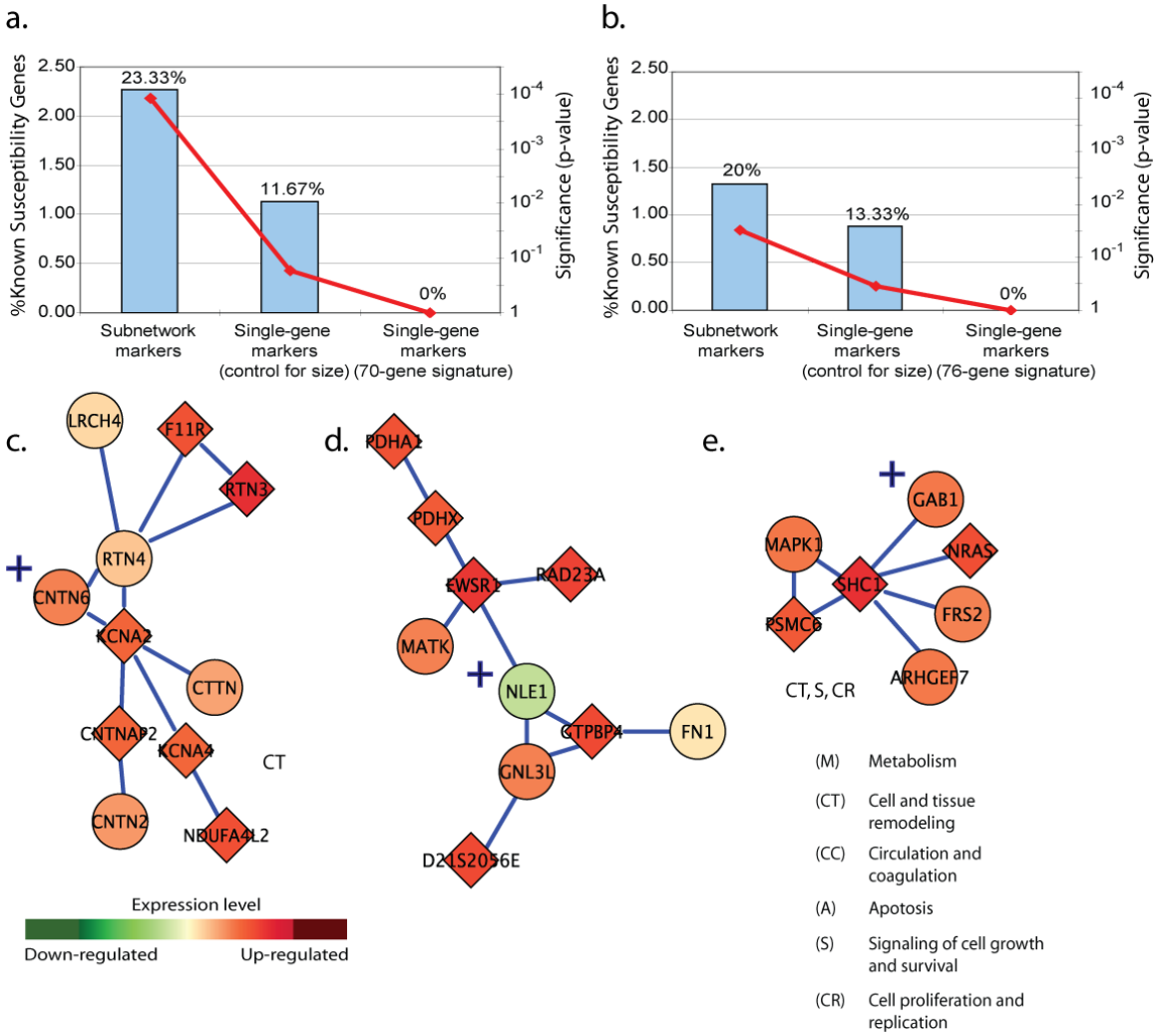


Figure 3.8. Detection of 60 known disease genes in breast cancer.

The enrichment of disease genes is shown for subnetworks or individual genes selected from van de Vijver *et al.*[104] **(a)** or Wang *et al.*[23] **(b)**. Blue bars chart the percentage of disease genes among all genes covered in the markers on the left axis; the red line charts the hypergeometric p-values of enrichment on the right axis. Numbers above the bars are the recovery rates of the known susceptibility genes in each marker set. **(c-e)** Example discriminative subnetworks containing genes with breast cancer mutations listed in Sjoblom *et al.*[45] Mutation genes are marked by a plus sign.

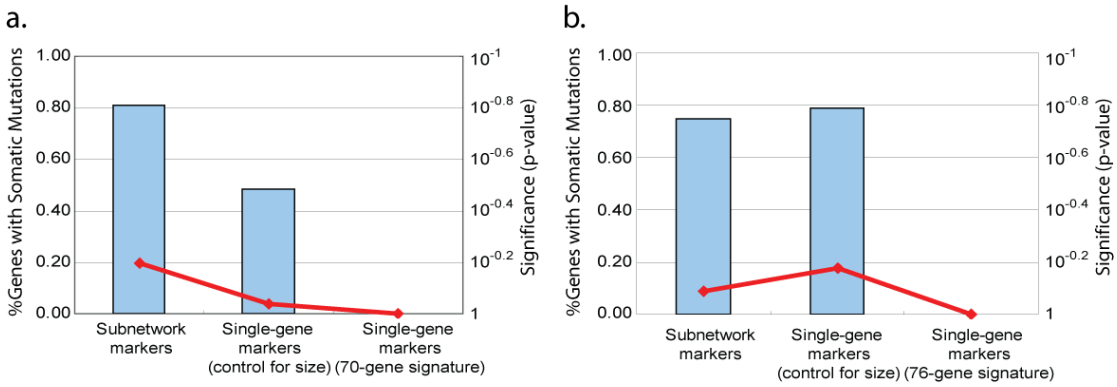


Figure 3.9. Detection of 71 genes with somatic mutations associated with breast cancer in Sjoblom *et al.*[119].

The enrichment of disease genes is shown for subnetworks selected from van de Vijver *et al.*[104] **(a)** or Wang *et al.*[23] **(b)**. Blue bars chart the percentage of disease genes on the left axis; the red line charts the hypergeometric *p*-values of enrichment on the right axis.

Table 3.1. Classification accuracies of the 70-gene selected by van't Veer *et al.* and 76-gene selected by Wang *et al.* in their original studies.

Extracted from Table 2 in van de Vijver *et al.*[104]; patients who were part of the previous study (van't Veer *et al.*[15]) for selecting the 70 genes were excluded.

		Gold standards	
		Metastatic	Non metastatic
Prediction	Metastatic	39 (TP)	65 (FP)
	Non metastatic	3 (FN)	73 (TN)
Total		42	138

Sensitivity = (TP/TP+FN) = 93%

Specificity = (TN/TN+FP) = 53%

Accuracy = (TP+TN) / (TP+FP+FN+TN) = 62 %

Extracted from Table5 in Wang *et al.*[23]

		Gold standards	
		Metastatic	Non metastatic
Prediction	Metastatic	52 (TP)	60 (FP)
	Non metastatic	4 (FN)	55 (TN)
Total		56	115

Sensitivity = 93%

Specificity = 48%

Accuracy = 63 %

Table 3.2. List of 60 breast cancer susceptibility genes.

Gene	Entrez	Locus	Recorded in OMIM
BRCA1	672	17q21	Y
BRCA2	675	13q12.3	Y
TP53	7157	17p13.1	Y
ESR1	2099	6q25.1	Y
PPM1D	8493	17q22-q23	Y
PIK3CA	5290	3q26.3	Y
SLC22A18(BWSCR1A)	5002	11p15.5	Y
RB1CC1	9821	8q11	Y
AR	367	Xq11-q12	Y
RAD54L	8438	1p32	Y
CDH1	999	16q22.1	Y
KRAS2	3845	12p12.1	Y
PHB	5245	17q21	Y
ATM	472	11q22-q23	
PTEN	5728	10q23.3	
STK11	6794	19p13.3	
HRAS	3265	11p15.5	
NAT1	9	8p23.1-p21.3	
NAT2	10	8p22	
GSTM1	2944	1p13.3	
GSTP1	2950	11q13	
GSTT1	2952	22q11.23	
CYP1A1	1543	15q22-q24	
CYP1B1	1545	2p21	
CYP17A1	1586	10q24.3	
CYP19A1	1588	15q21.1	
PGR	5241	11q22-q23	
COMT	1312	22q11.21	
UGT1A1	54658	2q37	
TNF	7124	6q21.3	
HFE	3077	6q21.3	
TFRC	7037	3q29	

Table 3.2. Continued

VDR	7421	12q13.11
APC	324	5q21-q22
APOE	348	19q13.2
CYP2E1	1571	10q24.3
HSD17B1(EDH17B2)	3292	17q11-q21
ERBB2(HER2)	2064	17q21.1
CHEK2	11200	22q12.1
XRCC1	7515	19q13.2
XRCC3	7517	14q32.3
RAD51	5888	15q15.1
LIG4	3981	13q33-q34
SOD2	6648	6q25.3
PPARG	5468	3q25
ITGB3	3690	17q21.32
ITGA2	3673	5q23-q31
MMP3	4314	11q22.3
TGFB1	7040	19q13.1
HSPA1L	3305	6q21.3
HSPA1B	3304	6q21.3
DPYD(DHP)	1806	1p22
TYMS(TS)	7298	18p11.32
CYP2C8	1558	10q23.33
BARD1	580	2q34-q35
NCOA3	8202	20q12
LOH11CR2A(BCSC-1)	4013	11q23
NCOA6(ASC2)	23054	20q11
TSG101	7251	11p15.2-p15.1
TK1	7083	17q23.2-q25.3

4. NETWORK-BASED ANALYSIS OF CHRONIC LYMPHOCYTIC LEUKEMIA IDENTIFIES PATHWAYS THAT CONTRIBUTE TO DISEASE EVOLUTION

The clinical course of chronic lymphocytic leukemia (CLL) is heterogeneous. Gene expression profiling of CLL cells has the potential to discriminate subgroups of patients at different risks for disease progression and immanent therapy. Here, we develop a new system for stratifying patients at different risk levels, based on analysis of CLL gene expression profiles in the context of defined protein interaction networks. We find that gene expression profiles of protein interaction networks can discriminate CLL patients who are at different risks for requiring treatment after tissue collection. We identify 38 networks that can stratify patients more accurately than established markers. In addition to their predictive power, the networks represent an array of disease pathways and suggest novel molecular mechanisms governing CLL progression. We also find increased similarity in expression of these networks over time, regardless of the initial risk category at diagnosis. These results suggest that degenerate pathways may converge into common pathways that govern disease progression. Presently, decisions about the need for therapy are based on the time of diagnosis— our results, based on the time of tissue sampling, have implications for understanding cancer evolution and for the development of novel treatment strategies for patients with CLL.

4.1 Background and significance

Chronic lymphocytic leukemia (CLL), the most common leukemia in the western world, is characterized by accumulation of monoclonal B cells in the blood, marrow, and

secondary lymphoid tissues. The clinical course of patients with CLL is highly variable. Some patients are free of symptoms for many years, during which time treatment is typically not necessary. For others the disease is relatively aggressive and requires therapy soon after diagnosis. Because standard therapies are associated with potential morbidity and are not considered curative, current recommendations are to withhold treatment until the patient manifests disease-related complications or clear evidence of disease progression [120].

Several prognostic markers have been defined that can identify patients with poor prognosis at early stages of the disease. For example, patients segregate into two major subgroups based on whether their leukemia cells express immunoglobulin heavy chain variable region (IGHV) genes that have incurred somatic mutations[121]. Patients with CLL cells that express IGHV lacking mutations generally have a more aggressive clinical course than patients with CLL cells that express IGHV that have incurred somatic mutations[122, 123]. Similarly, patients that have CLL B cells that express high-levels of CD38 or the zeta-associated protein of 70 kD (ZAP-70) progress on average more rapidly than those with CLL cells that have low or undetectable levels of these proteins[122, 124-129].

For many cancers, an increasing number of prognostic markers have been identified through analysis of genome-wide expression profiles[14, 53, 95, 96, 130-135]. Marker sets are selected by scoring each individual gene for how well its expression pattern discriminates between different classes of disease. Several microarray studies have reported sets of genes that are useful as surrogate markers for known prognostic factors in CLL, such as the IGHV mutational status [132, 136-141]. Other studies have instead correlated

gene expression levels directly with median time of patient survival or progression-free survival[142, 143].

Despite their promise, expression-based biomarkers continue to face serious challenges due to their variable accuracy for predicting patient outcomes[22]. In addition, the marker sets obtained by different research groups often share few genes in common. Two landmark studies, Rosenwald et al. (2001) and Klein et al. (2001), each identified approximately 100 genes that were expressed differentially by CLL cells with mutated versus unmutated IGHV. However, only four marker genes were identified in common between studies. One reason for this discrepancy may be genetic heterogeneity across patients, referring to the fact that different genes may be dysregulated in different patients [2-4]. Another reason is that changes in expression of the relatively few genes governing disease progression may be subtle compared to those of the downstream effectors, which can vary considerably from patient to patient [17, 60, 97].

As an alternative approach for identifying disease markers, several groups have integrated gene expression measurements over sets of genes that encode proteins known to interact within protein networks or pathway databases[28-33, 37, 98]. Such prognostic profiles are not listings of individual genes or proteins, but the aggregate expression of subnetworks of genes or proteins within a vast interaction network. These subnetworks can identify gene expression differences between different populations of patients that account for their diverse clinical behavior and— unlike conventional analysis— the roles of these genes in disease are interpretable in the context of networks and pathways.

Here, we pursue a network-based analysis of gene expression profiles to discriminate between groups of patients with disparate risks for CLL progression. The clinical characterization of patients, blood sample preparation, and microarray processing all follow the unified protocol implemented by the Microarray Innovations in LEukemia (MILE) program[144-146], which has proposed standards for microarray-based assays in the diagnosis and sub-classification of leukemia. Unlike conventional prognosis using known factors or gene markers, we make no assumptions about the time of oncogenesis. Rather, the data lead us to propose an alternative clinical variable to assess patients' risk of treatment need, based on the treatment-free survival from the date of tumor sampling rather than from the date of diagnosis. From an initial cohort of 130 patients, we identify 38 prognostic subnetworks that can reliably predict the relative risk for disease progression from the time of sample collection. The prognosis power of these subnetworks is validated on a second cohort of patients in the MILE study and on a published data from CLL patients outside the MILE program. From our own serial samples as well as a published longitudinal study of CLL patients, we find evidence that the subnetwork signatures may evolve over time in the low-risk patient population, converging on a high-risk profile just prior to onset of severe disease.

4.2 Gene expression profiling of peripheral blood from CLL patients

We profiled genome-wide mRNA expression of leukemia-cell samples of 130 CLL patients registered at the Moores Cancer Center (La Jolla, CA, USA) on Affymetrix HG-U133 plus 2 GeneChips (referred to as the UCSD cohort). Lymphocytes were purified from the peripheral blood samples of patients that had not received treatment on

the dates of blood withdrawn, as per the MILE protocol[145]. Expression data were gathered from samples found to have a CLL cell population with greater than 90% CD5+CD19+, as assessed via flow cytometry. Total RNA was isolated and hybridized to Affymetrix HG-U133+2 GeneChips. An independent cohort of 17 patients was selected from 2 European sites in the MILE study (Rome and Munich) and their gene expression profiles were obtained using the same protocol as the UCSD cohort. Of the total of 20,606 genes represented on the microarray, a total of 15,348 had expression levels that were reliably detected in at least 8 patients (5% of the cohort).

4.3 IGHV mutation status cannot reliably predict treatment-free survival from sample collection

As in most CLL studies, the time from diagnosis (DX) to sample collection (SC) (abbreviated as DX→SC) varied significantly among the 130 patients in the UCSD cohort (**Figure 4.1A**). Since leukemia samples were obtained at various times after diagnosis, but prior to therapy, only about 40% of the cohort was sampled within one year of diagnosis and a large proportion (16.9%) of the patients had samples collected five years or more after diagnosis. As expected, patients with leukocytes that used unmutated IGHV had a shorter median time from diagnosis (DX) to therapy (TX) (abbreviated as DX→TX) than did patients with leukocytes that used mutated IGHV (*p*-value = 10^{-5} in **Figure 4.2**). However, the IGHV mutation status was not predictive of the time from sample collection to therapy (abbreviated as SC→TX) for patients whose SC was more than a year after DX (*p*-value = 0.16 in **Figure 4.1B**), reflecting perhaps the fact that IGHV mutation status is a static marker not evolving over time. Therefore, even

patients who have CLL cells that use mutated IGHV ultimately may require therapy even though they continue to have the so-called “good” prognostic feature.

4.4 Transcriptional activity converges between patients of different IGHV status as disease advances

Next, we examined the relationship between sample collection time and the expression profiling data. Given the large variation in the time length of DX→TX across the patient population, we normalized every patient’s sample collection time relative to the DX→TX interval and designated this as the relative sampling time (RST, **Figure 4.1C**). Some patients were sampled at a relatively early disease stage (RST <20%; green bars in **Figure 4.1C**) whereas others were sampled at a relatively disease stage (RST ≥80%; red bars in **Figure 4.1C**). We then compared the expression profiles of CLL cells with unmutated versus mutated IGHV and that were collected at similar RST. The comparison showed that the level of differential gene expression between the two subgroups became lower as RST approached TX (**Figure 4.1D**), suggesting that transcriptional differences between CLL cells of different IGHV mutation status converge with disease progression. Interestingly, the expression levels of only 279 genes differed significantly between early versus late RST for CLL cells that used unmutated IGHV, but 1103 genes differed in expression levels between early versus late RST for CLL cells that used mutated IGHV (FDR ≤20% from a two-tailed t-test; upper inset in **Figure 4.1C**).

4.5 Previous microarray studies yield gene markers of equivalent or less predictive power as IGHV status

We next sought to evaluate whether sets of marker genes proposed by previous studies are prognostic of SC→TX. On the UCSD cohort, five of ten CLL marker sets published previously (see **Section 4.1**) were able to separate the patients into two risk groups with an acceptable difference on their median times of DX→TX (p-value ≤ 0.01 in five-fold cross validation in **Figure 4.1E**, see **Section 4.10**). However, none of them reached the same statistical significance as did the IGHV mutation status. Moreover, only two gene sets, both of which were from studies that took SC into consideration, showed prognostic power on SC→TX (the right-most two sets of bars in **Figure 4.1E**).

4.6 Protein networks stratify CLL patients into different risk groups

Figure 4.3 shows the overall process of network-based disease prognostics, which involves identification of informative subnetworks (**Figure 4.3A**), clustering of patients into subgroups on the basis of their subnetwork activities (**Figure 4.3B**), and Kaplan–Meier survival analysis to assign low-risk and high-risk labels to each subgroup (**Figure 4.3C**). To obtain a human protein interaction network, we assembled a pooled data set comprising 45,526 experimentally-validated interactions among 9,800 human proteins, integrated from yeast two-hybrid experiments[46, 47] [46, 47] and curation of the literature for both protein-protein and protein-DNA binding [44, 82, 84-86, 147]. Of the total of 15,348 genes reliably detected in CLL, a total of 7,589 are covered in the protein network. We overlaid the expression values of each gene on its corresponding protein in the network,

allowing us to consider subnetworks of connected genes whose expression profiles could be aggregated into subnetwork "activity" scores (see **Section 4.9**).

Using this framework, we searched for subnetworks whose activities across the 130 patients in the UCSD cohort were associated with the treatment-free survival SC→TX. A total of 38 prognostic subnetworks were identified from this cohort covering a total of 230 genes and based on a panel of three separate tests for statistical significance (see **Section 2.2**; see **Figure 4.4** for example subnetworks). The prognostic subnetworks included proteins involved in ***WNT*** signaling[148] (**Figure 4.4A**), resistance to apoptosis[149] (**Figure 4.4B**) or cell metabolism[150, 151] (**Figure 4.4Q**), all of which are known factors in CLL pathogenesis. Clustering of the patients by subnetwork activity resulted in one cluster of 54 patients for which the median treatment-free survival was low and a second cluster of 76 patients for which the median SC→TX was substantially higher (**Figure 4.5A**).

We found that the low- and high- risk groups had a strong association with IGHV status: Among all low-risk patients, ~63% had CLL cells that used mutated IGHV (with less than 98% germ-line sequence homology), versus ~40% for high-risk patients (association *p*-value = 0.008 using a Fisher's exact test, **Figure 4.5C**). On the other hand, over one-third of the patients in each group were categorized differently by the subnetwork profiles than by their IGHV mutation status. Interestingly, we found that the low-risk group could be further divided into two clear subgroups, designated low-risk I and II, with very different subnetwork activity profiles (**Figure 4.5A**). The low-risk I patients, whose subnetwork profiles were almost perfectly anti-correlated with those of

the high-risk patients, were also associated with longest treatment-free survival SC→TX (**Figure 4.5B**).

Twenty-two of the 38 significant subnetworks had increased activity in the defined high-risk group (referred to as pro-onconets; see **Figures 4.4A-O** for examples) whereas the other 16 had decreased activity (referred to as anti-onconets; see **Figures 4.4P-T** for examples). Among the protein functions significantly enriched within the 38 subnetworks, the majority related to cell metabolism (45.4%), cell survival, proliferation or death (36.7%), and cell signal transduction (13.2%, **Figure 4.5D**). Several key signaling proteins implicated in CLL literature, such as ***MAPK***/***ERK***, ***TGFβ***, ***CREB*** and ***WNT***, were involved in regulation of multiple subnetworks (**Figure 4.5E**; p-value $\leq 5\times10^{-4}$ from NCBI DAVID analysis).

4.7 Predicting the timing of therapy from the date of sample collection

We next explored the power of the subnetwork markers to make predictions for individual patients. For this purpose, a patient's average gene expression level was calculated for each of the 38 subnetworks; the list of 38 average levels was designated as the patient's subnetwork profile. This profile was predicted as "high-risk" if it correlated with the average subnetwork profiles of the high-risk group better than those of the low-risk group. Conversely, the patient subnetwork profile was predicted as "low-risk" if it better correlated with the average subnetwork profiles of the low-risk group (see **Section 4.10**).

Cross validation within the UCSD cohort showed good predictive performance (p-value = 3.5×10^{-6}; red lines in **Figure 4.6A**). We used a five-fold cross validation

procedure in which four-fifths of patients were randomly selected for subnetwork identification, and the prediction accuracy of these subnetworks was tested on the remaining one-fifth of patients (see **Section 4.10**). A similar cross validation procedure was applied using individual gene expression markers instead of subnetworks. Although these gene-based markers also held prognostic value (p-value = 5.24×10^{-4}; green lines in **Figure 4.6A**), they were significantly less robust than the network-based approach at predicting risk for disease progression. Both prognostics compared favorably with either the IGHV mutation status (p-value = 0.01) or those reported in previous microarray studies (**Figure 4.1E**).

Although cross validation is a useful starting point, it can inflate estimates of accuracy since both the training and testing phases are performed on the same cohort of patients. Therefore, we also examined the data collected from the independent cohort of 17 CLL patients evaluated at other sites participating in the MILE study in Europe (referred as the European cohort). The activity signatures of the 38 subnetworks identified from the UCSD cohort were able to deliver a robust prognosis on the European cohort (p-value = 0.027, **Figure 4.6B**). However, the gene expression markers failed to correctly identify European patients who were at high risk (p-value = 0.714, **Figure 4.6B**). Use of the IGHV mutation status also failed to segregate these patients (p-value = 0.681 in **Figure 4.7**). Strikingly, these markers actually mis-segregated the high risk patients into a subgroup that had a longer treatment-free survival than that of the other patients (**Figures 4.6B** and **4.7**). Furthermore, none of the ten previously-published gene marker sets could stratify patients in this European cohort into subgroups that differed significantly in their intervals of SC→TX.

As yet another independent test of prediction accuracy, we examined an external data set drawn from a previous study outside of the MILE program[152]. The subnetwork signature was validated to stratify patients on the Friedman et al.[152] independent patient cohort (*p*-value = 0.035 in **Figure 4.6C**). However, neither the individual gene expression markers nor the IGHV mutation status were indicative of ~~SC~~X of this patient cohort.

4.8 Convergence of dynamic cll subnetwork transcriptome with disease progression

Thus far, patients were sampled at only one time point. To investigate the correlation between dynamic subnetwork activities and CLL progression, we sought to examine the overall activity changes of all the 38 subnetworks in a previous genome-wide longitudinal expression study in CLL[142] (**Figure 4.8A**). In this study, 13 patients were profiled at each of two time points, one obtained at diagnosis and the other just prior to therapy. On average, more than half of the pro-onconets increased in activity between the time of diagnosis and the time of therapy. Conversely, the anti-onconets decreased in activity over the course of the disease. Remarkably, among the 22 pro-onconets, eleven showed significant activity induction prior to therapy (*p*-value ≤ 0.05 from a paired t-test in **Figures 4.4A**, **4.4C**-**D**, **4.4G**, **4.4I**-**L**, **4.4N** and **4.4O**); three of the 16 anti-onconets were significantly repressed prior to treatment (**Figures 4.4R**-**T**).

Using our own longitudinal samples, we next measured expression changes of the genes implicated in those significant onconets indicative of the disease course of the patients in Fernandez et al.[142] (**Figure 4.8B**, see **Section 4.11**). Leukemia cells of fourteen UCSD patients were sampled serially, at two different time points after DX but

prior to TX, and probed using RT-PCR against a panel of 22 genes. Genes were selected based on their involvement in the predictive subnetworks (pro- or anti-onconets) related to cell cycle (**Figures 4.4C** and **D**), ***MYC*** regulation (**Figures 4.4E-F** and **4.4N**), G-protein signaling (**Figures 4.4I** and **4.4L**), macromolecule metabolism (**Figures 4.4G-H** and **4.4S**) or apoptosis (**Figures 4.4R** and **4.4T**).

We found that genes involved in pro-onconets increased expression over time in most patients (**Figure 4.8B**). Conversely, genes in anti-onconets decreased expression over time in approximately half of the patients but, strikingly, increased expression in almost as many others. Interestingly, the expression pattern of one patient (Patient 14 in **Figure 4.8B**) was completely opposite those of the others.

To determine whether the activity changes inferred from transcription have a functional effect on CLL progression, we selected a ***MYC***-associated subnetwork involved in cell cycle (**Figure 4.4E**), as an example, to examine the serial protein expression in sixteen CLL patients (thirteen patients are the same as in **Figure 4.8B**; see **Section 4.12** for flow cytometry in **Figure 4.9A** and immuno-blotting). Most patients had elevated protein expression level on ***MYC*** (**Figure 4.9B**) and its interacting partner ***CSNK2A1*** over time (**Figure 4.9C**); ***TNFRSF7***, another member gene in the same subnetwork, also showed higher probability of increasing protein expression (2:1 patient ratio of increased versus decreased expression in the middle panel of **Figure 4.9B**). Five of the sixteen patients had both ***MYC*** and ***TNFRSF7*** proteins expressed at higher level in the later-stage samples (Patients 1-3, 6 and 15 in **Figure 4.9B**). Another metabolism-

related subnetwork gene ***MCP*** showed a slight but persistent elevation in protein expression as disease progressed (**Figure 4.9B**).

4.9 Scoring, searching, and pruning subnetworks

A subnetwork is defined as a gene set that induces a single connected component in the protein interaction network. Given a particular subnetwork *M*, let *a* represent its vector of activity scores over the patients, and let *T* represent the corresponding vector of treatment-free survival (SC→TX). To derive *a*, expression values g_{ij} are normalized to *z*-transformed scores z_{ij} which for each gene *i* have μ=0 and σ=1 over all samples *j* (**Figure 2.2**). The individual z_{ij} of each member gene in the subnetwork are averaged into a combined *z*-score, which is designated the activity a_j. The predictive score *S*(*M*) is an estimation of the statistical significance of a as the sole predictor variable on a patient's treatment need in a Cox proportional hazard model on *T*:

$$\frac{H(t)}{H_0(t)} = e^{ka}$$

, where *H*(*t*) is the hazard function at time *t* and $H_0(t)$ is the baseline hazard for an individual when the value of a equals zero. *S*(*M*) is defined as –log *p*-value of a χ^2 test on the above model of the hazard over a null model of only the baseline hazard. Given the predictive score function *S*, a greedy search is performed to identify subnetworks within the protein interaction network for which the scores are locally maximal. Candidate subnetworks are seeded with a single protein and iteratively expanded, with every protein serving as a seed in a separate search. At each iteration, the search considers addition of a protein from the neighbors of proteins in the current subnetwork. The addition that yields the maximal score increase is adopted. After each

addition, the search considers deletion of each protein from the current subnetwork (except those proteins essential to subnetwork connectivity), and deletions that yield higher score are accepted. The search ends when no addition or deletion increases the score over a specified improvement rate *r*. The parameter *r* is chosen as 0.1 to avoid over-fitting to the expression data used. To assess the significance of the identified subnetworks, three tests of significance are performed. In this study, significant subnetworks are selected that satisfy all three tests with $p1<0.05$, $p2<0.05$, and $p3<5\times10^{-5}$. See **Section 2.2** for details on estimation of a null distribution of *S* by permuting the network and expression data as well as mergence of overlapped subnetworks.

4.10 Prognosis evaluation

Given a set of subnetwork markers, patient samples in a training set are clustered into two subgroups by a 2-means clustering method based on similarity in activity. The two clusters of the training samples are labeled as low- or high- risk groups according to their treatment-free survival curves in a Kaplan–Meier analysis. A nearest shrunken centroid classifier[153] is then trained on the subnetwork activity matrix (significant subnetworks versus patient samples) with the risk labels learned from the clustering analysis. For a new patient of unknown prognosis, the expression profile is first transformed into a subnetwork activity profile in the same way as for the training samples. The nearest shrunken centroid classifier assigns the new activity profile to one of the two risk groups whose shrunken mean activity of subnetworks over training samples is more similar to the activity of the new sample. For gene markers, similar risk

stratification and outcome prediction procedures are performed on the original gene expression matrix.

Subnetwork and gene markers were evaluated using two approaches: (1) cross validation within the UCSD cohort and (2) an independent validation using the UCSD cohort as training data and the European cohort and the cohort in Friedman et al[152] as test data. In the cross validation, one-fifth of the UCSD samples were designated as 'test' data and withheld during risk group assignment and classifier training. Subnetwork markers and top gene markers were identified using only the training data. Each of the five patient subsets in the UCSD cohort was evaluated in turn as the test set, while training on the other four sets. The risk-group predictions among the five test sets were pooled to plot two treatment-free survival curves in a Kaplan–Meier analysis. In the second validation approach, subnetwork markers or top gene markers were selected using the whole UCSD data set. Patients in the European cohort or the cohort in Friedman et al[152] were assigned to one of the two risk groups by the classifier learned from the UCSD cohort. In both validation schemes, a log-rank test was used to estimate the significance level of the difference between the survival curves in a Kaplan–Meier analysis.

4.11 Real time PCR for serial gene expression

Total RNA was prepared from frozen PBMCs using Trizol (Invitrogen). Two micrograms of RNA was reversely transcribed using SuperScript III First-Strand Synthesis System (Invitrogen). Expression levels of each gene were measured in triplicates by iQ5 Real-Time PCR Detection System (Bio-Rad) using SYBR Green (Invitrogen) with the primers listed in the **Table 4.1**. Fold change of a gene in two

subsequent samples of the same patient was calculated using the Pfaffl method [154] where β-actin was used as the reference gene to normalize the expression levels across different samples.

4.12 Protein expression analysis using flow cytometry and immuno-blotting

Single-cell suspensions were first stained for surface expression with PE-labeled ***CD5*** and FITC-labeled ***CD19*** antibodies (Pharmingen) to gate on leukemia B-cells. For ***MYC*** expression, cells were then underwent fixation and permeabilization using the Fix&Perm kit (Caltag) and stained with APC-labeled monoclonal human ***c-MYC*** antibody (Cell Signaling). For ***TNFRSF7*** and ***MCP*** expression, cells were stained with PE-labeled ***CD27*** (Abnova) and APC-labeled ***CD46*** (Abnova) monoclonal antibodies, respectively. Flow cytometry analyses were performed using FACScalibur (Becton-Dickinson) with FLOWJO software (Tree Star). The difference in median fluorescence intensity (MFI) between real and isotope stains (referred as δMFI) was used to quantify the protein expression (**Figure 4.9A**). Immuno-blotting was performed using antibodies specific for ***CSNK2A1*** (Abnova) and β-actin (Santa Cruz) on lysates from primary CLL cells.

4.13 Discussion

In this study, we find that the resulting subnetworks provide models charting the molecular mechanisms underlying CLL disease progression. For example, ***MAPK***/***ERK*** signaling cascade has 20 member genes found in our subnetworks. Activation of ***ERK*** functions in cellular proliferation and differentiation[155, 156]. Aberrations in the ***MAPK***/***ERK*** cascade have been implicated in a high proportion of human cancers[157, 158]

and its deregulation leads to the generation of mitogenic signals in essentially all hematologic malignancies[159, 160]. The observations of the five ***MYC***-participating subnetworks and the 14 ***CREB*** target genes included in the subnetworks also suggest the impact of ***MAPK***/***ERK*** signaling on CLL disease progression, given that ***MAPK*** can phosporylate ***MYC*** and ***CREB***. Another prominent signaling protein is ***TGFβ***, which induces apoptosis in numerous cell types[161]. It acts as an antiproliferative factor at early stages of oncogenesis; however later it enhances tumor progression. The participation of ***TGFβ*** in several pro-onconets implies its promoting role in tumor progression, consistent with the observation that in vitro addition of ***TGFβ*** does not increase spontaneous apoptosis of B cells in CLL patients[162, 163], but rather serves as an endogenous growth inhibitor[164]. Same number of pro-onconets and anti-onconets in our subnetwork signature include genes involved in ***TGFβ*** signaling, supporting the potential dual role of ***TGFβ*** in CLL development and progression.

Although genes with known cancer mutations, such as ***MYC*** and ***TGFβ***, are typically not detected through analysis of differential expression, they play a central role in the protein network by interconnecting many expression-responsive genes. We observed that many known cancer genes were connected with each other inside a subnetwork. In all, ~27% of the genes in CLL subnetworks (62 of 230 genes total) were of known cancer contribution (hypergeometric $p = 2\times10^{-15}$ in **Figure 4.10**, see **SUPPLEMENTAL METHODS**). This fraction was very high compared to conventional expression analysis, for which we found 16.5% of genes (38 of top 230 genes) were of known cancer contribution. This higher enrichment was not due solely to

the bias of using a literature-curated network (compare to random networks in **Figure 4.10**). As one explanation for why network analysis performs better, we found that the majority of the cancer genes identified by network analysis (49 of 62) did not exhibit an altered expression pattern as disease progressed ($p > 0.01$ from an uni-variate Cox hazard model on SG→TX). Rather, they were included in the subnetwork s because of their connectivity—i.e., they were required to interconnect many expression-responsive genes (**Figure 4.4**).

The inferior reproducibility of gene-expression based prognosis, not only did the top candidate genes analyzed in this study but also the gene sets published previously, may provide evidences to a long-discussed hypothesis that CLL is caused by complex interactions among different pathways and factors[165]. The concordant changes of the subnetworks in two series of longitudinal patient samples, one from our site and one from a previous study in Spain[142], demonstrate the utilization of subnetwork markers in studying molecular pathways involved in cancer progression.

Many of the prognosis indicators used for segregating CLL patients into different risk categories for disease progression define subgroups that differ in the median times from diagnosis to initial therapy. However, many patients are asymptomatic at diagnosis, but are detected through incidental laboratory findings. Some patients who receive infrequent medical evaluations may have undetected CLL for years prior to diagnosis, potentially shortening the interval between diagnosis and initial therapy. The established clinical staging systems are most useful in predicting outcomes in patients with advanced disease[166]. Patients who are asymptomatic at diagnosis have the greatest requirement for

biomarkers that can predict whether the disease will be indolent or aggressive. Taking such uncertainties and needs into consideration, we sought to identify prognostic markers that reliably could predict the time from sample collection to initial treatment. Unlike many prior microarray studies, which segregated patients using established prognostic markers, we instead focused on defining markers associated with the treatment-free survival intervals of patients. This joint learning of gene expression profiles and clinical variables, often categorized as semi-supervised learning[167, 168], identifies prognostic markers and risk groups simultaneously and has higher potential in research on diseases in which patient subtyping is critical for effective treatment, but precise classification is still under development.

The success of correlating treatment-free survival from the date of sample collection with CLL subnetwork transcriptome suggests the association between inner cell states and the disease stages. The idea of cancer as an evolutionary process is not new[169, 170], but little attention has been drawn on the applications of understanding and predicting neoplastic progression. The association observed here between treatment-free survival and the subnetwork transcriptome supports the notion that transcriptional activity of these subnetworks contributes to, or results from, the dynamic evolution of leukemic cells. With proper normalization on the diverse clinical courses between patients, we find considerable differential gene expression between CLL cells that use mutated IGHV versus unmutated IGHV at diagnosis that fads as the disease progresses to the point of requiring therapy. That the transcriptome difference fades when the two subgroups progress, albeit at different rates, supports the idea of cancer evolution. Putting these together, we re-challenge the “two distinct disease” hypothesis and speculate that 1) the

CLL disease transcriptome evolves over time to reach a state associated with disease requiring treatment, 2) leukemia cells that use unmutated IGHV have a higher risk for rapid evolution to develop the transcriptome associated with disease requiring treatment, and 3) the transcriptome of leukemia cells that use mutated IGHV transforms gradually to a subnetwork transcriptome similar to that of leukemia cells that use unmutated IGHV prior to therapy. Regardless of their IGHV mutation status, our serial patient samples as well as those in a previous longitudinal CLL study both demonstrate elevated expression of the pro-onconets and declining expression of the anti-onconets in the identified subnetwork signature over time, further suggesting that degenerate pathways may converge into common pathways that govern disease progression.

4.14 Acknowledgement

Chapter 4, in full, is a re-editing of the materials submitted for publication. Chuang, H.Y., Rassenti, L., Salcedo, M., Licon, K., Ideker, T., Kipps, T. Network-based analysis of Chronic Lymphocytic Leukemia identifies pathways that contribute to disease evolution, **submitted**. The dissertation author was the primary investigator and author of this paper.

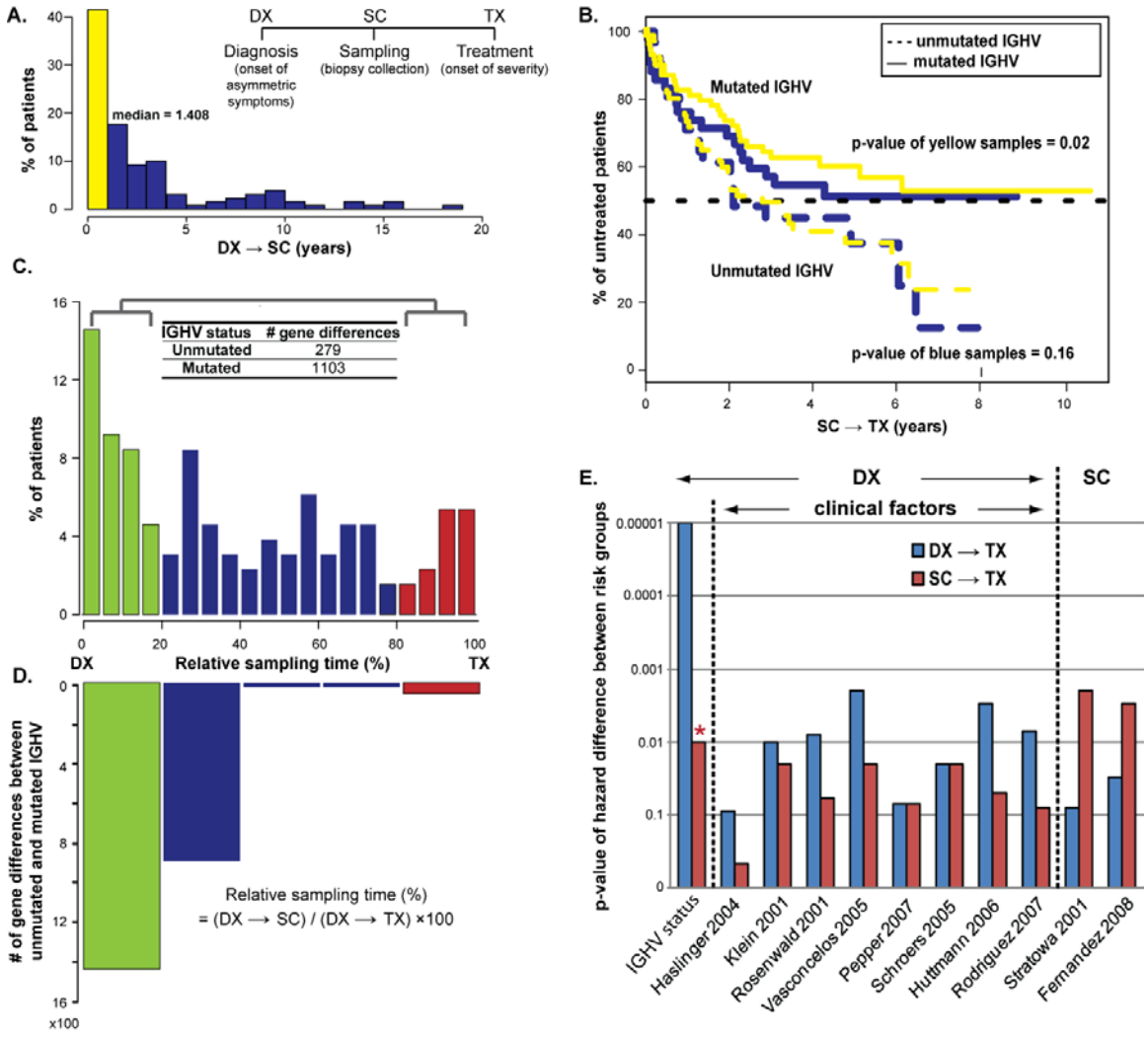


Figure 4.1. Disparity between the date of diagnosis (DX) and the date of tumor sample collection (SC) for patient stratification.

(**A**) A histogram of the time from DX to SC is shown for the 130 patients in the UCSD cohort. (**B**) Survival analysis on SC→TX of the patients whose DX→SC > 1 year versus DX→SC < 1 year with regard to the two risk groups defined by IGHV mutation status [blue samples versus yellow samples in (**A**)]. (**C**) Distribution of relative sampling time (DX→SC normalized by the total time DX→TX) among the 130 patients. Upper inset tabulates the number of differentially expressed genes between early- and late-stage patients among each IGHV subgroup. (**D**) Gene expression differences between IGHV subgroups at different stages of CLL. (**E**) Survival analyses of all 130 UCSD patients using a panel of previously-published marker sets. Bars chart the *p*-value of the difference between the low- and high-risk groups, defined by each marker set reported previously**.** Each marker set is evaluated on both DX→TX (blue bars) and SC→TX (red bars).

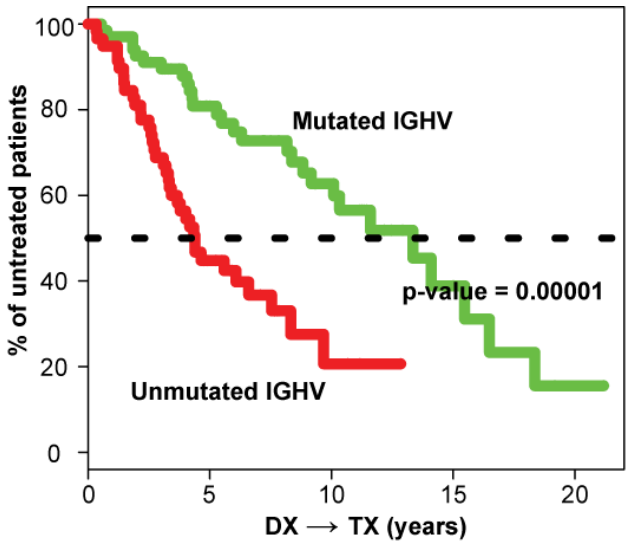


Figure 4.2. Survival analysis on DX→TX of the UCSD patients with regard to the two risk groups defined by IGHV mutation status.

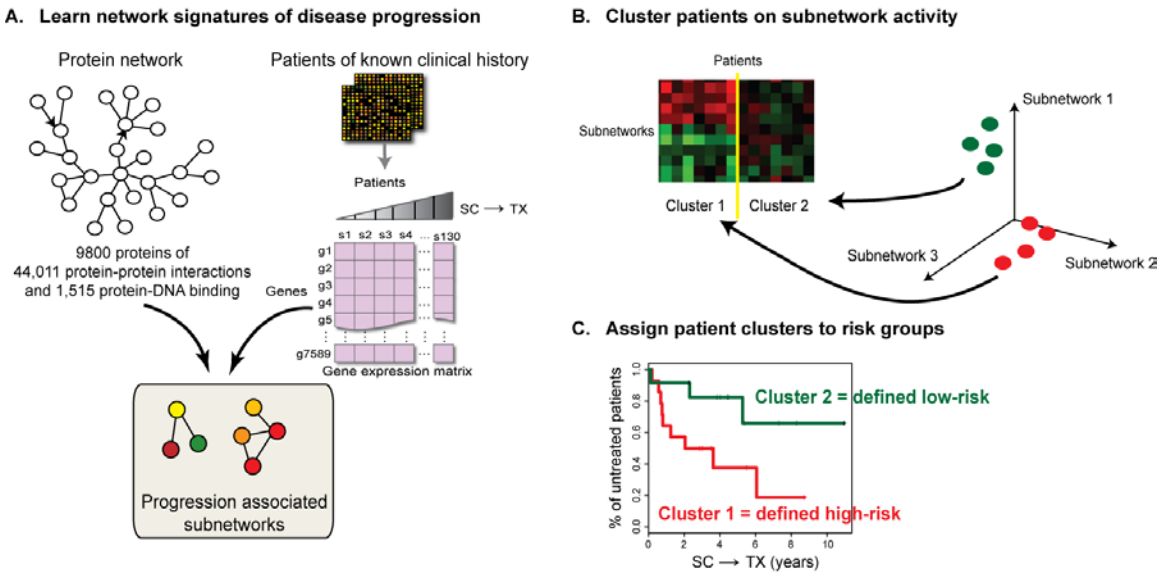
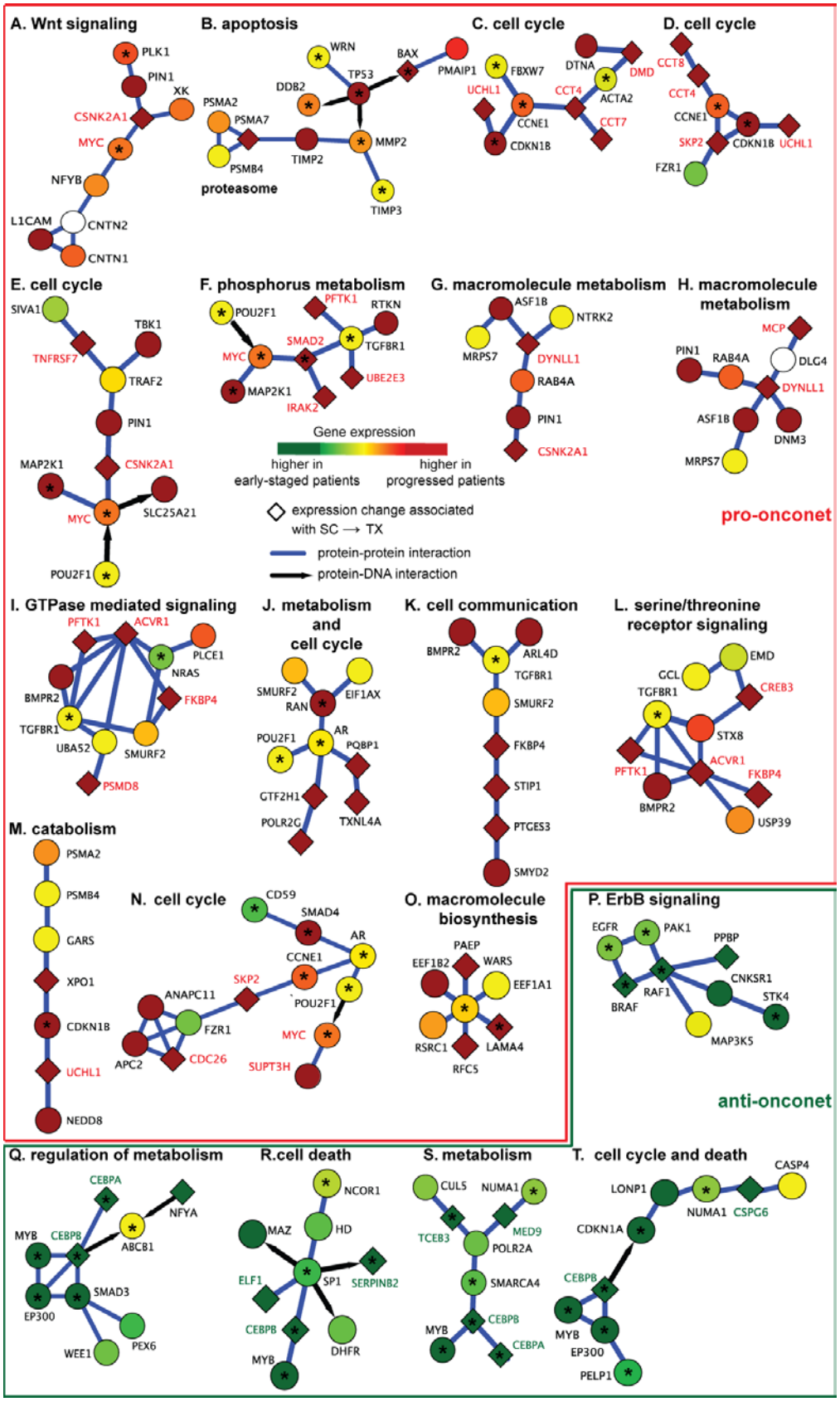


Figure 4.3. Schematic overview of subnetwork identification and definition of risk groups.

The expression profile of each gene is projected onto its corresponding protein in a protein-protein interaction network. A greedy search is performed to find subnetworks for which the activities are associated with the time from Sample Collection to Treatment (SC →TX). Significant subnetworks are selected based on null distributions estimated from permuted data. Subnetworks are used to identify disease genes, and the subnetwork activity is used to characterize the signatures of different risk groups. (**B**) K-means clustering segregates patients by their distinct subnetwork activity patterns. (**C**) Patient clusters are assigned high versus low risk based on median treatment-free probabilities in a Kaplan–Meier analysis.

Figure 4.4. Example subnetworks of CLL disease progression enriched for the hallmarks of cancer.

(**A**-**O**) are pro-onconets and (**P**-**T**) are anti-onconets. Nodes and links represent human proteins and protein physical interactions, respectively. Blue links indicate protein-protein interactions; black arrows indicate protein-DNA binding. The color of each node scales with the change in gene expression in patients of shorter treatment-free survival intervals versus longer: red represents upregulation in patients of shorter intervals whereas green represents down-regulation. The predominant cellular functions are indicated next to each subnetwork. Known cancer susceptibility genes are marked by a black asterisk. Genes of names marked in red/green are further probed for serial expression in an additional patient cohort (red is genes in pro-onconets and green in anti-onconets).


A. Wnt signaling
PLK1
PIN1
XK
CSNK2A1
MYC
NFYB
L1CAM
CNTN2
CNTN1
B. apoptosis
WRN
BAX
PMAIP1
TP53
DDB2
PSMA2
PSMA7
MMP2
TIMP2
PSMB4
proteasome
TIMP3
C. cell cycle
FBXW7
DTNA
DMD
UCHL1
CCT4
ACTA2
CCNE1
CCT7
CDKN1B
D. cell cycle
CCT8
CCT4
CCNE1
SKP2
CDKN1B
UCHL1
FZR1
E. cell cycle
SIVA1
TBK1
TNFRSF7
TRAF2
PIN1
MAP2K1
CSNK2A1
MYC
SLC25A21
POU2F1
F. phosphorus metabolism
PFTK1
RTKN
POU2F1
SMAD2
TGFBR1
MYC
UBE2E3
MAP2K1
IRAK2
Gene expression
higher in early-staged patients
higher in progressed patients
G. macromolecule metabolism
ASF1B
NTRK2
MRPS7
DYNLL1
RAB4A
PIN1
CSNK2A1
H. macromolecule metabolism
MCP
PIN1
RAB4A
DLG4
DYNLL1
ASF1B
DNM3
MRPS7
expression change associated with SC → TX
protein-protein interaction
protein-DNA interaction
pro-onconet
I. GTPase mediated signaling
PFTK1
ACVR1
PLCE1
NRAS
BMPR2
FKBP4
TGFBR1
UBA52
SMURF2
PSMD8
J. metabolism and cell cycle
SMURF2
EIF1AX
RAN
POU2F1
AR
PQBP1
GTF2H1
TXNL4A
POLR2G
K. cell communication
BMPR2
ARL4D
TGFBR1
SMURF2
FKBP4
STIP1
PTGES3
SMYD2
L. serine/threonine receptor signaling
EMD
GCL
TGFBR1
CREB3
STX8
ACVR1
PFTK1
FKBP4
BMPR2
USP39
M. catabolism
PSMA2
PSMB4
GARS
XPO1
CDKN1B
UCHL1
NEDD8
N. cell cycle
CD59
SMAD4
AR
CCNE1
SKP2
ANAPC11
POU2F1
FZR1
MYC
APC2
CDC26
SUPT3H
O. macromolecule biosynthesis
PAEP
EEF1B2
WARS
EEF1A1
RSRC1
LAMA4
RFC5
P. ErbB signaling
EGFR
PAK1
PPBP
CNKSR1
RAF1
STK4
BRAF
MAP3K5
anti-onconet
Q. regulation of metabolism
CEBPA
NFYA
MYB
CEBPB
ABCB1
SMAD3
EP300
WEE1
PEX6
R.cell death
NCOR1
MAZ
HD
ELF1
SP1
SERPINB2
CEBPB
DHFR
MYB
S. metabolism
CUL5
NUMA1
TCEB3
MED9
POLR2A
SMARCA4
CEBPB
MYB
CEBPA
T. cell cycle and death
CASP4
LONP1
NUMA1
CSPG6
CDKN1A
CEBPB
MYB
EP300
PELP1

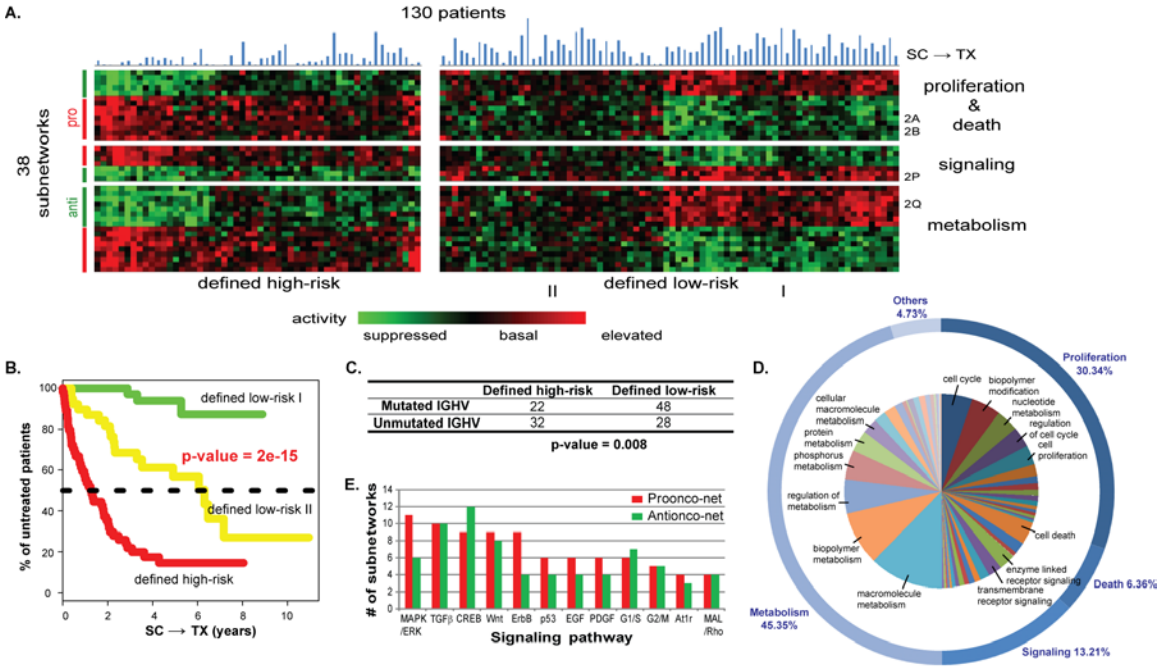


Figure 4.5. Subnetwork signatures of CLL disease progression.

(**A**) Activity of the 38 significant subnetworks (rows) across the 130 patients (columns). The color of each block scales with the activity level of a subnetwork in a particular patient. Patients are clustered into high/low risk groups and subnetworks are clustered into three functional categories (proliferation and death, signaling, and metabolism). Blue bars above the heatmap show the treatment-free survival time of each patient. (**B**) Kaplan–Meier analysis yields treatment-free probabilities with regard to the three risk groups defined by subnetwork activity patterns. (**C**) Comparison of patient stratification by subnetwork prognosis versus IGHV mutation status. (**D**) Distribution of the predominant cellular functions associated with the 38 subnetworks. Related functions are clustered into categories named on the outer circle. The marked functions in the inner circle are associated with at least 2% of the subnetworks. See **Figure 4.11** for all enriched functions. (**E**) Top enriched signaling cascades. Bars show numbers of the 38 subnetworks which have member genes involved in each pathway.

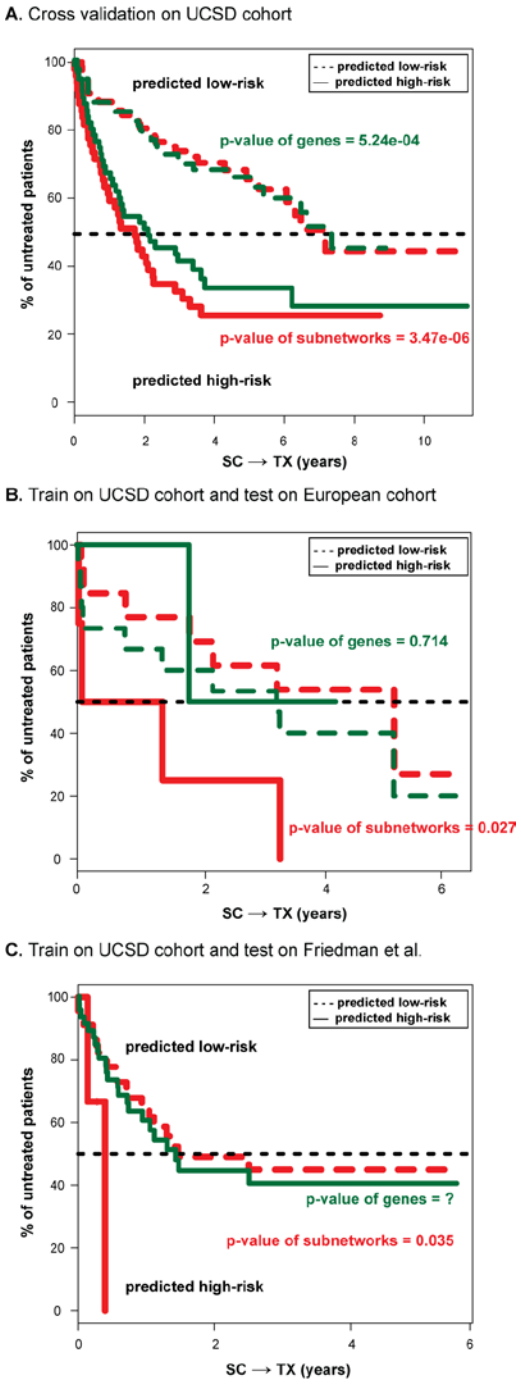


Figure 4.6. Prognosis of new patients.

(**A**) Five-fold cross validation on the 130 patients from UCSD. Survival analyses on SC→TX are shown for both the low (dashed lines) and high (solid lines) risk groups predicted by subnetwork signatures (red lines) or by gene signatures (green lines). (B-C) Survival curves on SC→TX for the 17 European patients (**B**) or for the patient cohort in Friedman et al[152] (**C**). The two risk groups are predicted by two sets of markers developed on the UCSD cohort, including the 38 subnetworks (red lines) and the top 230 genes (green lines).

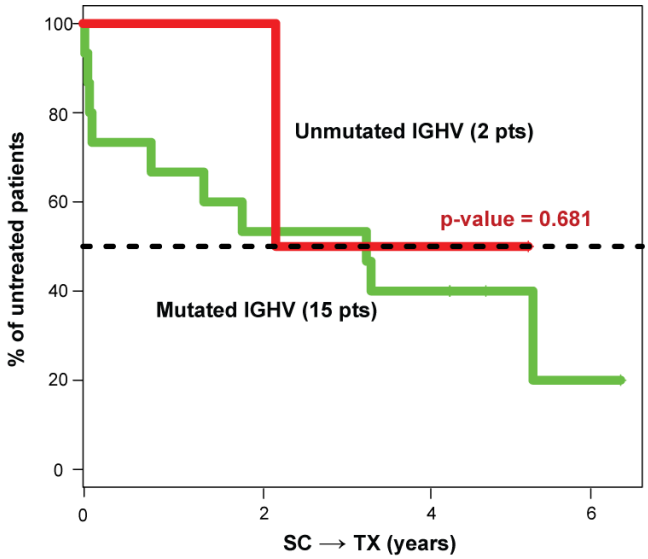
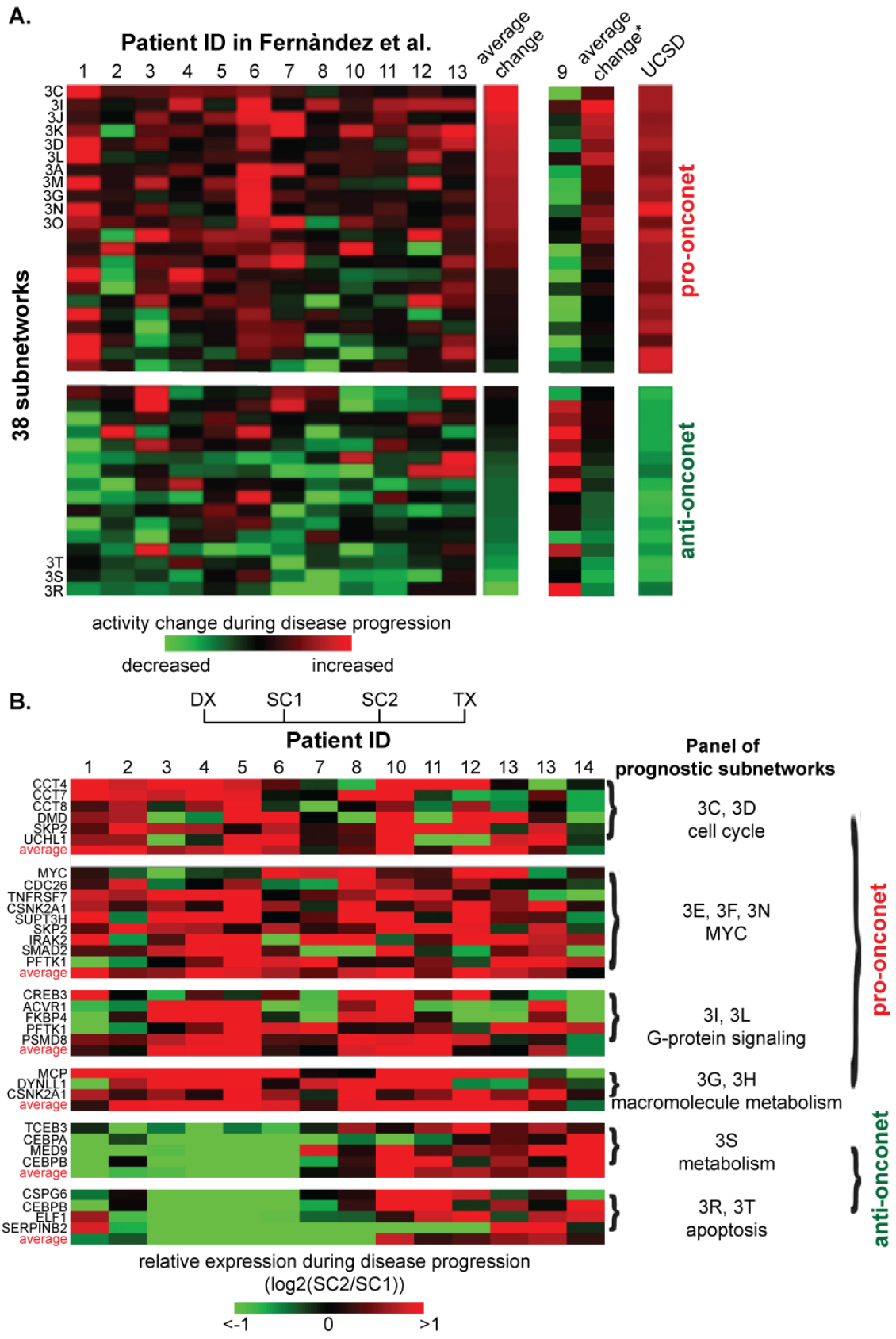


Figure 4.7. Survival analysis on SC→TX of the European cohort with regard to the two risk groups defined by IGHV mutation status.

Figure 4.8. Serial expression of example subnetwork genes and the subnetwork signature along disease progression.

(**A**) Subnetwork activity changes in serial samples of thirteen patients from Fernandez et al.[142] Rows and columns represent subnetworks and patients, respectively. The color of each block scales with the activity change in a subnetwork from the early-stage sample to the progressed sample of a particular patient. The heatmap of patient #9 is separately displayed due to its contrasting pattern versus the other 12 patients. The average change column illustrates the averaged activity change in a subnetwork across patients: the column of an asterik sign represents the average of all the 13 patients and the other one excludes patient #9. The right-most column denotes the prognosis power of the 38 subnetworks on UCSD samples (the coefficient of each subnetwork as the predictor in an uni-variate Cox hazard model on SG→TX). The subnetworks that are significantly differentially activated between early-stage and progressed samples in Fernandez et al.[142] (p-value < 0.05 from a one-tailed t-test) are indicated by the figure panels in which they are displayed (3C, 3I, etc.). (**B**) Gene expression changes in serial samples of fourteen additional patients registered at UCSD. Rows and columns represent genes and patients, respectively. The color of each block scales with the log2-transformed ratio of a gene in the earlier sample as compared to the later sample of a particular patient. The "average" rows illustrate the averaged expression change of genes in similar subnetworks across patients. Genes participating in similar subnetworks are clustered together and the figures of the corresponding subnetworks are indexed next to each cluster.


A.
Patient ID in Fernàndez et al.
1 2 3 4 5 6 7 8 10 11 12 13
average change
9
average change*
UCSD
38 subnetworks
3C
3I
3J
3K
3D
3L
3A
3M
3G
3N
3O
3T
3S
3R
pro-onconet
anti-onconet
activity change during disease progression
decreased
increased
B.
DX SC1 SC2 TX
Patient ID
1 2 3 4 5 6 7 8 10 11 12 13 13 14
Panel of prognostic subnetworks
CCT4
CCT7
CCT8
DMD
SKP2
UCHL1
average
3C, 3D
cell cycle
MYC
CDC26
TNFRSF7
CSNK2A1
SUPT3H
SKP2
IRAK2
SMAD2
PFTK1
average
3E, 3F, 3N
MYC
CREB3
ACVR1
FKBP4
PFTK1
PSMD8
average
3I, 3L
G-protein signaling
MCP
DYNLL1
CSNK2A1
average
3G, 3H
macromolecule metabolism
TCEB3
CEBPA
MED9
CEBPB
average
3S
metabolism
CSPG6
CEBPB
ELF1
SERPINB2
average
3R, 3T
apoptosis
pro-onconet
anti-onconet
relative expression during disease progression
(log2(SC2/SC1))
<-1
0
>1

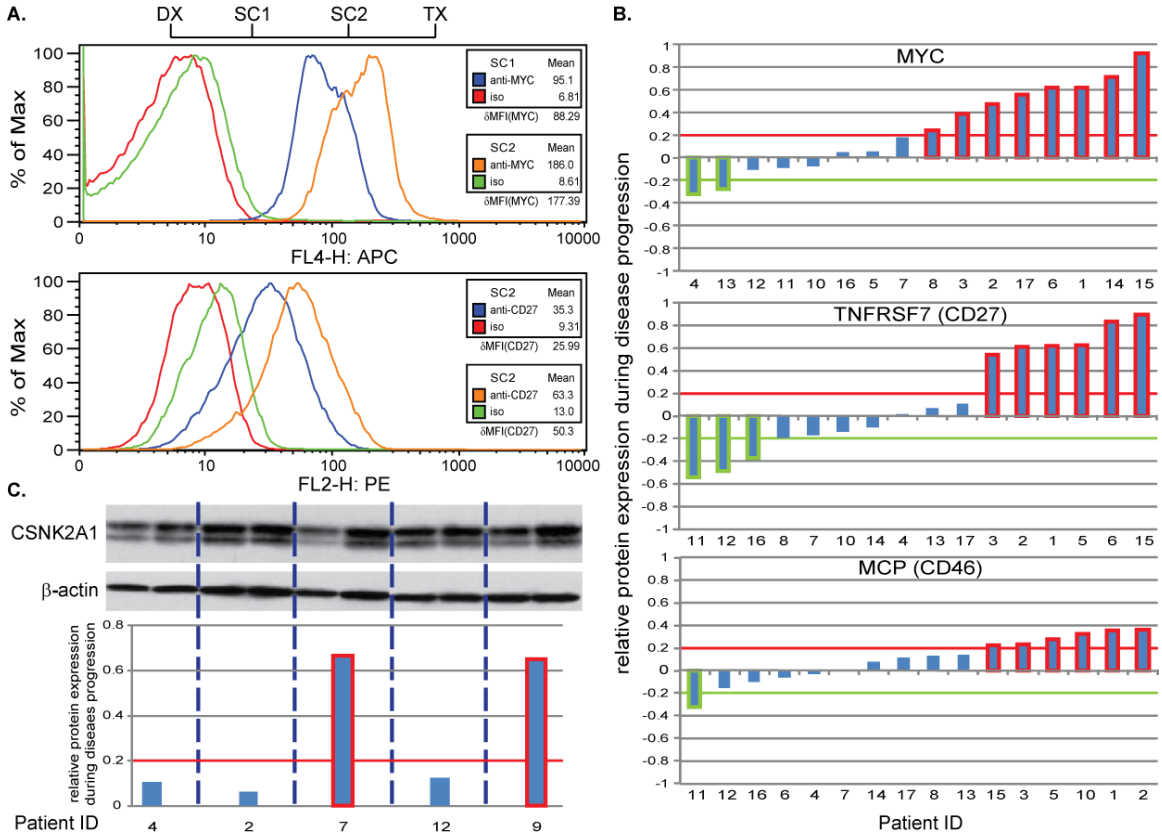


Figure 4.9. Serial protein expression of example subnetwork genes during disease progression.

(**A**) Example of protein expression quantification using median florescence intensity measured by flow cytometry. (**B**) Protein expression changes of ***MYC***, ***TNFRSF7*** and ***MCP*** in serial samples of sixteen patients registered at UCSD. Bars chart the log2-transformed δMFI ratio of a protein in the earlier sample as compared to the later sample of a particular patient (see **Section 4.12**). A threshold value of ±0.2 is selected to highlight patients of differential protein expression over time. (**C**) Immuno-blotting of ***CSNK2A1*** in serial samples of five patients. Bars chart log2-transformed ratio of ***CSNK2A1*** expression (normalized by β-actin) in the earlier sample as compared to the later sample of a particular patient.

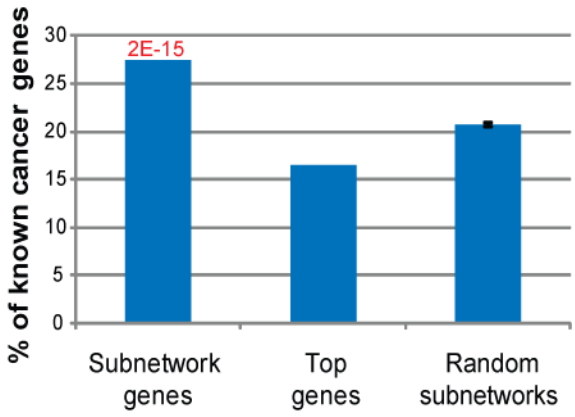


Figure 4.10. Cancer gene enrichment in each marker set.

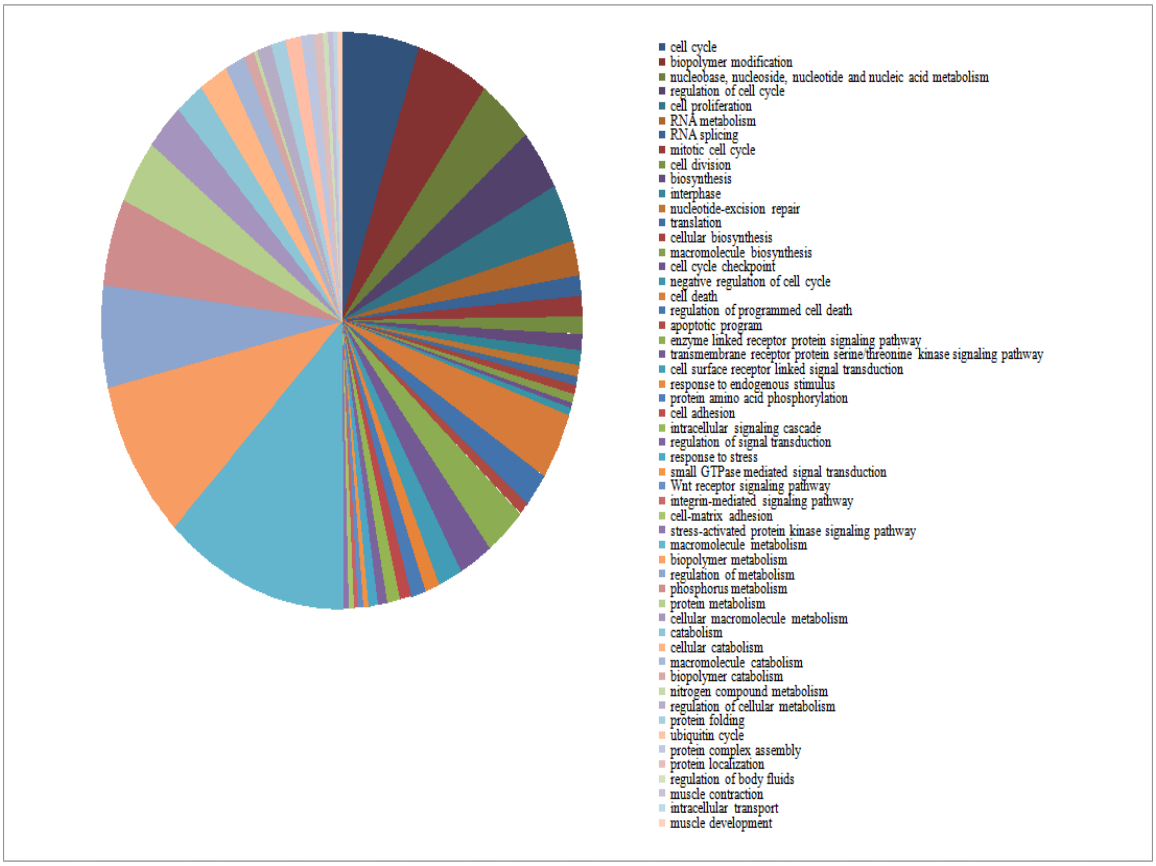


Figure 4.11. Predominant cellular functions associated with the 38 subnetworks.

Pie charts the relative frequency of each cellular function listed on the right.

5. INFERRING PATHWAY ACTIVITY TOWARD PRECISE DISEASE CLASSIFICATION

The advent of microarray technology has made it possible to classify disease states based on gene expression profiles of patients. Typically, marker genes are selected by measuring the power of their expression profiles to discriminate among patients of different disease states. However, expression-based classification can be challenging in complex diseases due to factors such as cellular heterogeneity within a tissue sample and genetic heterogeneity across patients. A promising technique for coping with these challenges is to incorporate pathway information into the disease classification procedure in order to classify disease based on the activity of entire signaling pathways or protein complexes rather than the expression levels of individual genes or proteins.

We propose a new classification method based on pathway activities inferred for each patient. For each pathway, an activity level is summarized from the gene expression levels of its condition-responsive genes (CORGs), defined as the subset of genes in the pathway whose combined expression delivers optimal discriminative power for the disease phenotype. We show that classifiers using pathway activity achieve better performance than classifiers based on individual gene expression, for both simple and complex case-control studies including differentiation of perturbed from non-perturbed cells and subtyping of several different kinds of cancer. Moreover, the new method outperforms several previous approaches which use a static (i.e., non-conditional) definition of pathways. Within a pathway, the identified CORGs may facilitate the

development of better diagnostic markers and the discovery of core alterations in human disease.

5.1 Background and significance

Analysis of genome-wide expression profiles has become a widespread technique for identifying diagnostic markers of various disease states, outcomes, or responses to treatment[14, 15, 23, 53, 95]. Markers are selected by scoring each individual gene for how well its expression pattern can discriminate between different classes of disease or between cases and controls. The disease status of new patients is predicted using classifiers tuned to the expression levels of the marker genes.

One challenge of expression-based classification is that cellular heterogeneity within tissues and genetic heterogeneity across patients in complex diseases may weaken the discriminative power of individual genes[17, 21, 61, 97]. In addition, marker genes are typically selected independently although proteins are known to function coordinately within protein complexes, signaling cascades, and higher-order cellular processes. Thus, the resulting expression-based classifiers may contain unnecessarily many marker genes with redundant information which may lead to decreased classification performance[171].

Due to these types of difficulties, several groups have hypothesized that a more effective means of marker identification may be to combine gene expression measurements over groups of genes that fall within common pathways[25-27, 29, 30, 37, 98]. The pre-defined functional groupings of genes are drawn from canonical pathways curated from literature resources such as the Gene Ontology[57] and KEGG databases[99] or experimentally-defined gene lists from microarray studies[20, 30, 37]. Recently, pathway-

based analysis has been extended to perform disease classification of expression profiles. Some approaches use gene expression parametrically by representing pathway activity with a function summarizing the expression values of member genes[34, 172], while others estimate probabilities of pathway activation based on the consistency of changes in gene expression[64, 65]. Alternative approaches engineer normal cells to activate pre-selected oncogenic pathways to determine gene signatures which can distinguish tumor characteristics[20, 66]. These methods have demonstrated classification accuracies that are comparable to conventional gene-based classifiers, while providing a strong biological interpretation for why the expression profile is associated with a particular type of disease (i.e., based on the pathways found to be perturbed). On the other hand, a potential shortcoming of current pathway-based classifiers is that the pre-defined set of genes making up a pathway may be derived from conditions irrelevant to the disease of interest. Moreover, not all the member genes in a perturbed pathway are typically altered at the mRNA level.

Here, we propose a novel gene-expression-based diagnostic that incorporates pathway information in a condition-specific manner (Pathway Activity inference using Condition-responsive genes, PAC). The markers are encoded not as individual genes, nor as static literature-curated pathways, but as subsets of condition-responsive co-functional genes (Condition-Responsive Genes, CORGs). To optimally discriminate samples of different phenotypes, we identify CORGs from each static pathway in the context of the specific disease in question. The combined expression levels of the CORGs are treated as the pathway "activity" and used to build classifiers for predicting the disease status of new patients. We show that our pathway-based approach

outperforms previous analyses of differential expression in classifying samples across seven different datasets. Moreover, we show that pathway activities inferred using only CORGs lead to better classification performance as compared to pathway activities inferred using various types of summary statistics of all genes which participate in a common pathway. The resulting pathway markers and their CORGs also provide models of the molecular mechanisms which define the disease of interest.

5.2 Datasets

We obtained previously-published mRNA expression datasets covering seven different disease classification scenarios: 24 expression profiles of HeLa cells after stimulation by Tumor Necrosis Factor (TNF)[173], expression profiles of 62 primary prostate tumors and 41 normal prostate specimen[174], expression profiles of 143 acute lymphoblastic leukemia (ALL) patients[175], breast cancer expression profiles for 295 patients from the Netherlands[104] and 286 patients from the USA[23], and lung cancer expression profiles for 86 patients from Michigan[176] and 62 patients from Boston[177].

Each dataset was divided into two populations of distinct phenotypes as per the original publications (**Table 5.1**). For the TNF study[173], 12 samples had normal IkB proteins (labeled "Wildtype") and 12 samples expressed mutant IkB blocking NF-kB signaling (labeled "Mutant"). For the prostate cancer study[174], 62 samples were retrieved from primary tumors (labeled "Cancer") and 41 samples were from normal prostate specimen (labeled "Normal"). For the ALL study[175], 79 patients suffered from one subtype resulting from a t(12;21)(p12,q22) reciprocal translocation (labeled "TEL-AML1") and the other 64 patients showed hyperdiploid hyperdip >50 (labeled "HH").

For the two breast cancer datasets, metastasis had been detected in 78[104] and 106[23] patients during follow-up visits within five and seven years after surgery (labeled "Metastatic"); the remaining 217 and 180 patients were still metastasis free (labeled "Non-metastatic"). For the two lung cancer datasets, we defined the two phenotype populations according to[37], who labeled 24 patients in the Michigan dataset and 31 patients in the Boston dataset as having a "Poor" prognosis, while the remaining 62 and 31 patients were labeled as having a "Good" prognosis.

For pathway information, we used the C2 functional set downloaded from MsigDB v1.0[37]. This set includes 472 canonical metabolic and signaling pathways pooled from eight manually-curated databases along with 50 co-expressed gene clusters obtained from various microarray studies. Each pathway or gene cluster defines a set of genes (gene clusters are henceforth also called "pathways"). In total, the available pathways covered 5602 genes, most but not all of which were measured in the seven gene expression datasets, due to the various array platforms used.

5.3 Condition-responsive gene identification and pathway activity inference

To integrate the expression and pathway data sets, we overlaid the expression values of each gene on its corresponding protein in each pathway. Within each pathway, we searched for a subset of member genes whose combined expression across the samples were highly discriminative of the phenotypes of interest (**Figure 5.1**). For a particular gene set *G*, let *a* represent its vector of activity scores over the samples in a study, and let *c* represent the corresponding vector of class labels (e.g. good vs. poor prognosis). To derive *a*, expression values g_{ij} are normalized to z-transformed scores z_{ij}

which for each gene *i* have mean $\mu=0$ and standard deviation $\sigma=1$ over all samples *j*. The individual z_{ij} of each member gene in the gene set are averaged into a combined z-score which is designated the activity a_j (the square root of the number of member genes is used in the denominator to stabilize the variance of the mean). Many types of statistic, such as the Wilcoxon score or Pearson correlation, could be used to score the relationship between a and c. In this study, we defined the discriminative score $S(G)$ as the *t*-test statistic[178] derived on a between groups of samples defined by *c*.

For a given pathway, a greedy search was performed to identify a subset of member genes in the pathway for which $S(G)$ was locally maximal. We refer to this subset as the set of "condition-responsive genes" (CORGs) representing the majority of the pathway activation under the relevant conditions. To identify the CORG set, member genes were first ranked by their *t*-test scores, in ascending order if the average *t*-score among all member genes was negative, and in descending order otherwise. The CORG set *G* was initialized to contain only the top member gene and iteratively expanded. At each iteration, addition of the gene with the next best *t*-test score was considered, and the search was terminated when no addition increased the discriminative score $S(G)$. The activity vector *a* of the final CORG set was regarded as the pathway activity across the samples.

5.4 Previous gene-set ranking approaches and other pathway-based classification methods

We also used a method proposed by Tian *et al.*[30] to assess the probability of a pathway being altered in disease based on the correlation between the expression of all its

member genes and the disease phenotype. For each pathway *P* in MsigDB, Tian *et al*. calculated a score *T* by averaging the *t*-test statistic scores of all member genes. Higher *T* was indicative of stronger pathway correlation with the disease status. The top 10% of pathways (52 pathways) in each dataset were selected for further analysis and for classification. The decision whether a pathway has been disrupted by disease was assessed on the basis of the discriminating power of the member genes between the groups of interest (using a *t*-test statistic). However, there may be some signatures of pathway disruption that are independent of the classification task at hand. To detect such signatures, a number of statistical functions[17, 179] can be adopted in the framework of Tian *et al*. Unlike the *t*-test, these functions are designed to detect perturbed patterns rather than mean expression changes.

To compare our PAC with other activity inference schemes, we implemented three other expression summarization methods, including a principal component analysis (PCA) similar to that used in Bild *et al*.[20] and the mean and median approaches used in Guo *et al*.[172]. Bild *et al*. used the first principal component of the expression of the member genes to represent the activation of a given pathway, while Guo *et al*. summarized the expression levels of member genes by using simple statistics like mean and median.

5.5 Marker robustness evaluation

For each dataset, 100 alternative two-fold splits were generated of each mRNA expression profile in the dataset. Pathways were ranked on each fold using the method of Tian *et al*.[30], and CORGs for each pathway were identified using the samples in a single

fold. Individual genes were also ranked by their discriminative power on each fold. The robustness was estimated as the average degree of overlap among top pathways/genes derived from the two folds of samples across the 100 splits.

5.6 Classification evaluation

Logistic regression models[108] were trained on both the pathway activity matrix (pathways versus samples) and the original gene expression matrix (genes versus samples—i.e., conventional gene-based classification). For within-dataset experiments, the expression samples in a dataset were divided so that four-fifths of the samples were used as the training set to build the classifier, and one fifth were used as the test set (five-fold cross validation). Each of the five subsets in the dataset was evaluated in turn as the test set and withheld during marker selection (including CORG identification) and classifier training. In order to train a generalized classifier and to minimize over-fitting, we further split the training set into three smaller subsets of equal size: two subsets were used as the marker selection set to rank markers (pathways or genes) as well as identify CORGs (pathways only), and one subset was used as the validation set for assessing which marker set was significant for classification. Thus the CORGs might be different for a specific pathway, depending on the samples used in the marker selection set. Pathways or genes were ranked by the *p*-value of discriminative power to classify samples in the marker selection set, after which the logistic regression model was built by adding markers sequentially in increasing order of *p*-value (sequential selection). The number of markers used in the classifier was optimized by evaluating its Area Under ROC Curve (AUC, see Swets *et al.*[109] for details) on the validation set. The AUC metric

captured performance over the entire range of sensitivity/specificity values. The final classification performance was reported as the AUC on the test set using the classifier optimized from the validation set. For unbiased evaluation, we generated 100 alternative five-fold splits of samples in each dataset and ran cross validation on each split. The final reported AUC values were averaged across 500 randomly selected ways of partitioning the data into four-fifths training and one-fifth test samples.

For cross-dataset experiments, markers (pathways or genes) were selected using the whole first dataset and then tested on the second dataset (or vice versa). CORG identification was also performed on the first dataset. As for the within-dataset experiments, the patient samples in the second dataset were divided into five subsets of equal size: four subsets were designated as the "training" set to build the classifier using markers from the first dataset, and one subset was held for testing. One hundred alternative five-fold splits were generated to partition samples in the second dataset into four-fifths for training and one-fifth for testing. Therefore, we learned 500 classifiers for each of these two datasets, in which each classifier was associated with its own pathway marker set. The averaged AUC values among the 500 classifiers built on the second dataset were reported as the final classification performance for each marker set identified from the first dataset. Among the 500 classifiers, the pathway marker set used in classification could be different depending on which training samples were used in the second dataset. However, the CORGs of each pathway were the same across these 500 classifiers because the identification was done using the whole first dataset.

In this study, for pathway-based classifiers, the input marker set was defined as the top 10% of pathways in MSigDB ranked by Tian *et al.*[30] using a designated training set. In order to compare pathway and gene based methods in a fair manner that controls for the number of genes used, we provided the gene-based classifiers with the same number of top ranked genes as the number of CORGs pooled from the significant pathways selected by Tian *et al.*[30].

5.7 Pathway markers amplify signals over multiple weak gene markers

We first tested the robustness of the pathway markers selected by the method of Tian *et al.*[30]. The agreement between the significant pathways was higher than that between the individually scored gene markers (**Figure 5.2**). The CORGs within the top pathways were also more consistent than individually scored gene markers in different subsets of samples. The observed robustness of CORGs might imply that some non-differentially expressed genes, which are often dropped in conventional analysis, do have associations with the disease of interest.

We hypothesized that pathway information could be used to restrict the search space for truly perturbed genes whose aggregated expression is more predictive for disease status than individually considered. We began by analyzing the breast and lung cancer datasets (four datasets in total), since each dataset has available two separate cohorts of patients studied by different researchers. The top 10% of pathways were selected for each of the four datasets (see **Section 5.3**). We identified the CORGs for each top pathway and aggregated their expression levels into a single activity value for each sample. By design, the inferred pathway activities had more discriminative power

in distinguishing samples with different disease phenotypes than did the individual expression levels of the member CORGs (PAC versus CORGs in **Figures 5.3A**, **5.3C**, **5.3E**, and **5.3G**). However, the discriminative power fell when the pathway activity was inferred using not only the CORGs but all member genes associated with each pathway (PAC_all in **Figures 5.3A**, **5.3C**, **5.3E**, and **5.3G**). This result suggests that, as might be expected, not all genes in a significant pathway are transcriptionally altered or associated with the phenotype of interest.

We then compared our pathway markers to the individual gene markers selected without pathway information. We found that the PAC activity scores outperformed individual gene markers in terms of discriminating samples with different disease phenotypes in both the source datasets used for marker identification (PAC versus Genes in **Figures 5.3A**, **5.3C**, **5.3E**, and **5.3G**) and the independent verification datasets (**Figures 5.3B**, **5.3D**, **5.3F**, and **5.3H**). In the verification datasets, the CORGs demonstrated almost the same discriminative power as did the top genes, although the top genes were more powerful in the original datasets. These comparisons suggest that aggregating the perturbed genes in a pathway leads to a better marker for discriminating disease phenotypes. Although the expression of a single gene might not be a strong predictor, pathway integration provides a means to amplify individual weak signals at the transcriptional level.

5.8 Pathway markers increase the classification accuracy

We next tested that the inferred pathway activity levels could be used in the classification of disease status for a new expression profile. To use pathway information

for classification, pathway activities were used as feature values in a classifier based on logistic regression. The technique of five-fold cross validation was applied to test the predictive power of the pathway markers (see **Section 5.6**). In each run of cross validation, we only considered the top 10% of pathway markers selected by Tian *et al.*[30] using the designated training data.

As shown in **Figure 5.4A**, our pathway-based classifiers (PAC) significantly outperformed the conventional gene-based classifiers (Gene). The improved performance was not simply due to grouping multiple gene expression measurements, as shown by comparing our performance with that of random groups of genes (PAC_random; averaged AUCs of 1000 sets of same-size random gene sets as the significant pathways). Classifiers using pathway activity inferred by the mean or median of the member gene expression[180] or the 1st principle component (PCA)[20] had higher predictive power than those using random gene sets (PAC_random), but only comparable power to the conventional gene-based classifiers. These results indicate that there are at least two critical factors in developing an advanced molecular diagnostic: (1) a biologically meaningful definition of pathways and (2) inference of condition-specific pathway activity.

Next, we tested the reproducibility of the pathway markers selected across different microarray platforms or different cohorts of patients. For this purpose, we used expression profiles of the two lung cancer datasets and the two breast cancer datasets generated from different groups. For each cancer, significant pathways and their CORGs were identified using the whole first dataset and then tested on the second dataset, or vice

versa (**Figure 5.4B**). Our pathway-based classifiers again significantly outperformed the gene-based classifiers.

To show that the better performance of PAC was not dependent on the chosen classification algorithm, we evaluated all markers and pathway activity inference methods using three additional classification approaches: k-nearest neighbors, naïve Bayes, and linear discriminative analysis. Moreover, forward selection method was also employed to show our superior performance was not beneficial from the feature selection method used. All further analyses demonstrated the same trends, i.e., our CORG-based pathway classifiers outperformed other gene-based and pathway-based classifiers (**Figures 5.5** and **5.6**).

5.9 Pathway markers and their corgs provide biologically informative models for lung cancer prognosis

Beyond achieving better classification performance, the discriminative pathway markers and their CORGs can lend insight into the biological basis for why samples are classified as a specific disease status. As an example, we examined the pathway markers selected in the above two cross-dataset experiments for classification of lung cancer prognosis (for a similar analysis of breast cancer metastasis, see **Table 5.2** and **Figure 5.7**). We counted the frequency with which each pathway in MSigDB was selected over the 500 classifiers, and we identified the top most frequent pathways having over 100 occurrences (**Table 5.3**).

Pathways involved in glucose metabolism ("Glycolysis" in **Table 5.3**) and estrogen signaling ("Breast cancer estrogen signaling" and "Estrogen receptor modulators

down-regulated genes") were frequently used in classifying lung cancer patients, and over-expression of these pathways had poor prognosis in both datasets (**Figure 5.8**). Constitutively up-regulated glycolysis has been observed in most primary and metastatic cancers and further explored to develop potential therapeutic targets[181-183]. Up-regulated glycolysis enables unconstrained proliferation and invasion and may lead to a more aggressive type of lung cancer[182]. Estrogen signaling has been known to promote cell proliferation and suppresses apoptosis, and its role in the late steps of lung metastasis has recently been suggested[184]. As shown in **Table 5.3**, many pathways could be represented by CORGs of the size from two to four, although some required more than eight genes (**Figure 5.9**). Especially for larger CORG sets, it would be computationally infeasible to identify these combinations to have maximal discriminative power in the absence of prior pathway knowledge.

5.10 Acknowledgement

Chapter 5, in full, is a re-editing of the materials published in Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T., Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Compu Biol.* 4(11): e1000217 (2008). The dissertation author was the primary investigator and author of this paper.

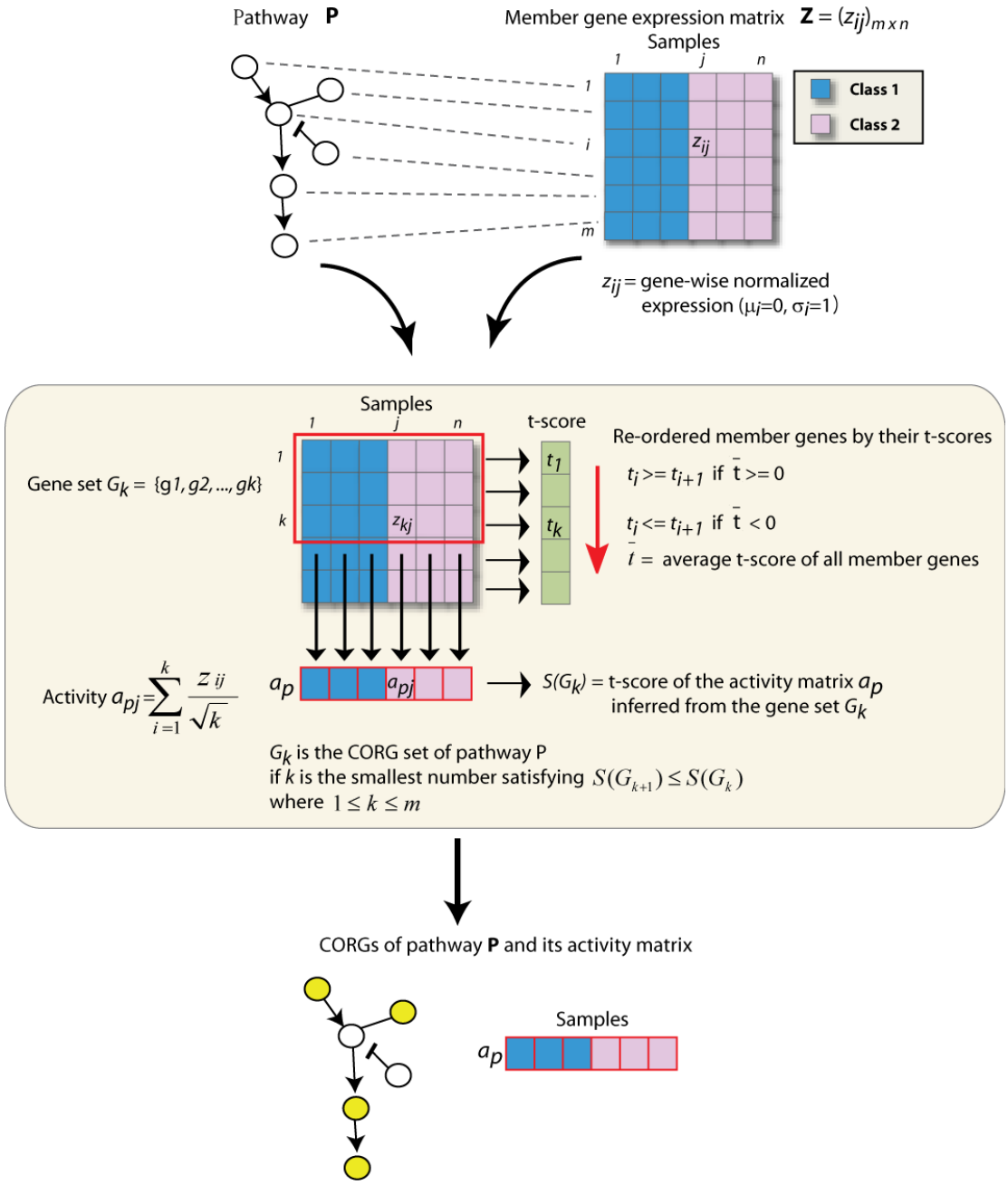


Figure 5.1. A schematic diagram of key gene identification and activity inference.

Selected significant pathways are further subject to CORG identification corresponding to the phenotype of interest. Gene expression profiles of patient samples drawn from each subtype of diseases (e.g., good or poor prognosis) are transformed into a "pathway activity matrix". For a given pathway, the activity is a combined z-score derived from the expression of its individual key genes. After overlaying the expression vector of each gene on its corresponding protein in the pathway, key genes which yield most discriminative activities are found via a greedy search based on their individual power (see **Section 5.3**). The pathway activity matrix is then used to train a classifier.

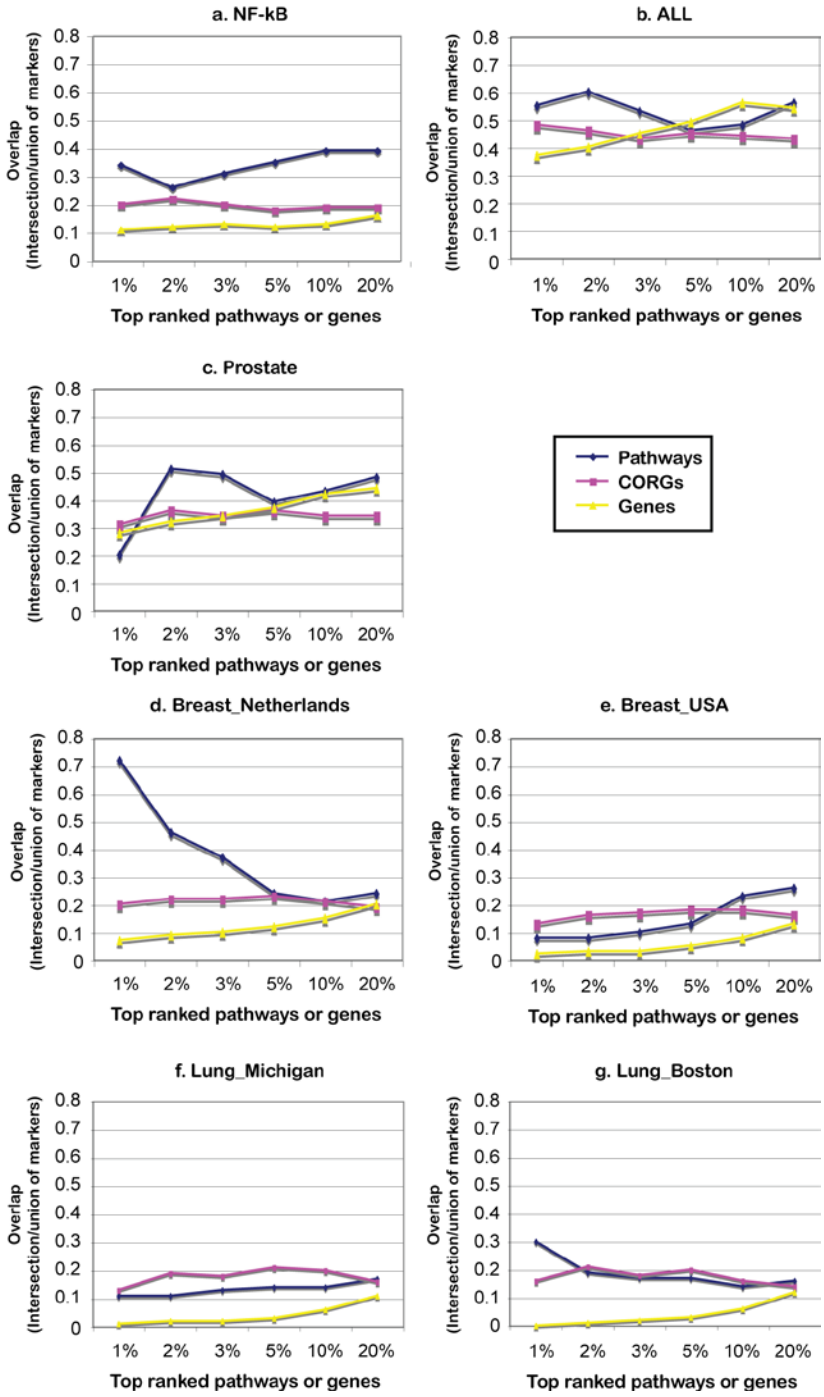


Figure 5.2. **Marker reproducibility of pathway-based and gene-based selection in (a)** NF-kB dataset, **(b)** ALL dataset, **(c)** Prostate dataset, **(d)** Netherlands dataset, **(e)** USA dataset, **(f)** Michigan dataset and **(g)** Boston dataset.

Blue and yellow lines chart the magnitude of overlap among top n markers for pathways ranked by Tian *et al* [30] and genes ranked by conventional *t*-test, respectively. Purple lines chart the magnitude of overlap among member CORGs for the top n pathways. The performance of the 100 alternative splits is denoted by its mean.

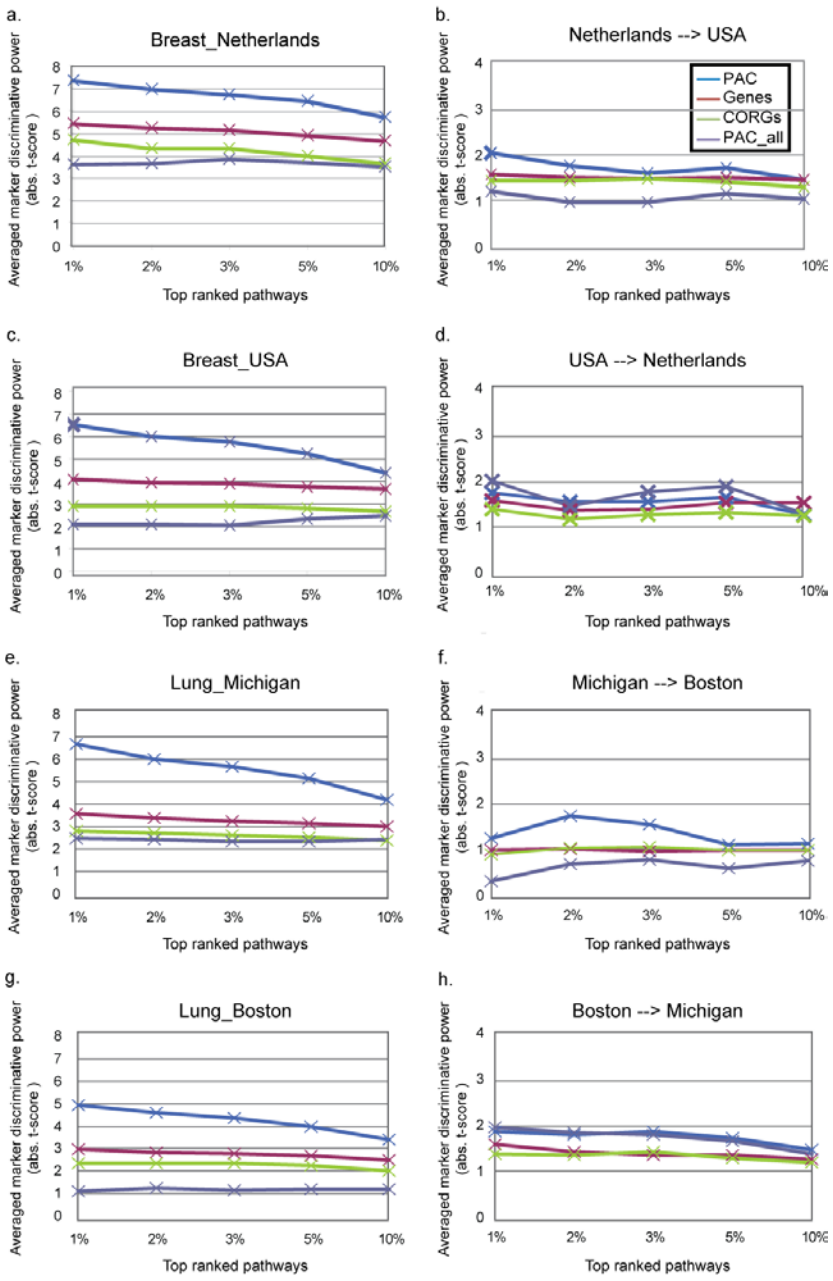


Figure 5.3. Discriminative power of pathway and gene markers in the breast and lung cancer datasets.

Mean absolute *t*-scores against phenotypes were compared between four marker sets in the training set, which was used to rank markers (**a** and **c** for the two breast cancer datasets and e and g for the two lung cancer datasets), or in an independent test set (**b, d, f** and **h**). Pathway markers were ranked by using their absolute *t*-scores from a two-tail *t*-test on activity levels (see *S(G)* in **Section 5.3**) in the training dataset between the two phenotypes of interest. Pathway activities were estimated using only CORGs (PAC) or all member genes (PAC_all). The individual predictive power of CORGs in the top pathways was also evaluated using the same *t*-test on their gene expression levels (CORGs). A similar analysis was performed using the same number of top discriminative genes as the number of CORGs covered by the pathway markers (Genes).

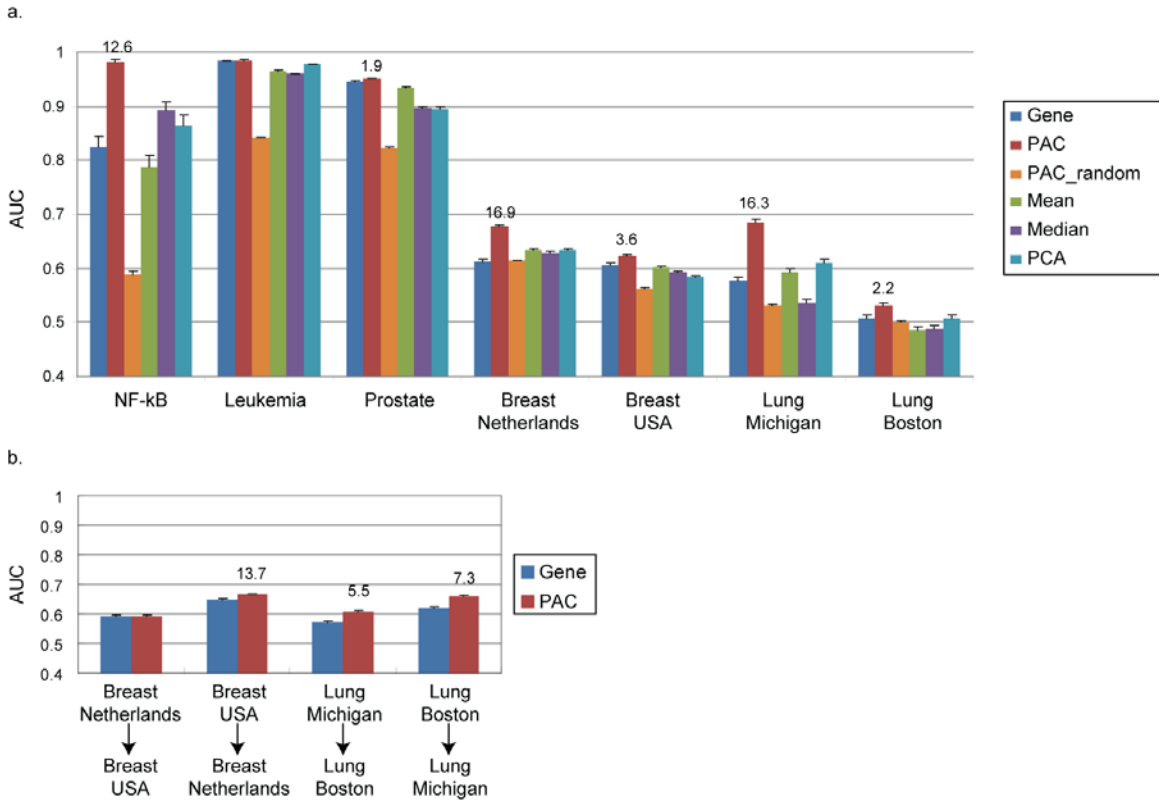


Figure 5.4. **Classification accuracy (a) within- and (b) across- datasets.**

Bar chart of Area Under Curve (AUC) classification performance of CORG-based pathway markers (PAC), conventional pathway markers (Mean, Median, and PCA) and individual genes (Gene; same number of top discriminative genes as the number of CORGs in pathway markers). Classification performance is summarized as mean±ste of AUC over 100 runs of 5-fold cross-validation within a dataset. To compute PAC_random, the AUC values of 1000 sets of random gene sets were averaged. Numbers above the red bars are log (*p*-value) from a Wilcoxon signed-ranked test on the 500 AUCs of "PAC" against those of "Gene" (only the ones with *p*-value < 0.05 are shown). The *p*-values measure the significance of difference between PAC and gene-based classification.

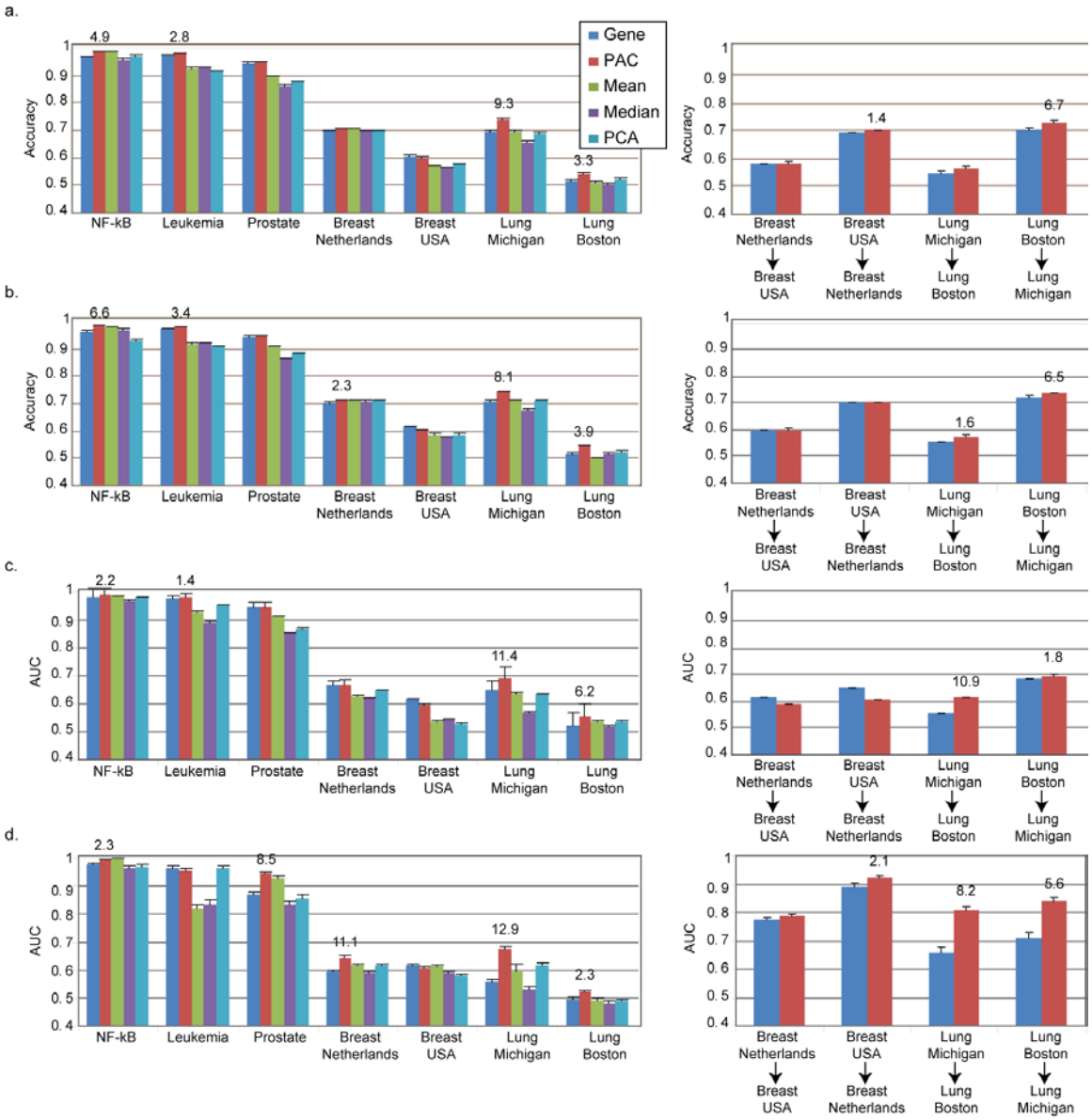


Figure 5.5. **Classification accuracy within- and across- datasets using different classifiers.**

(a) k-nearest neighbor with k = 3, **(b)** k-nearest neighbor with k = 5, **(c)** naïve bayes and **(d)** linear discriminative analysis. Bar chart of classification accuracy in **(a)** and **(b)** and Area Under Curve (AUC) performance in **(c)** and **(d)**. Classification performance is summarized as mean±ste of accuracy/AUC over 100 runs of 5-fold cross-validation within a dataset. Numbers above the red bars are $-\log$ (*p*-value) from a Wilcoxon signed-ranked test on the 500 accuracies/AUCs of "PAC" against those of "Gene" (only the ones with *p*-value < 0.05 are shown).

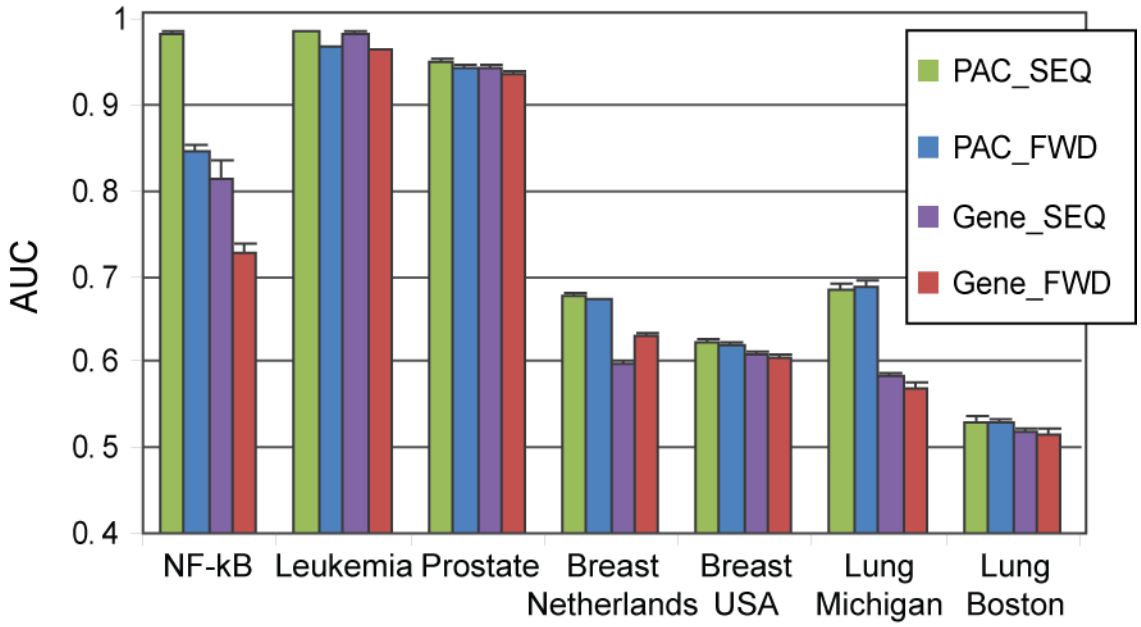


Figure 5.6. Classification accuracy within- and across- datasets using sequential selection (SEQ) or forward selection (FWD).

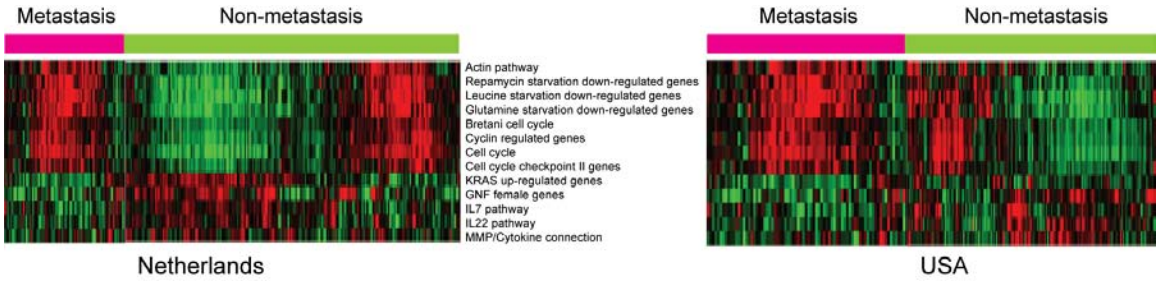


Figure 5.7. Pathway activity of the top frequently-used markers in the two breast cancer datasets.

Activities were inferred from CORGs identified from each dataset. Green/red blocks indicate pathways (rows) that are up-/down- regulated in patients (columns) of specific phenotype (above color bars: pink and green indicate metastasis and non-metastasis, respectively). Pathways are clustered based on the similarity of their activities across patients.

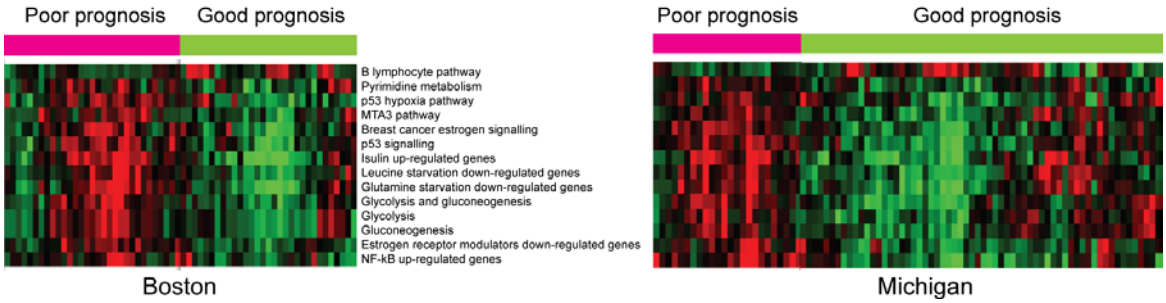


Figure 5.8. **Pathway activity of the top frequently-used markers in the two lung cancer datasets.**

Activities were inferred from CORGs identified from each dataset. Green/red blocks indicate pathways (rows) that are up-/down- regulated in patients (columns) of specific prognosis (above color bars: pink and green indicate poor and good prognosis, respectively). Pathways are clustered based on the similarity of their activities across patients.

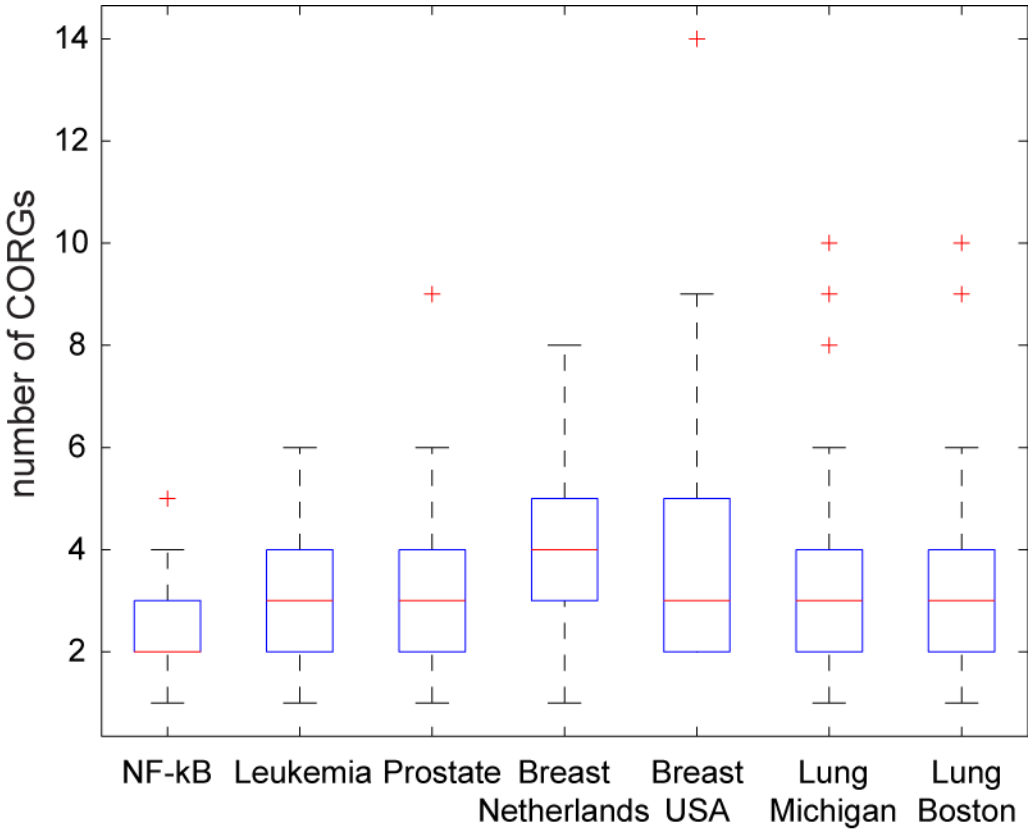


Figure 5.9. Distribution of numbers of CORGs in top 10% pathways.

Table 5.1. The seven data sets used in method evaluation.

Name	Publication	Number of samples in each class
NF-kB	Tian et al. 2005a[173]	Wildtype: 12 samples Mutant: 12 samples
Prostate cancer	Lapointe et al. 2004[174]	Normal: 41 samples Cancer: 62 samples
Leukemia	Yeoh et al. 2002[175]	TEL-AML1: 79 samples HH: 64 samples
Breast_Netherland	van de Vijver et al. 2002[104]	Metastatic: 78 samples Non-metastatic: 217 samples
Breast_USA	Wang et al. 2005[23]	Metastatic: 106 samples Non-metastatic: 180 samples
Lung_Boston	Beer et al. 2002[176]	Poor prognosis: 31 samples Good prognosis: 31 samples
Lung_Michigan	Bhattacharjee et al. 2001[177]	Poor prognosis: 24 samples Good prognosis: 62 samples

Table 5.2. Frequently selected pathway markers for breast cancer prognosis.

Pathway Name	Frequency	# genes *	CORGs
From Netherlands to USA			
Cyclin regulated genes	416/500	2/13	E2F1 *CCNE2*
IL7 pathway	200/500	3/16	BCL2 STAT5A IL7
Cell cycle	197/500	3/84	E2F1 *ESPL1* CCNB2
ActinY pathway	142/500	3/19	PIR PSMA7 ACTR3
GNF female genes	123/500	3/85	RPS4X RPS6 RPL6
From USA to Netherlands			
Cell Cycle	500/500	4/84	CCNE2 *ESPL1* MAD2L1 CDK2
Brentani cell cycle	282/500	4/86	CCNE2 MAD2L1 CDK2 MXD1
Cyclin regulated genes	280/500	3/13	*CCNE2* CDK2 CCNA2
KRAS up-regulated genes	202/500	3/84	TUFT1 P4HA2 COL4A1
Cell cycle checkpoint II genes	200/500	2/10	CCNE2 FANCG
Glutamine down-regulated genes	167/500	6/313	TCEB1 KPNA2 CYCS TMED9 UTP18 MORF4L2
MMP/Cytokine connection	148/500	5/15	DEAF1 TNFRSF1B CD44 IL1B TGFB2
Leucine down-regulated genes	136/500	6/180	TCEB1 KPNA2 CYCS TDG CCT6A CSE1L
IL22 pathway	124/500	3/13	SOCS3 STAT5A STAT3
Rapamycin down-regulated genes	111/500	4/229	STAU1 CYCS RAE1 MORF4L2

* The number of CORGs and member genes are specified.

** Pathways/Genes in italics are shared between datasets

Table 5.3. Frequently selected pathway markers for lung cancer prognosis.

Pathway Name	Frequency	# genes*	CORGs
From Michigan to Boston			
Glutamine up-regulated genes	433/500	5/313	NP LDHA BZW1 TUBA1 LAMB3
Gluconeogenesis	247/500	2/32	LDHA ENO2
Glycolysis	245/500	3/22	*ENO2 PGK1* ALDOA
Breast cancer estrogen signaling	203/500	3/101	VEGF *KRT18 KRT19*
Glycolysis and gluconeogenesis	176/500	5/55	GAPD LDHA ENO2 ALDH3B2 ALDH3B1
Estrogen receptor modulators down-regulated genes	138/500	4/74	ARHE STC1 *KRT7* COPEB
Leucine down-regulated genes	134.500	4/180	NP LDHA TUBA1 CCNA2
B lymphocyte pathway	102/500	4/11	CR2 ITGAL HLA-DRA CR1
From Boston to Michigan			
Breast cancer estrogen signaling	481/500	6/101	*KRT18 KRT19* GAPD MT3 CDKN2A TFF1
Pyrimidine metabolism	258/500	3/45	POLR2E NP RRM1
Glycolysis	258/500	2/22	*ENO2 PGK1*
MTA3 pathway	238/500	3/16	TUBA1 GAPD MTA1
Insulin up-regulated genes	165/500	10/235	PGAM1 ARF4 ARCN1 DNCL1 EIF2S2 PSMA6 YWHAH PSMA3 ZNF9 CLNS1A
P53 hypoxia pathway	148/500	3/20	FHL2 IGFBP3 HIF1A
Glutamine down-regulated genes	133/500	4/313	PGAM1 ERH PAICS BZW1
p53 signalling	114/500	6/101	HIF1A FADD GAPD APEX1 CDKN2A CSNK2B
Estrogen receptor modulators down-regulated genes	108/500	3/74	*KRT7* DUSP4 MMD
NFKB up-regulated genes	103/500	2/111	KRT7 GBP1

* The number of CORGs and member genes are specified.

** Pathways/Genes in italics are shared between datasets

6. DISSECTING DISEASE PROGRESSION OF CHRONIC LYMPHOCYTIC LEUKEMIA USING AN INTEGRATED QUANTITATIVE PROTEOMIC AND GENOMIC ANALYSIS

The clinical course of patients with chronic lymphocytic leukemia (CLL) is heterogeneous. Mapping the pathways that lead to variable courses in progression is one of the key challenges of CLL research. Microarray studies have highlighted differences in mRNA levels found between such CLL subgroups; however, different study identifies a different set of marker genes. One reason for the discrepancy seen in gene array studies is that mRNA measurements are often noise and do not necessarily correlated with the activity of the corresponding proteins.

To address this limitation, we applied a shotgun proteomic method using iTRAQ isobaric tags combined with LC-MS/MS to study differential expression at protein level in CLL patients of different prognosis. A total of 2556 proteins were uniquely identified in at least 2 of the 5 pairs of aggressive and indolent CLL patients. Different from conventional analyses that select marker proteins based on fold changes, we developed a method, Significance Analysis of Mass spectrum based protein quantification (pSAM), that assigns a score to each protein on the basis of change in expression relative to the standard deviation of repeated measurements across replicate MS/MS runs. Between the two CLL subtypes, pSAM identified 69 proteins of significantly changed expression. We demonstrated that the protein markers selected by pSAM have 2 times higher chance to function in a same pathway than those selected by conventional fold-change methods. Moreover, our protein markers are of a more correlated change on gene expression and of

a more robust expression pattern across different cohorts of patients. Furthermore, the MS2 measurements enabled by iTRAQ regents were shown to be more sensitive and reliable than previous MS1 measurements using oxygen isotopes.

We next showed that protein differential expression is coherent with their interaction. By incorporating protein interaction networks into gene and protein expression of aggressive and indolent CLL patients, we identified several differentially expressed subnetworks of consistent or variable changes between genes and proteins. Some subnetworks support known notations of the mechanisms underlying disease progression; while others provide novel hypotheses that are currently testing using immunoblotting and knockdown assays in lab. We also proposed strategies for utilizing comprehensive gene expression, limited protein expression and interaction networks to predict the differential expression of proteins that have not been measured in the shotgun proteomic analysis. The prediction can then be used to guide the design of next targeted proteomic experiments.

6.1 Background and significance

The advent of the whole-genome microarray technology has made mRNA expression profiling become the method of choice for identifying gene markers able to diagnose the severity of disease and predict future disease outcome. However, for some cancers, mRNA-based classification has yet to achieve high accuracy. For example, two large-scale breast cancer studies[15, 23] each identified sets of ~70 gene markers that were only 60-70% accurate for prediction of metastatic versus non-metastatic tumors and shared only three genes in common. One reason for these problems is that mRNA

measurements are often noisy and do not necessarily correlate with the activity of the corresponding protein. A solution to this difficulty is to obtain direct measurements of protein levels and states. Proteins are relevant biomarkers because of their easy assay in the clinic practice. Furthermore, proteins are the proximal cause of most pathology and are the targets of most drugs. Genome-wide discovery of protein biomarkers has been held back due to the lack of an established method that is sensitive, quantitative, and applicable to mammalian cells.

From the genome and transcriptome sequences, we have been presented opportunities to derive a relatively complete list of proteins, by extending genes, thus revolutionizing the field of proteomics. Recent advances in mass spectrometry based protein-profiling technologies[185] have allowed a high throughput systemic analysis of the cellular machinery. Although mass spectrometry is effective in the identification of peptides but not of complete proteins, by matching mass spectra to genomic sequences[186], the high-performance combinations of chromatography and tandem mass spectrometry (MS/MS) enable the identification of a growing number of proteins and can also identify when they have complex secondary modifications[187-189]. Quantitative analysis of protein and peptide constituents can be achieved by isotopic labeling of proteins and peptides, such as $^{16}O/^{18}O$ oxygen isotopes[190], Isotope Coated Affinity Tags (ICAT)[191], stable isotope labeling by amino acids in cell culture (SILAC)[192], and isobaric tag for relative and absolute quantification (iTRAQ)[193, 194], or by label-free quantification of derived mass spectra.

Quantitative MS/MS holds great promise for disease diagnosis, as a method for rapidly identifying proteins whose expression levels or modification states are significantly altered between diseased and normal individuals (i.e., protein biomarkers of disease)[195-197]. As with many instruments, a key challenge of mass-spectrometry-based proteomics is to increase measurement sensitivity—i.e., the lower limit of detection for protein abundance. Another set of issues involves biological and technical reproducibility. Even if the protein identifications and abundances are reproducible over different measurements of the same sample (technical replicates), they may be less so when the entire experiment is repeated over multiple individuals or cell/tissue isolation steps (biological replicates). In complex conditions such as cancer, it is likely that the same disease can be achieved by many different combinations of aberrant proteins, and that this particular combination differs from patient to patient. These issues, and especially the difficulties posed by biological variability across patients, make identification of protein biomarkers a hard problem.

To date, the above challenges have been addressed for MS/MS predominantly through new chemistry, hardware, and intelligent post-processing[198-200]. Here, we describe a promising alternative statistical method, called Significance Analysis of Mass spectrum based protein quantification (pSAM), to clean and interpret raw protein measurements from shotgun proteomic experiments. To identify proteins of statistically significant expression changes between phenotypes of interest, pSAM carries out spectrum specific *z*-tests and computes a discriminative score for each protein, which is a summary statistic from all assigned spectrums. We first apply pSAM to identify proteins that are differentially expressed between CLL patients of different progression risk levels.

Using multidimensional LC-ESI-tandem mass spectrometry with iTRAQ labeling, we profile fifteen CLL patients, five individuals of each of the three risk groups, across which more than 3,000 proteins are identified. The same individuals have also been profiled with mRNA expression arrays. We then perform comparative analyses between expression of mRNA versus protein to identify proteins of correlated mRNA expression as well as proteins of no mRNA differential expression, i.e. those who can only be detected by proteomic analyses. We further propose strategies to develop follow-up targeted proteomics analysis by leveraging the substantial information present in protein interaction networks.

6.2 MS-based shotgun proteomics with isobaric tag for relative and absolute quantification

Analogous to the shotgun sequencing approach in genomics, the most widely used method for protein identification is referred to as MS-based shotgun proteomics[201]. In a shotgun proteomic experiment, protein samples are first digested with trypsin. The labeled peptides are separated by multidimensional liquid chromatography (LC), and the resolved peptides are subjected to an electric potential, which causes a spray to be formed, leading to the desolvation and ionization of the peptides (electrospray ionization; ESI). Mass to charge (m/z) ratios are measured from peptide ions in the mass spectrometer (MS). Specific ions are selected, depending on the sampling technique used, for a physical collision into smaller fragment ions and subsequently measured in the second mass analyser in tandem mass spectrometry (MS/MS).

MS/MS fragment ion information from the second stage (MS2) contains sequence information that can be compared with sequences from in silicon digested protein sequence databases for peptide and subsequent protein identification. For binary isotopic labeling or label-free quantification, the peak areas or intensities observed in the MS spectra of eluted peptides in the first stage (MS1) are used to quantify relative peptide abundance. However, overlapping spectra of the same peptides from different samples require further analytical techniques to deconvolute the resulting spectrum and the associated protein/peptide abundances[202], increasing the likelihood of systematic errors in measurements. In addition, the binary labeling techniques add complexity to the acquired spectra and to their interpretation by introducing additional peaks into the mass spectra.

In contrast to MS1-based quantification, shotgun proteomic methods involving iTRAQ enable simultaneous identification and quantification of peptides using MS2 and permit parallel proteome analysis of more than two samples[194, 203]. iTRAQ uses four amine specific isobaric reagents to label the primary amines of peptides from four different biological samples (**Figure 6.1a**). The labeled peptides from each sample are mixed, separated using LC and analyzed using MS/MS. Different from other differential labeling techniques, the same peptide from each sample appears as a single peak in the MS1 spectrum because of the isobaric nature of these reagents. This reduces the complexity in the MS1 spectrum and thus decreases the systematic noise introduced by labeling. Upon the physical collision prior MS2, the iTRAQ-tagged peptides fragment to release reporter ions (at 114.1, 115.1, 116.1 and 117.1 m/z) and b- and y-ion series among other fragments. The peak height of the reporter ions are measured in MS2 and

used to assess relative abundance of peptides (consequently the proteins from which they are derived) (**Figure 6.1b**).

6.3 pSAM − Significance Analysis of Mass spectrum based protein quantification

pSAM is a method for identifying differentially expressed proteins using the relative abundance of two iTRAQ reporter ions acquired in a MS2 spectra. This process involves several scoring steps for both within and between MS/MS experiments. For a given MS/MS spectrum, let p be the intensity of reporter 1 and q be that of reporter 2, where each reporter is used to label a different sample. The ratio of p over q is represented as r while the product of p and q is as m. Let $\sum_j$ be the set of spectrum i mapped to the unique peptides derived from protein j. Let d_i be the log2 transform of r_i. We first calibrate d_i by its intensity dependent variation $\sigma_{d,mi}$:

$z_i = \frac{d_i}{\sigma_{d,m_i}}$, where $\sigma_{d,mi}$ is the standard deviation of the d_k's whose m_k are of similar levels as m_i.

The fold change of a specific protein between the two samples is then defined as the weighted average of the normalized reporter ratios among the spectra which have been mapped to peptides derived from the protein:

$$k_j = \frac{\sum_{i \in \sum_j} z_i}{\sqrt{|\sum_j|}}$$

If replicate experiments are performed, the protein ratios from individual experiments are then integrated into a single score to assess the significance and robustness of the observed changes:

$s_j = \frac{\frac{\left|\sum_p^n k_{j,p}\right|}{\sqrt{n}}}{\sigma_j + \sigma_{med}}$, where n is the number of experiment p detecting protein j reliably, σ_j is the standard deviation of k_j across these n runs of experiments, and σ_{med} is the median of σ_j of every protein across all the replicates. Because MS profiling is an open system and has many uncharacterized systematic noise, a protein might not be detected in every MS run and may have large variation in measurements between runs. The *s* score awards proteins with more times of detection and less variation across runs.

Figure 6.1c provides an illustration of this hierarchical scoring procedure. To estimate the *p*-values of *s* of a protein *j*, the protein score s_j is compared to the random scores sampled from a permutation test of 300 random trials across multiple MS/MS experiments. For each MS/MS experiment in a random trial, we randomly assign the intensity ratio of two reporter ions z_i to $\sum_l$ where l is a protein rather than j and then perform the same scoring procedure as that on real data. Such permutation disrupts the correlation between spectra generated from the same peptide. The score of each protein s_j is indexed on the null distribution of all random scores.

6.4 Experimental Design

Multidimensional LC-ESI-tandem mass spectrometry with iTRAQ labeling is applied to compare protein levels across three classes of mature B-cells: (a) aggressive CLL, (b) indolent CLL, and (c) normal B cells. Peripheral blood monocytes from patients were lysed and proteins were digested by trypsin. Three equal aliquots of digested protein samples, one from a patient in each of the three classes of mature B-cells, were treated each with one of distinct N-terminal iTRAQ isotopes. The three

labeled lysates are mixed together and fractionated by LC following by ESI and tandem mass spectrometry. At MS1, for each time point, we select 15 peaks of the highest signal intensity and submit the peaks into MS2. A database search is then performed using the fragmentation data from MS2 to identify the labeled peptides and hence the corresponding proteins. The fragmentation of the attached iTRAQ tags generates a low molecular mass reporter ion whose intensities quantify the relatively abundance of the peptides which they originated.

The raw MS/MS spectra are extracted and searched using Spectrum Mill (Agilent) against the International Protein Index database. Each spectrum is assigned to a unique peptide of the best match. A concatenated forward-reverse database is constructed to calculate the in-situ false discovery rate (FDR) of matched spectrum-peptide pairs. Cutoff scores are dynamically assigned for each dataset from a MS/MS experiment to maintain FDR less than 3.5%; spectrum-peptide pairs of a match score less than the cutoff are discarded. Peptides shared by multiple proteins are further removed before quantification. Relative peptide quantification is calculated as the iTRAQ reporter ion intensity ratios *r* of any two samples. Relative protein quantification is performed by pSAM on these *r*'s. In total, five patients of each B-cell class are profiled in a way that each MS/MS experiment has one patient from each class, i.e., five runs of MS/MS are performed (**Figure 6.2a**). After mapping the IPI protein IDs to NCBI Entrez gene IDs, there are 2,556 unique proteins detected at least twice among the five runs of MS/MS experiments.

To investigate the differences in protein identification and quantification obtained in isobaric tagging experiments compared with those observed in other shotgun proteomic experiments, we also perform another two sets of MS1-based protein quantification experiments in additional ten aggressive and ten indolent CLL samples using binary $^{16}O/^{18}O$ labeling. One set of the MS1-based quantification is a pooled design where in a MS run, five aggressive samples are mixed together and then labeled with ^{16}O. The mixture of five indolent samples is labeled with ^{18}O. The MS experiment is repeated three times on aliquots from these two labeled lysates and also repeated in a dye-swapped fashion for another three times (i.e., the aggressive samples are labeled with ^{18}O whereas the indolent samples are labeled with ^{16}O). Therefore, this set contains six MS experiments (the left panel in **Figure 6.2b**). Another set is using individually labeled samples in a MS experiment. Every MS experiment measures peptides from only one aggressive sample labeled with ^{16}O and one indolent sample labeled with ^{18}O; a dye-swapped experiment is performed using vice versa labeling of the same samples. In all, five aggressive and five indolent samples are used in this set so ten MS experiments are performed (the right panel in **Figure 6.2b**). The LC-ESI step and the database search procedure are the same as those used in the above iTRAQ experiments, but the relative quantification of $^{16}O/^{18}O$ labeling is calculated using spectra counts on MS1 spectra as contrast to peak intensities on MS2 spectra for iTRAQ quantification. Spectra count ratio is digitized into either up (ratio $\geq$ 1), down (ratio $\leq$ 1), or undetected (ratio = 0) categories. The sum of the digitized spectra count ratio is used to estimate the level of protein expression changes across multiple runs.

6.5 pSAM removes selection biases on protein abundance or size

Ratio of peak heights of two reporter ions *r*, where proteins differing by more than an arbitrary cut-off value in abundance are considered to be differentially expressed, is commonly used in analyzing proteomics data. Such a concept of fold changes is intuitive and easy to interpret. However, simple ratios have a tendency in selecting proteins of naturally lower abundance regardless the phenotypes of interest (**Figure 6.3a**), as the phenomenon seen in microarray data[204]. Moreover, in MS experiments, proteins of larger size are digested into more peptides on average than proteins of smaller size, thus also more likely to be classified as differentially expressed by the ratio method (**Figures 6.3b**). As well discussed in microarray literature[204], a normalization on the variance of the ratios is needed to remove such an intensity-dependent bias (**Figure 6.3c**) in analyses of differential expression.

To stabilize the intensity-dependent variation on ratios, pSAM first normalizes the log ratio of two reporter ions based on the log product of the two intensities, similar to the lowess normalization procedure on microarray data[204] (**Figure 6.3f**). As a result, we can see that pSAM has a more even chance in selecting proteins over a wide range of abundance and size, reducing false positives for peptides with naturally low intensities or smaller size and false negatives for peptides with naturally high intensities or larger size (**Figures 6.3d-e**). **Figure 6.3g** shows that most proteins have correlated ratios and pSAM scores while some are only considered to be differentially expressed by either method.

6.6 Protein markers selected by pSAM are more functionally correlated and coherent with gene expression changes

Proteins work in a concert to carry out cellular functions. Any molecular alterations effecting cell phenotypes should involve proteins functioning in the same signaling cascade or complexes. Among the 2,556 proteins which are detected at least twice in the five runs of MS/MS experiments on CLL samples, for aggressive versus indolent CLLs, the top 100 proteins of largest expression changes ranked by pSAM scores are more likely to be participating in the same pathways than those ranked by simple ratios (**Figure 6.4a**; p-value ≤ 0.05 for protein functional enrichment analyzed using NCBI DAVID). We also examine the functional enrichment of the two sets of top proteins ranked by spectra counts, each from the pooled and individual MS1-based quantification (**Figure 6.2b**). In terms of MS1 methods, the pooled experimental design seems to be an effective means to reveal robust and common differences between phenotypes as compared to the individual design (**Figure 6.4a**). However, the pooled design is not able to recover an established CLL protein marker ***ZAP70*** (**Figure 6.4b**). In all, the proteins selected by pSAM on MS2 measurements are more functionally correlated and have a larger overlap with the proteins selected by MS1 quantification, as compared to the other three methods (**Figure 6.4c**).

We further check the correlation between our mRNA and protein quantification data. For a fair comparison, two microarray methods, simple ratios and SAM[205], are used to select differentially expressed genes between aggressive and indolent CLLs on mRNA level and the results are compared with the proteins selected by ratios and pSAM,

respectively. When compared to the top gene markers, both ratios and pSAM have a low rate of false positives which are defined as top proteins with opposite direction of gene expression changes (**Figure 6.4d**). Although the expression level between genes and proteins are not well correlated in our data as noted elsewhere[200], the proteins selected by pSAM have a more consistent change at both the transcription and translation levels than those selected by simple ratios.

6.7 *HIP1R* and *CD74* are promising novel protein markers of CLL progression risk

Among the 2,556 proteins detected twice of high confidence in our iTRAQ data, pSAM identifies about 1,700 proteins significantly differentially expressed in B-cells between CLL patients and normal people, but only fourty proteins different between aggressive versus indolent CLLs (**Figure 6.5a**; FDR = 30% in **Figure 6.5b**). This finding supports the notion researchers have observed from gene expression data that these two forms of CLLs are truly the same disease but might be with some subtle difference in pathology.

To demonstrate the detection power of pSAM, we select two example proteins, ***HIP1R*** and ***CD74***, out of the thirty proteins that are differentially expressed not only between aggressive versus indolent CLLs but also between CLLs and normal B-cells. ***HIP1R***, Huntingtin-interacting protein 1, has an averaged 6.6 fold increase at protein level in our aggressive CLLs when compared to the indolent samples (pSAM *p*-value = 0.003). ***HIP1R*** has been shown to be capable of stabilizing receptor tyrosine kinases on cell surface that may contribute to alteration in cell growth and survival[206, 207]. B-cell receptor signaling pathway, essential to B-cell survival, is activated through an antigen

binding to the cell surface receptor tyrosine kinases[208] and is found to be enhanced in aggressive CLLs[124].

To study its potential contribution in stabilization of BCR signaling in aggressive CLLs, we probe the protein expression of ***HIP1R*** in the original samples used in MS/MS experiments as well as additional 34 CLL patients by using low-throughput immuno-blotting (**Figure 6.6a**). We found that the patients of ***HIP1R***-expressing CLL cells have a shorter treatment-free survival as compared to those of no ***HIP1R*** expression (*p*-value = 0.006 in **Figure 6.6b**). The same patient cohort is not able to be separated into two distinct risk groups by ***ZAP70*** expression (*p*-value = 0.944 in **Figure 6.6c**). **Figure 6.6d** suggests that potential regulation of ***HIP1R*** induction can be a gene dosage effect, given its chromosomal location 12q24. Trisomy 12 is a known indicator of poor prognosis in CLL[209]. Another explanation can be the down-regulation of miR-29 observed in aggressive CLLs[210, 211], which has a target site in the 3'-end of ***HIP1R*** gene.

In contrast to ***HIP1R***, ***CD74*** has only 1.4 fold increase in aggressive CLLs. However, it is selected by pSAM as a protein marker indicative of CLL progression because of its consistent expression elevation across multiple replicates (*p*-value = 0.005). To capture this subtle expression change, we quantify the protein expression of ***CD74*** by flow cytometry. This slight expression elevation in aggressive CLLs is validated by a higher percentage of CD74 positive CLLs (**Figure 6.7a**). In a validation cohort of 73 new CLL patients, we find that ***CD74*** expression is highly associated with ***ZAP70*** expression (**Figure 6.7b**) and can be used to separate the cohort into two groups of different risk levels of disease progression (**Figure 6.7c**).

Activation of cell-surface ***CD74*** by ***MIF*** initiates a survival cascade to induce ***NF-kB*** activation, contributing to CLL tumorgenesis[212]. A phase I clinical trial of anti-CD74 therapy in B cell malignancies is undergoing (a collaboration between Cornell University and Immunomedics). Previously, CLL lymphocytes are shown to over-express CD-74 as compared to normal B-cells but the expression level is not different between different risk groups[212]. We re-analyze the data in Binsky *et al.*[212] and find that ***CD74*** is indeed differentially expressed at both gene and protein levels between early-staged ***ZAP70***-positive versus ***ZAP70***-negative patients (**Figure 6.7d**).

6.8 Differential protein expression is coherent with protein interactions

By mapping the differentially expressed proteins between aggressive and indolent CLLs (176 proteins of pSAM p-value $\leq$ 0.05) on the protein interaction network in **Figure 2.2**, we find several subnetworks formed by direct interactions between these proteins. Three subnetworks of size greater than four are not possibly formed by random (Figure 6.8a; p-value $\leq$ 0.05 as compared to a distribution of sizes of random subnetworks formed by 100 trails of randomly withdrawn 176 proteins). This might suggest dependence between protein differential expression and the physical interactions between proteins.

To systematically examine this correlation, we define a test statistic r as the number of protein pairs where the two proteins are both differentially expressed. Among the 21,150 protein pairs where the two proteins are directly interacting or have at least one common interacting partner, there are 2,085 protein pairs where the two proteins are also differentially expressed (i.e., r = 2,085). The null hypothesis here is that in any

random 21,150 protein pairs we can also find a similar number of protein pairs of differentially expressed proteins. Estimated by 1000 sets of random 21,150 protein pairs, the null distribution of r has a mean at 1623.8 and a standard deviation of 2.5 (**Figure 6.7b**). Indexed on this null distribution, r = 2,085 from the real interacting protein pairs has a p-value close to 0, suggesting that a protein is more likely to be differentially expressed if its interacting partner has differentia expression.

6.9 Integrated strategies for targeted proteomic

Given the coherence observed between protein differential expression and physical interactions, we test how much the information on the differential expression of interacting partners can help the prediction for an unmeasured protein. A leave-one-out cross-validation technique (LOOCV) is used to evaluate the prediction performance, where one protein is withheld for test and the rest is used to train a predictor. In the protein interaction network, 1,516 proteins are measured at least twice in our iTRAQ experiments and of ≥ 1 interacting partner also measured in the data. We take the top 150 proteins of differential expression ranked by pSAM as the gold positive set and the bottom 150 proteins of least differential expression as the gold negative set for the LOOCV test.

We predict an unknown protein as differentially expressed if a portion of its interacting partners is measured as differentially expressed. The precision of the prediction (true positive rate) increases as the majority of the interacting partners of an unknown protein is differentially expressed (the red curve in **Figure 6.9a**). We next test the prediction based on gene expression changes. As expected, gene expression is a

strong predictor of protein expression (the blue curve in **Figure 6.9a**). However, the best prediction is made from the consensus of the predictions made based on gene expression alone and based on differential protein expression of the interacting partners, achieving a high true positive rate over 80% (the green curve in **Figure 6.9a**).

We further evaluate the feasibility of this integrated method by predicting the differential expression of the rest 7,877 unmeasured proteins in the interaction network. The enrichment of known cancer genes is used as the metric to benchmark the prediction performance (the list of cancer genes is assembled as described in **Chapter 4**). The combination of comprehensive gene expression data, limited protein expression data and protein interaction networks outperforms the other two prediction methods, predicting a set of differentially expressed proteins where ~30% are cancer related genes (**Figure 6.9b**). Some example unmeasured proteins but predicted as differentially expressed include ***TP53***, ***BRCA2***, ***ERBB2***, ***PIK3R1***, and ***IKBK1***.

6.10 Acknowledgement

Chapter 6, in full, is a re-editing of the materials currently being prepared for submission for publication. Chuang, H.Y., Shen, Z., Rassenti, L., Ideker, T., Kipps, T., Briggs, S. Dissecting disease progression of Chronic Lymphocytic Leukemia using an integrated quantitative proteomic and genomic analysis, **in preparation**. The dissertation author was the primary investigator and author of this paper.

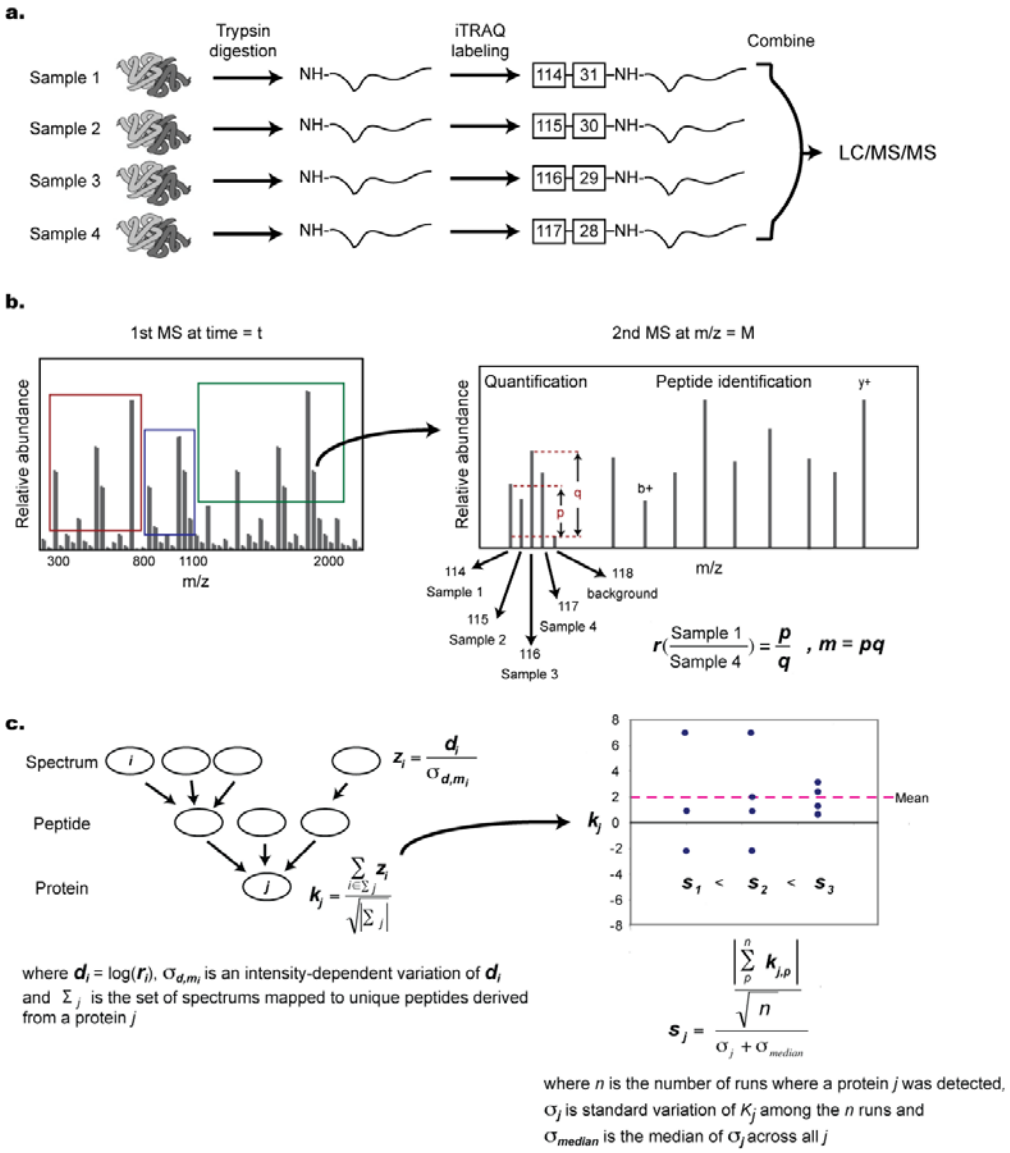


Figure 6.1. A schematic diagram of a MS/MS experiment using iTRAQ and pSAM.

(**a**) A iTRAQ molecule attaches to the N-terminus of a peptide. The proteins in each sample are first digested into peptides using trysin and labeled with individual iTRAQ reagent. Then the labeled peptides from several samples are pooled together and fractionated by LC following by MS/MS. (**b**) At MS1, for each time point, we select 15 peaks of the highest signal intensity from 3 m/z windows and submit the peaks into MS2. A database search is then performed using the fragmentation data from MS2 to identify the labeled peptides and hence the corresponding proteins. The fragmentation of the attached iTRAQ tags generates a low molecular mass reporter ion that is used to relatively quantify the peptides which they originated. (**c**) The relative abundance of two reporter ions is first normalized to stabilize the intensity dependent variation. We sum up the normalized relative abundance of two samples at spectra level to peptides and then to a whole protein. The ratios of a protein across multiple MS/MS runs are then summarized to stabilize the between-run variation (see **Section 6.3** for details).

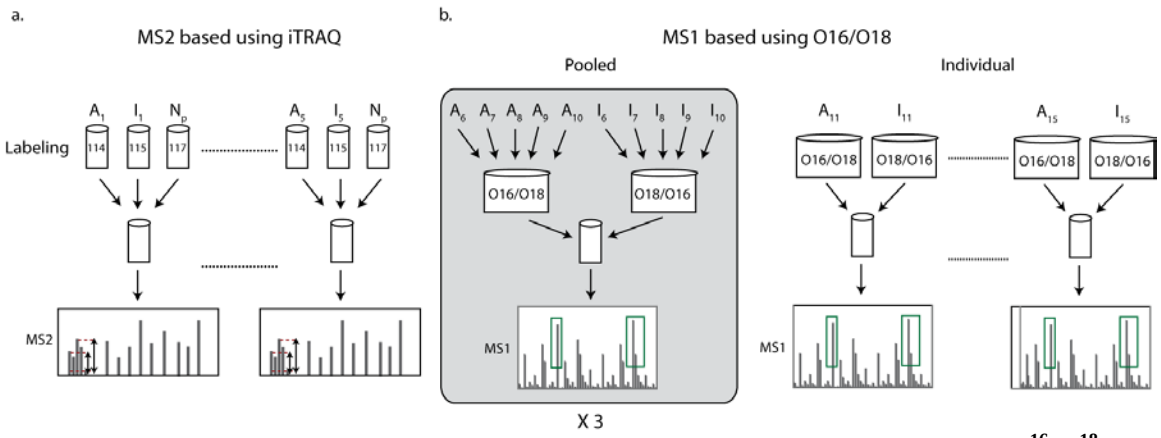


Figure 6.2. Experimental designs of MS/MS experiments using (a) iTRAQ and (b) $^{16}O/^{18}O$.

(**a**) The peptide samples from aggressive CLL, indolent CLL and normal B cells are labeled with iTRAQ isotopes of reporter ion m/z at 114, 115, and 117, respectively. Five sets of iTRAQ experiments are performed and the peptide abundance is quantified using peak heights of the iTRAQ ions in MS2 spetra. (**b**) Two designs of MS1 quantification are analysed, pooled versus individual. The pooled design is using pooled samples of five aggressive/indolent CLLs whereas the individual design has every CLL sample labeled individually with a oxygen isotopes. Multiple replicates and dye swapping are performed.

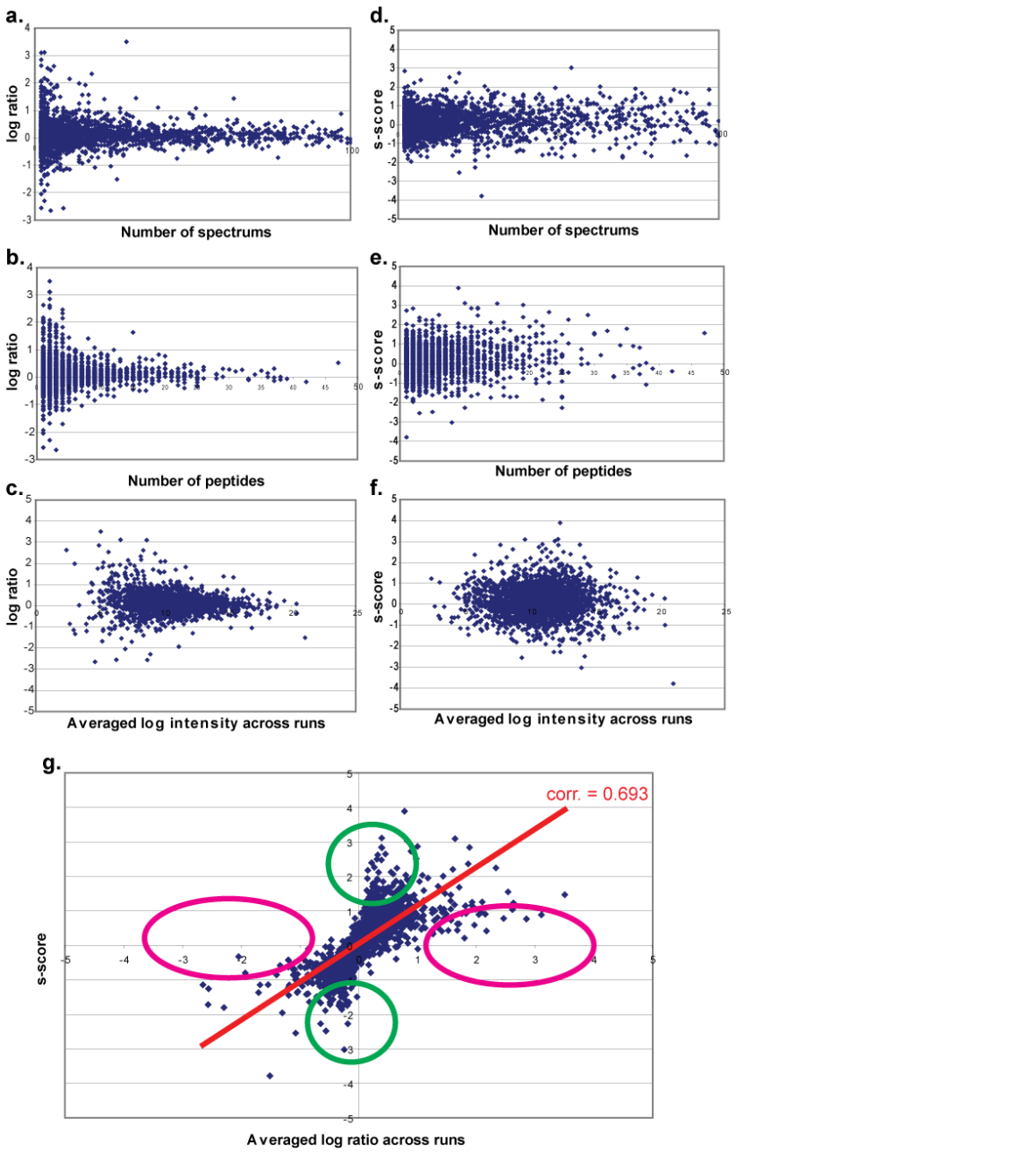
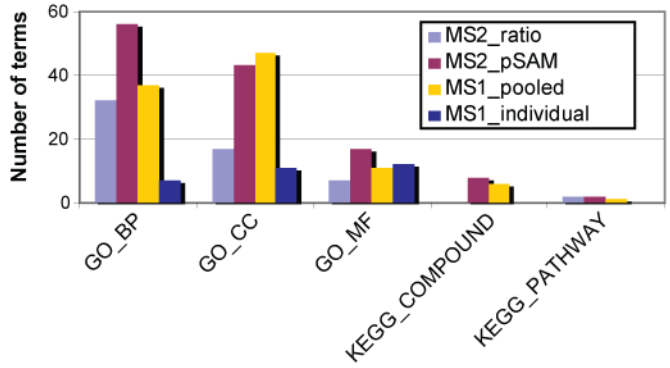


Figure 6.3. Protein selection bias when using simple ratios or pSAM.

Every data point is a protein in these scatter plots. The y-axis is either averaged log ratio across multiple runs (**a**-**c**) or pSAM *s*-score (**d**-**f**). The x-axis denotes natural protein abundance (number of spectrums), protein size (number of peptides), or the multiplication of the peak heights of the two reporter ions (log intensity). The pink areas circle out the proteins whose ratios are dramatic but *s*-scores are close to 0 whereas the green areas contain the proteins whose ratios are not interesting but *s*-scores are significant.

a.



b.

Methods	rank of ZAP70
MS2_Conventional	16
MS2_pSAM	23
MS1_pooled	405*
MS1_individual	30*

c.

Overlap	MS2_Conventional	MS2_pSAM	MS1_pooled
MS2_pSAM	25		
MS1_pooled	4	15	
MS1_individual	1	2	5

d.

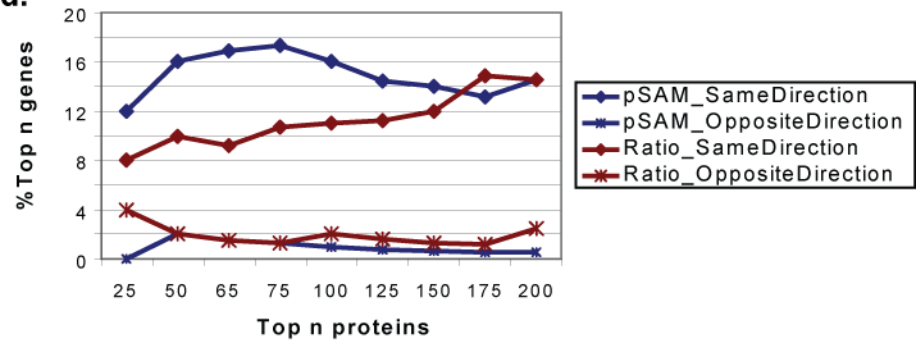


Figure 6.4. Comparisons of top proteins selected by MS1 and MS2 quantification.

(**a**) Functional enrichment of top 100 proteins by NCBI DAVID analysis. Bars chart the number of significantly associated functional groups from the databases listed on the x-axis. (**b**) Rank of ***ZAP70*** in each method. (**c**) Overlap of top 100 proteins selected by each method. Numbers are the proteins in common. (**d**) Overlap between top gene markers from mRNA data and top protein markers from MS data. Curves marked as "same direction" chart the percentage of top proteins whose gene expression are with the same direction of changes as protein expression between phenotypes whereas curves marked as "opposite direction" chart the percentage of those proteins whose gene expression change is inconsistent with the protein expression.

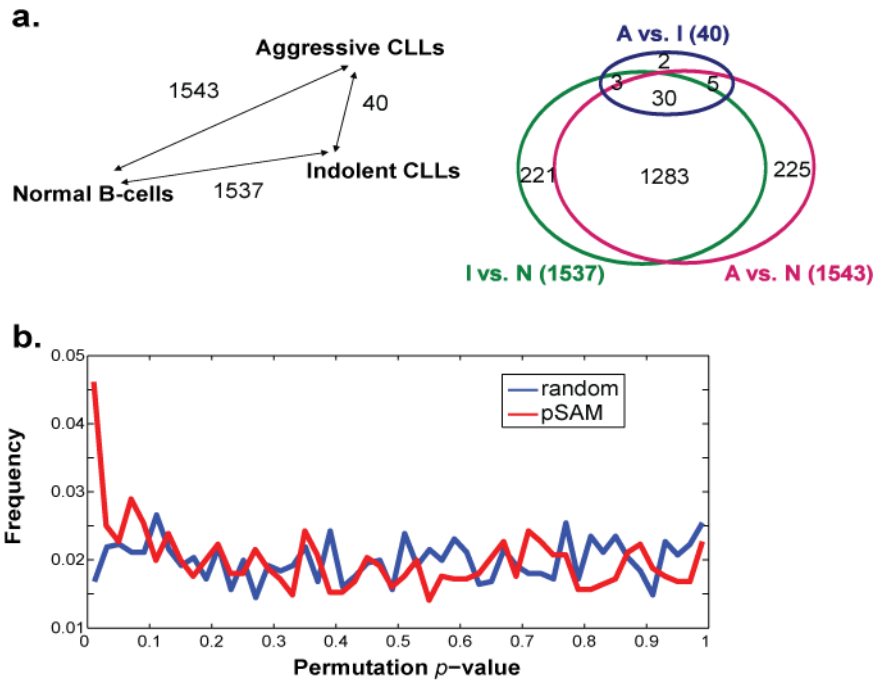


Figure 6.5. Differentially expressed proteins between the three classes of mature B-cells.

(**a**) Left panel: Number are the proteins selected by pSAM at FDR = 30% for distinguishing different classes of mature B-cells. Right panel is a Venn diagram of the overlap and uniqueness of the protein groups in the left panel. (**b**) Distributions of protein *p*-values (red) from real iTRAQ data versus random data (blue).

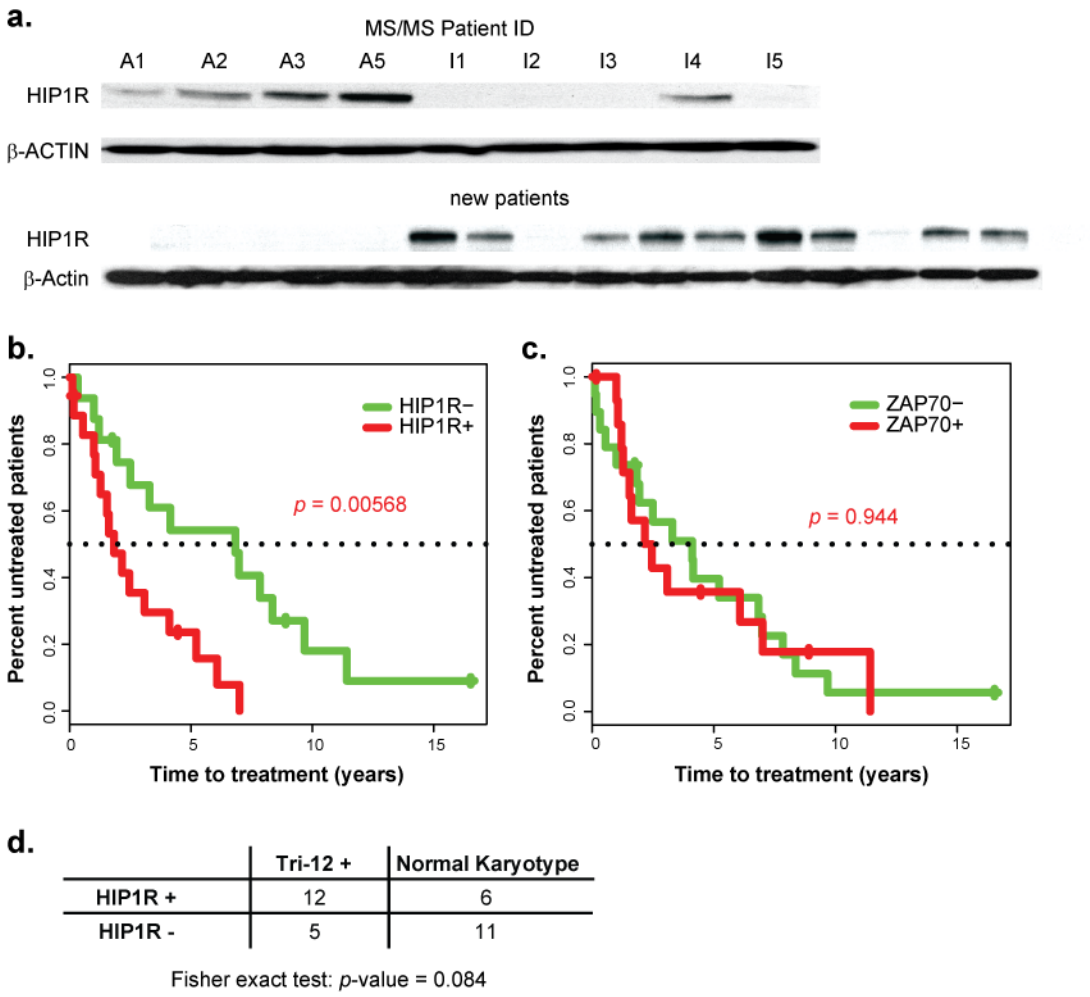


Figure 6.6. HIP1R protein expression in newly diagnosed patients.

(**a**) Immuno-blots of ***HIP1R*** and ***β-actin*** in the patient cohort used in the MS/MS experiments and in a new cohort for validation. (**b-c**) Survival analyses of the two risk groups of the new validation cohort defined by ***HIP1R*** protein expression or ***ZAP70*** protein expression. (**d**) Correlation between ***HIP1R*** expression and trisomy 12 in the new patient cohort.

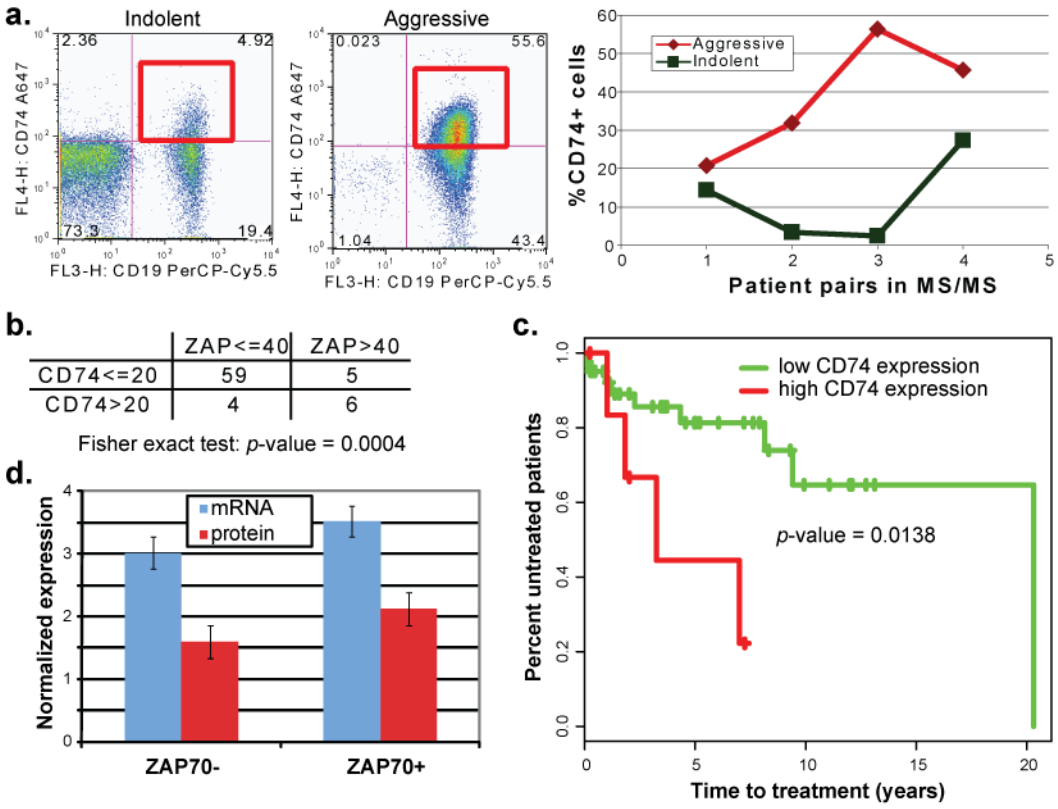


Figure 6.7. CD74 protein expression in newly diagnosed patients.

Flow cytometry of ***CD74*** in the patient cohort used in the MS/MS experiments (**a**) and in a new cohort for validation (**b**). (**c**) Survival analysis of the two risk groups of the validation cohort defined by CD74 expression (%CD74+ cells > 20 is called high expression). (**d**) Re-analysis of the qPCR and immune-blotting data of the patients of Rai stage I or II in Binsky *et al.*[212].

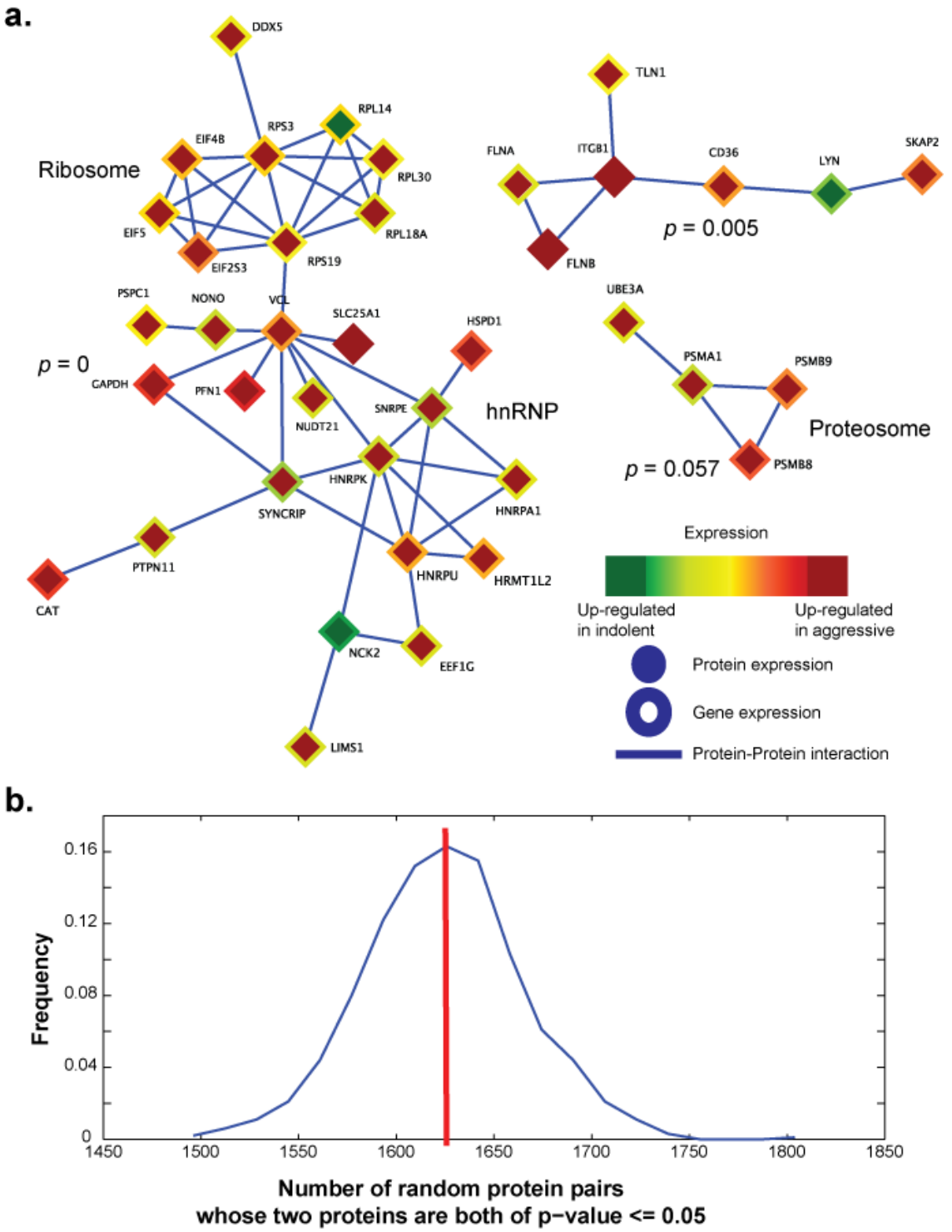


Figure 6.8. Correlation between protein differential expression and physical interactions.

(**a**) Significant subnetoworks of differentially expressed proteins between aggressive versus indolent CLLs (pSAM p-value ≤ 0.05). (**b**) The null distribution of r in random protein pairs (see **Section 6.8**).

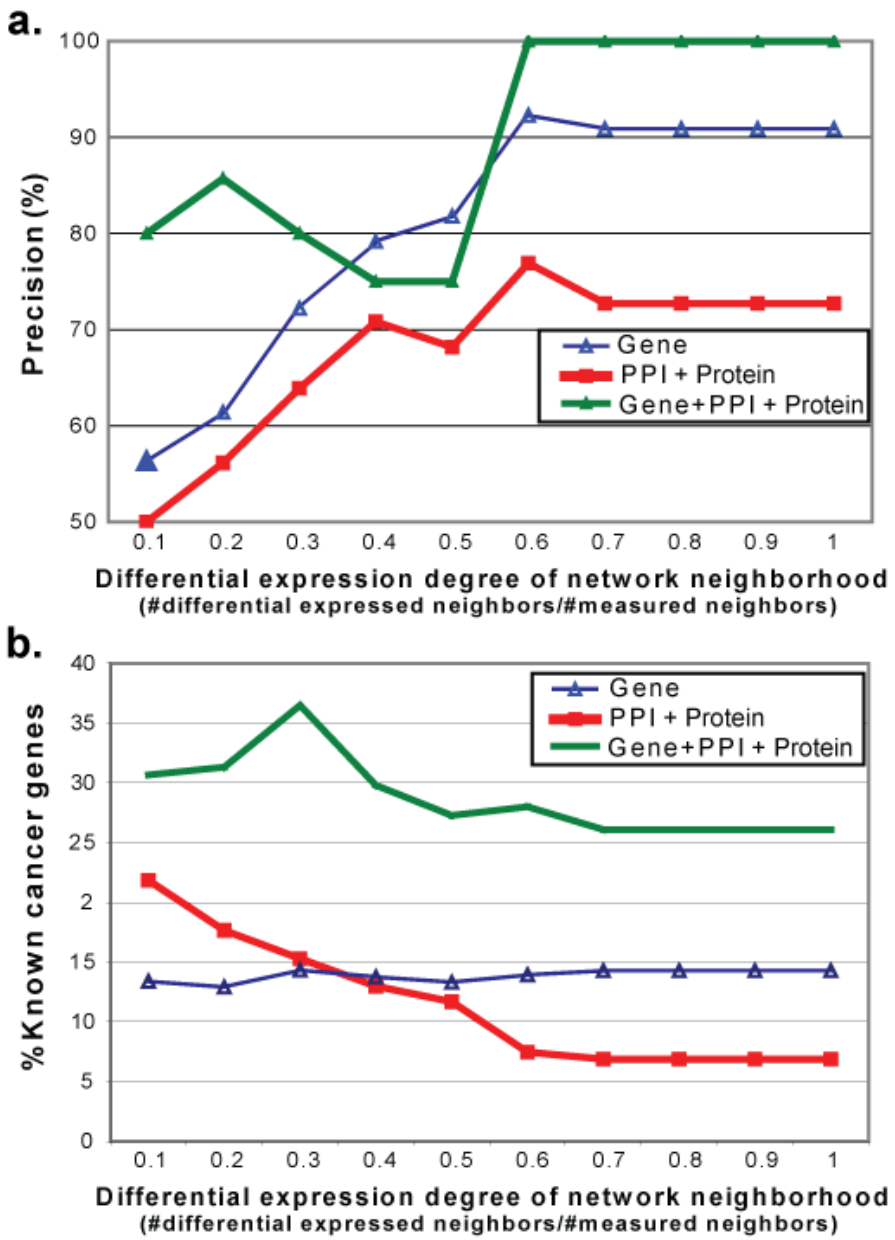


Figure 6.9. Strategies for targeted proteomics.

(**a**) LOOCV evaluation of prediction on differential expression of unmeasured proteins. The y-axis represents the precision of the prediction and the x-axis represents the parameter values used in the classifiers (see **Section 6.9**). (**b**) The cancer gene enrichment of predicted differentially expressed proteins in each method. The y-axis represents the percent known cancer genes of the prediction and the x-axis represents the parameter values used in the classifiers (see **Section 6.9**).

7. CONCLUSION

Deciphering the pathways that give rise to cancer development and progression involves dealing with the complex nature of the disease as it manifests itself in humans. Much progress experimentally is being made to develop high-throughput technologies in order to address this complexity from a holistic view of biological systems. To understand the collaborative effects involving interactions of multiple genes within complex pathways and their contributions towards a phenotype, computational approaches are needed to simulate known biological systems for hypothesis testing and account for what is not known from high-throughput data sets. Iterative systems approaches, connecting experimental data to computational approaches, provide deeper insights into human in vivo tumor behavior, and improve the development of better diagnostic and prognostic biomarkers for cancer. In this study, we discuss several promising techniques for coping with these challenges. In all, the proposed methods are not limited to cancer research, but are systems for studying complex genetic diseases.

7.1 Pathway-based molecular diagnosis

Protein interaction networks are a powerful framework for summarizing prior biological knowledge. In **Chapter 2**, we first demonstrate the utility of incorporating protein interactions into gene expression profiling for better understanding of cancer and development of precise prognostics. Gene expression profiles from large cohorts of patients are mapped to a huge human protein interaction network. A search over this network is performed to identify discriminative subnetworks which could be used to assess the aggravation risk of a patient.

In **Chapter 3**, we identify protein interaction subnetworks with coherent expression patterns of their component genes, which can distinguish the samples of patients which developed distant metastasis after surgery from those that did not. The subnetwork markers are more accurate in the classification of metastasis than the previous predictors and in the meantime provide potential insights into the molecular mechanisms involved in metastasis formation. In addition, the vast majority of the selected subnetworks contain highly interconnected proteins encoded by genes that are previously identified for breast cancer susceptibility but absent in the established predictors, because they are not detected as differentially expressed.

In **Chapter 4**, we develop a network-aided gene expression classification procedure to classify CLL progression status based on the activity of entire signaling pathways or protein complexes, rather than the expression levels of individual genes or proteins. The identified subnetwork markers that can reliably predict the relative risk for disease progression from the time of sample collection outperform previous risk-assessment markers on the prognosis of newly diagnosed patients. We also find evidence that the subnetwork signatures may evolve over time, suggesting degenerate pathways implicated in cancer evolution.

Besides de novo pathway reconstruction from protein interaction networks, we infer dynamic activity of canonical pathways for each patient based on the gene expression levels of condition-responsive gene in **Chapter 5**. We demonstrate an advanced diagnostic based on the activity of entire pathways to improve the prognosis

accuracy of lung cancer, prostate cancer, breast cancer and acute leukemia and to facilitate the discovery of molecular mechanisms underlying disease.

In summary, projection of gene expression profiles onto pathway databases or interaction networks is proving to be a powerful approach for understanding disease. The goal is to identify biomarkers not as lists of individual genes or proteins, but as functionally-related groups of genes or proteins whose aggregate expression accounts for the phenotypic differences between the different populations of patients. Conventional gene-expression analyses associate each individual of hundreds of genes with important parameters of cancer, but the functional correlations between the genes and the mechanisms of their control are largely unknown. Unlike those single gene markers analyzed in isolation, the "diagnostic pathway markers" provide a strong biological interpretation for why the expression profile is associated with a particular type of disease. The diagnostic pathways are more reproducible than single genes and can improve the prediction accuracy of disease states. Dissecting cancer transcriptome in a network-assisted view have identified features that are more closely tethered to the biology of disease progression, allowing us to observe mechanisms governing cancer evolution. At present, the success of network-based pathway identification and classification supports the notion that cancer is indeed a "disease of pathways"[1, 213], and that the keys for understanding at least some of these pathways are encoded in the protein network.

We believe this work is biologically significant and methodologically novel because of its integration of network analysis with microarray classification. Both of

these areas have individually received a great deal of attention in systems biology. However, we are aware of few if any studies which use protein networks to improve classification accuracy. Intuitively, one's ability to classify gene expression profiles should indeed be improved by introducing new and relevant biological information.

On the other hand, this study is preliminary and much work is needed before the approach can be translated into advanced diagnostics. One useful direction will be to complement expression and pathway connectivity with other large-scale data sets including information on genetic perturbations, epigenetic regulation, signal transduction, transcriptional control, protein expression, metabolism and so on. Integrating other types of genome-wide data holds further promise for determining cause and effect relationships within and between the degenerate pathways. The network method PINNACLE described here can serve as a systematic and integrative framework for incorporating heterogeneous data and outputting predictions.

Multiple improvements can also be made in terms of the computational perspective of PINNACLE. Currently, the iterative exploration of the high-dimensional space of all possible protein subnetworks seeded at all nodes of a highly branched and interconnected network is achieved by the use of a greedy algorithm. It is clearly that the identified subnetworks may be locally optimal at the discriminative power of disease classification and can be highly overlapped with each other. With the ever-increasing high-performance computing power becoming available within laboratory workstations, one can imagine that some global optimization algorithms, such as simulated annealing[42], can be incorporated to improve the subnetwork search procedure.

Moreover, a simple average is used to summarize member gene expression into subnetwork activity in this study. A caveat is the cancel-out between up-regulated and down-regulated member genes when inhibition occurs within a pathway. A more sophisticated mathematical function should be devised to capture the coherent dysregulation between genes and their suppressors. Besides expression coherence between interacting proteins, another interesting direction is to focus on the dysregulation on the interaction itself. For example, Taylor *et al.*[40] proposed to measure changes in interaction "coherence" between member genes in a subnetwork under different phenotypes. The interaction coherence in a sample was defined using the difference in expression of the central "hub" gene in a subnetwork with each of its interacting partners.

Finally, it is clear that many real and functionally-relevant interactions are missing in current protein-protein interaction datasets. Human interaction databases are growing dramatically through systematic yeast two-hybrid and transcriptional interaction screens[214]. Increased coverage, quality, and variety of human protein interaction data will, in turn, enable further opportunities for molecular characterization of human disease. Further insights can be expected from re-analysis of the same diseases as the data increase in coverage and quality.

7.2 Protein biomarker identification

The rise of proteomic technologies is particularly important to disease studies, because aberrations on the DNA level translate to the protein networks that perform cellular functions. **Chapter 6** addresses the cancer complexity from the proteomic angle. We develop a shotgun method to quantify protein expression correlated with progression

of CLL. Our protein markers are more functionally related, consistent with gene expression and robust across patient cohorts. We also find that protein differential expression is coherent with their interaction, enabling prediction of expression for unmeasured proteins in the shotgun experiments.

This work is the first proteome-wide study for disease progression in CLL. We believe it is also the most complete one in terms of protein coverage in cancer studies up to date. This work is biologically relevant because it identifies new prognostic protein markers that are not only associated with the pathways underlying the disease progression but also more robust than conventional methods of proteomic analysis. Furthermore, the work is novel in its integration of protein quantification with network analysis. We are aware of few if any studies which use protein networks to predict protein expression. This work demonstrates the feasibility of protein expression prediction from the integration of gene expression, protein expression and interaction network.

Due to the uncharacterized systematic errors in MS systems, the proposed MS analysis pSAM quantifies the expression changes of a protein between two samples within a single MS experiment and then summarize the changes across multiple replicates, to adjust for large across-experiment variability. However, it is desirable to be able to analyze data across multiple MS experiments since it allows studies to incorporate larger sample sizes, obtaining more accurate estimates of biological effects and thus having more power to detect meaningful differences. One future direction on pSAM development is first to sort out experimental factors associated with the noise as comparing relative measurements. Methods for controlling for sources of experimental

variation, such as analysis of variance (ANOVA)[215], can then be used to correct for experimental variability for quantative MS method using iTRAQ.

Our CLL proteomic study still concentrate on the selection of single protein markers. It is clear that single biomarkers are unlikely to provide information about tissue type and malignant transformation throughout the various stages of tumor development and progression, as we discuss very much through this dissertation. However, assembling such a panel of protein biomarkers is quite a challenge given the incomplete protein coverage due to the detection limits of shotgun MS-based proteomics. More recently, targeted MS proteomics workflows have been introduced to allow the selective detection and quantification of predetermined peptide ions, which are analogous to mRNA profiling using DNA microarrays[199, 216, 217]. A complement targeted MS experiment after shotgun MS proteomics is expected to increase the proteome coverage of the disease, if we know what peptides to measure in advance. Our work here in prediction of protein differential expression from an integration of comprehensive gene expression profiles, limited protein expression profiles from shotgun proteomics, and protein interaction network can help narrow down the large probing space in the following targeted MS proteomics. Increased coverage in quantitative MS methods will allow the possibility of joint learning between protein expression and other types of genome-wide data, in turn presenting opportunities in elucidating pathway dysregulation at translational level.

REFERENCES

1. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

2. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8 (2008).

3. Ding, L. et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75 (2008).

4. Parsons, D.W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807-12 (2008).

5. Mullighan, C.G. et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. *Nature* **446**, 758-64 (2007).

6. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**, 343-72 (2001).

7. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-4 (2002).

8. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

9. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).

10. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).

11. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9 (2000).

12. Chaurand, P., DaGue, B.B., Pearsall, R.S., Threadgill, D.W. & Caprioli, R.M. Profiling proteins from azoxymethane-induced colon tumors at the molecular level by matrix-assisted laser desorption/ionization mass spectrometry. *Proteomics* **1**, 1320-6 (2001).

13. Chuang, H.-Y., Hofree, M. & Ideker, T. A Decade of Systems Biology. *Annual Review of Cell and Developmental Biology* **26**, In press (2010).

14. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-7 (1999).

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

16. Lamb, J. et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929-35 (2006).

17. Tomlins, S.A. et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644-8 (2005).

18. Yao, J. et al. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. *Cancer Res* **66**, 4065-78 (2006).

19. Adler, A.S. et al. Genetic regulators of large-scale transcriptional signatures in cancer. *Nat Genet* **38**, 421-30 (2006).

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

21. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171-8 (2005).

22. Sotiriou, C. & Piccart, M.J. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat Rev Cancer* **7**, 545-53 (2007).

23. Wang, Y. et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671-9 (2005).

24. Calvano, S.E. et al. A network-based analysis of systemic inflammation in humans. *Nature* **437**, 1032-7 (2005).

25. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140 (2007).

26. Doniger, S.W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).

27. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. & Krawetz, S.A. Global functional profiling of gene expression. *Genomics* **81**, 98-104 (2003).

28. Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol* **4**, e1000217 (2008).

29. Pavlidis, P., Qin, J., Arango, V., Mann, J.J. & Sibille, E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* **29**, 1213-22 (2004).

30. Tian, L. et al. Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* **102**, 13544-9 (2005).

31. Wei, Z. & Li, H. A Markov Random Field Model for Network-based Analysis of Genomic Data. *Bioinformatics* (2007).

32. Nibbe, R.K., Markowitz, S., Myeroff, L., Ewing, R. & Chance, M.R. Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer. *Mol Cell Proteomics* **8**, 827-45 (2009).

33. Ulitsky, I., Karp, R.M. & Shamir, R. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. *Lecture Notes in Computer Science* **4955**, 347 (2008).

34. Breslin, T., Krogh, M., Peterson, C. & Troein, C. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics* **6**, 163 (2005).

35. Li, L. et al. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics* **85**, 16-23 (2005).

36. Ma, X., Lee, H., Wang, L. & Sun, F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* **23**, 215-21 (2007).

37. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

38. Tuck, D.P., Kluger, H.M. & Kluger, Y. Characterizing disease states from topological properties of transcriptional regulatory networks. *BMC Bioinformatics* **7**, 236 (2006).

39. Vert, J.P. & Kanehisa, M. Extracting active pathways from gene expression data. *Bioinformatics* **19 Suppl 2**, ii238-44 (2003).

40. Taylor, I.W. et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**, 199-204 (2009).

41. Chen, J. & Yuan, B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* **22**, 2283-90 (2006).

42. Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A.F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1**, S233-40 (2002).

43. Sharan, R. et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A* **102**, 1974-9 (2005).

44. Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-71 (2003).

45. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol* **6**, R40 (2005).

46. Rual, J.F. et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8 (2005).

47. Stelzl, U. et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-68 (2005).

48. Chung, C.H., Bernard, P.S. & Perou, C.M. Molecular portraits and the family tree of cancer. *Nat Genet* **32 Suppl**, 533-40 (2002).

49. Asyali, M.H., Colak, D., Demirkaya, O. & Inan, M.S. Gene Expression Profile Classification: A Review. *Current Bioinformatics* **1**, 55-73 (2006).

50. Quackenbush, J. Microarray analysis and tumor classification. *N Engl J Med* **354**, 2463-72 (2006).

51. Cheang, M.C.U., van de Rijn, M. & Nielsen, T.O. Gene Expression Profiling of Breast Cancer. *The Annual Review of Pathology: Mechanisms of Disease* **3**, 67-97 (2008).

52. Butte, A.J. & Kohane, I.S. Creation and implications of a phenome-genome network. *Nat Biotechnol* **24**, 55-62 (2006).

53. Ramaswamy, S., Ross, K.N., Lander, E.S. & Golub, T.R. A molecular signature of metastasis in primary solid tumors. *Nat Genet* **33**, 49-54 (2003).

54. Rhodes, D.R. et al. 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**, 9309-14 (2004).

55. Tomlins, S.A. et al. Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* **39**, 41-51 (2007).

56. Draghici, S. et al. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res* **31**, 3775-81 (2003).

57. Gene Ontology Database (GO), http://www.geneontology.org/.

58. Kim, R.D. & Park, P.J. Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol* **5**, R70 (2004).

59. Pavlidis, P., Li, Q. & Noble, W.S. The effect of replication on gene expression microarray experiments. *Bioinformatics* **19**, 1620-7 (2003).

60. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* **103**, 5923-8 (2006).

61. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).

62. Monti, S. et al. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood* **105**, 1851-61 (2005).

63. Segal, E., Friedman, N., Koller, D. & Regev, A. A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**, 1090-8 (2004).

64. Efroni, S., Schaefer, C.F. & Buetow, K.H. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE* **2**, e425 (2007).

65. Svensson, J.P. et al. Analysis of gene expression using gene sets discriminates cancer patients with and without late radiation toxicity. *PLoS Med* **3**, e422 (2006).

66. Glinsky, G.V., Berezovska, O. & Glinskii, A.B. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J Clin Invest* **115**, 1503-21 (2005).

67. Matyus, L. Fluorescence resonance energy transfer measurements on cell surfaces. A spectroscopic tool for determining protein interactions. *J Photochem Photobiol B* **12**, 323-37 (1992).

68. Fields, S. & Sternglanz, R. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet* **10**, 286-92 (1994).

69. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).

70. Cusick, M.E. et al. Literature-curated protein interaction datasets. *Nat Methods* **6**, 39-46 (2009).

71. Rain, J.C. et al. The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211-5 (2001).

72. Butland, G. et al. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433**, 531-7 (2005).

73. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).

74. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).

75. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).

76. Li, S. et al. A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-3 (2004).

77. Giot, L. et al. A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-36 (2003).

78. Bouwmeester, T. et al. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* **6**, 97-105 (2004).

79. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

80. Brown, K.R. et al. NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* **25**, 3327-9 (2009).

81. Hu, Z., Snitkin, E.S. & DeLisi, C. VisANT: an integrative framework for networks in systems biology. *Brief Bioinform* **9**, 317-25 (2008).

82. Alfarano, C. et al. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **33**, D418-24 (2005).

83. Bader, G.D., Betel, D. & Hogue, C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**, 248-50 (2003).

84. Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, D428-32 (2005).

85. Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-10 (2006).

86. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* **34**, D535-9 (2006).

87. Tourassi, G.D., Frederick, E.D., Markey, M.K. & Carey E. Floyd, J. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics* **28**, 2394-2402 (2001).

88. Fisher, R.A. Applications of “Student's” distribution. *Metron* **5**, 90-104 (1925).

89. Corder, G.W. & Foreman, D.I. Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach (Wiley, New Jersey, 2009).

90. Rodgers, J.L. & Nicewander, W.A. Thirteen ways to look at the correlation coefficient. *The American Statistician* **42**, 59-66 (1988).

91. Chen, P.Y. & Popovich, P.M. Correlation: Parametric and nonparametric measures (Sage Publications, Thousand Oaks, CA, 2002).

92. Collett, D. Modelling survival data in medical research (Chapman & Hall/CRC, 2003).

93. Gersten, M. et al. An integrated systems analysis implicates EGR1 downregulation in simian immunodeficiency virus encephalitis-induced neural dysfunction. *J Neurosci* **29**, 12467-76 (2009).

94. Weigelt, B., Peterse, J.L. & van 't Veer, L.J. Breast cancer metastasis: markers and models. *Nat Rev Cancer* **5**, 591-602 (2005).

95. Alizadeh, A.A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-11 (2000).

96. Ben-Dor, A. et al. Tissue classification with gene expression profiles. *J Comput Biol* **7**, 559-83 (2000).

97. Symmans, W.F., Liu, J., Knowles, D.M. & Inghirami, G. Breast cancer heterogeneity: evaluation of clonality in primary and metastatic lesions. *Hum Pathol* **26**, 210-6 (1995).

98. Pavlidis, P., Lewis, D.P. & Noble, W.S. Exploring gene expression data with class scores. *Pac Symp Biocomput*, 474-85 (2002).

99. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**, D277-80 (2004).

100. Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E. & Vert, J.P. Classification of microarray data using gene networks. *BMC Bioinformatics* **8**, 35 (2007).

101. Mendelsohn, A.R. & Brent, R. Protein interaction methods--toward an endgame. *Science* **284**, 1948-50 (1999).

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

103. Nikitin, A., Egorov, S., Daraselia, N. & Mazo, I. Pathway studio--the analysis and navigation of molecular networks. *Bioinformatics* **19**, 2155-7 (2003).

104. van de Vijver, M.J. et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* **347**, 1999-2009 (2002).

105. Tourassi, G.D., Frederick, E.D., Markey, M.K. & Carey E. Floyd, J. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics* **28**, 2394-2402 (2001).

106. Goebel BD, Z., Hagenauer, J. & Mueller, J.C. An Approximation to the Distribution of Finite Sample Size Mutual Information Estimates. *IEEE Internatioanl Conference on Communications* (2005).

107. Mak, H.C., Daly, M., Gruebel, B. & Ideker, T. CellCircuits: a database of protein network models. *Nucleic Acids Res* **35**, D538-45 (2007).

108. Agresti, A. Categorical data analysis (New York: Wiley, 1990).

109. Swets, J.A., Dawes, R. & Monahan, J. Psychological Science Can Improve Diagnostic Decisions. *Psychological Science in the Public Interest* **1** (2000).

110. Chang, C.-C. & Lin, C.-J. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (2001).

111. Molecular Signatures Database (MSigDB), http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2.

112. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. *Nat Rev Cancer* **4**, 814-9 (2004).

113. Kang, Y. et al. Breast cancer bone metastasis mediated by the Smad tumor suppressor pathway. *Proc Natl Acad Sci U S A* **102**, 13909-14 (2005).

114. Bachman, K.E. et al. The PIK3CA gene is mutated with high frequency in human breast cancers. *Cancer Biol Ther* **3**, 772-5 (2004).

115. Campbell, I.G. et al. Mutation of the PIK3CA gene in ovarian and breast cancer. *Cancer Res* **64**, 7678-81 (2004).

116. de Jong, M.M. et al. Genes other than BRCA1 and BRCA2 involved in breast cancer susceptibility. *J Med Genet* **39**, 225-42 (2002).

117. Lymberis, S.C., Parhar, P.K., Katsoulakis, E. & Formenti, S.C. Pharmacogenomics and breast cancer. *Pharmacogenomics* **5**, 31-55 (2004).

118. Online Mendelian Inheritance in Man, O.T. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), {6/30/2006}. http://www.ncbi.nlm.nih.gov/omim/

119. Sjoblom, T. et al. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. *Science* (2006).

120. Hallek, M. et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. *Blood* **111**, 5446-56 (2008).

121. Fais, F. et al. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest* **102**, 1515-25 (1998).

122. Damle, R.N. et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* **94**, 1840-7 (1999).

123. Hamblin, T.J., Davis, Z., Gardiner, A., Oscier, D.G. & Stevenson, F.K. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* **94**, 1848-54 (1999).

124. Chen, L. et al. Expression of ZAP-70 is associated with increased B-cell receptor signaling in chronic lymphocytic leukemia. *Blood* **100**, 4609-14 (2002).

125. Crespo, M. et al. ZAP-70 expression as a surrogate for immunoglobulin-variable-region mutations in chronic lymphocytic leukemia. *N Engl J Med* **348**, 1764-75 (2003).

126. Orchard, J.A. et al. ZAP-70 expression and prognosis in chronic lymphocytic leukaemia. *Lancet* **363**, 105-11 (2004).

127. Rassenti, L.Z. et al. ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia. *N Engl J Med* **351**, 893-901 (2004).

128. Rassenti, L.Z. et al. Relative value of ZAP-70, CD38, and immunoglobulin mutation status in predicting aggressive disease in chronic lymphocytic leukemia. *Blood* **112**, 1923-30 (2008).

129. Stamatopoulos, B. et al. Quantification of ZAP70 mRNA in B cells by real-time PCR is a powerful prognostic factor in chronic lymphocytic leukemia. *Clin Chem* **53**, 1757-66 (2007).

130. Bilban, M. et al. Deregulated expression of fat and muscle genes in B-cell chronic lymphocytic leukemia with high lipoprotein lipase expression. *Leukemia* **20**, 1080-8 (2006).

131. Heintel, D. et al. High expression of lipoprotein lipase in poor risk B-cell chronic lymphocytic leukemia. *Leukemia* **19**, 1216-23 (2005).

132. Huttmann, A. et al. Gene expression signatures separate B-cell chronic lymphocytic leukaemia prognostic subgroups defined by ZAP-70 and CD38 expression status. *Leukemia* **20**, 1774-82 (2006).

133. Nuckel, H. et al. Lipoprotein lipase expression is a novel prognostic factor in B-cell chronic lymphocytic leukemia. *Leuk Lymphoma* **47**, 1053-61 (2006).

134. Oppezzo, P. et al. The LPL/ADAM29 expression ratio is a novel prognosis indicator in chronic lymphocytic leukemia. *Blood* **106**, 650-7 (2005).

135. van't Veer, M.B. et al. The predictive value of lipoprotein lipase for survival in chronic lymphocytic leukemia. *Haematologica* **91**, 56-63 (2006).

136. Haslinger, C. et al. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J Clin Oncol* **22**, 3937-49 (2004).

137. Klein, U. et al. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *J Exp Med* **194**, 1625-38 (2001).

138. Pepper, C. et al. Highly purified CD38+ and CD38- sub-clones derived from the same chronic lymphocytic leukemia patient have distinct gene expression signatures despite their monoclonal origin. *Leukemia* **21**, 687-96 (2007).

139. Rosenwald, A. et al. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. *J Exp Med* **194**, 1639-47 (2001).

140. Schroers, R. et al. Combined analysis of ZAP-70 and CD38 expression as a predictor of disease progression in B-cell chronic lymphocytic leukemia. *Leukemia* **19**, 750-8 (2005).

141. Vasconcelos, Y. et al. Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes. *Leukemia* **19**, 2002-5 (2005).

142. Fernandez, V. et al. Gene expression profile and genomic changes in disease progression of early-stage chronic lymphocytic leukemia. *Haematologica* **93**, 132-6 (2008).

143. Stratowa, C. et al. CDNA microarray gene expression analysis of B-cell chronic lymphocytic leukemia proposes potential new prognostic markers involved in lymphocyte trafficking. *Int J Cancer* **91**, 474-80 (2001).

144. Kohlmann A, K.T., Rassenti LZ, Downing JR, Shurtleff SA, Mills KI, Gilkes AF, Hofmann WK, Basso G, Dell'orto MC, Foà R, Chiaretti S, De Vos J, Rauhut S, Papenhausen PR, Hernández JM, Lumbreras E, Yeoh AE, Koay ES, Li R, Liu WM, Williams PM, Wieczorek L, Haferlach T. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol* **142**, 802-807 (2008).

145. Kohlmann, A. et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the Microarray Innovations in LEukemia study prephase. *Br J Haematol* **142**, 802-807 (2008).

146. Haferlach, T. et al. Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report From the International Microarray Innovations in Leukemia Study Group. *J Clin Oncol* (2010).

147. Bader GD, D.I., Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* **29**, 242-245 (2001).

148. Lu, D. et al. Activation of the Wnt signaling pathway in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* **101**, 3118-23 (2004).

149. Danilov, A.V., Danilova, O.V., Klein, A.K. & Huber, B.T. Molecular pathogenesis of chronic lymphocytic leukemia. *Curr Mol Med* **6**, 665-75 (2006).

150. Franks SE, S.M., Arias-Mendoza F, Shaller C, Padavic-Shaller K, Kappler F, Zhang Y, Negendank WG, Brown TR. Phosphomonoester concentrations differ between chronic lymphocytic leukemia cells and normal human lymphocytes *Leukemia research* **26**, 919-926 (2002).

151. Franks, S.E. et al. Phosphomonoester concentrations differ between chronic lymphocytic leukemia cells and normal human lymphocytes *Leukemia research* **26**, 919-926 (2002).

152. Friedman, D.R. et al. A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia. *Clin Cancer Res* **15**, 6947-55 (2009).

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

154. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29**, e45 (2001).

155. Seger R, K.E. The MAPK signaling cascade. *FASEB J* **9**, 726-735 (1995).

156. Seger, R. & Krebs, E.G. The MAPK signaling cascade. *FASEB J* **9**, 726-735 (1995).

157. Dancey J, S.E. Issues and progress with protein kinase inhibitors for cancer treatment. *Nat Rev Drug Discov* **2**, 296-313 (2003).

158. Dancey, J. & Sausville, E.A. Issues and progress with protein kinase inhibitors for cancer treatment. *Nat Rev Drug Discov* **2**, 296-313 (2003).

159. Platanias LC. Map kinase signaling pathways and hematologic malignancies. *Blood* **101**, 4667-4679 (2003).

160. Platanias, L.C. Map kinase signaling pathways and hematologic malignancies. *Blood* **101**, 4667-4679 (2003).

161. Bierie, B. & Moses, H.L. TGF-beta and cancer. *Cytokine Growth Factor Rev* **17**, 29 (2006).

162. Douglas RS, C.R., Lamb RJ, Nowell PC, Moore JS. Chronic lymphocytic leukemia B cells are resistant to the apoptotic effects of transforming growth factor-beta. *Blood* **89**, 941-947 (1997).

163. Douglas, R.S., Capocasale, R.J., Lamb, R.J., Nowell, P.C. & Moore, J.S. Chronic lymphocytic leukemia B cells are resistant to the apoptotic effects of transforming growth factor-beta. *Blood* **89**, 941-947 (1997).

164. Lotz, M., Ranheim, E. & Kipps, T.J. Transforming growth factor beta as endogenous growth inhibitor of chronic lymphocytic leukemia B cells. *J Exp Med* **179**, 999-1004 (1994).

165. Carlucci, F. et al. A 57-gene expression signature in B-cell chronic lymphocytic leukemia. *Biomed Pharmacother* **63**, 663-71 (2009).

166. Kaufman, M., Rubin, J. & Rai, K. Diagnosing and treating chronic lymphocytic leukemia in 2009. *Oncology (Williston Park)* **23**, 1030-7 (2009).

167. Bair E, T.R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* **2**, e108 (2004).

168. Bair, E. & Tibshirani, R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* **2**, e108 (2004).

169. Merlo LM, P.J., Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-935 (2006).

170. Merlo, L.M., Pepper, J.W., Reid, B.J. & Maley, C.C. Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-935 (2006).

171. Saeys, Y., Inza, I. & Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* (2007).

172. Guo, Z. et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* **6**, 58 (2005).

173. Tian, B., Nowak, D.E., Jamaluddin, M., Wang, S. & Brasier, A.R. Identification of direct genomic targets downstream of the nuclear factor-kappaB transcription factor mediating tumor necrosis factor signaling. *J Biol Chem* **280**, 17435-48 (2005).

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

175. Yeoh, E.J. et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133-43 (2002).

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

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

178. Fisher, R.A. Applications of "Student's" distribution (1925).

179. Mani, K.M. et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol* **4**, 169 (2008).

180. Guo, Z. et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* **6**, 58 (2005).

181. Gambhir, S.S. Molecular imaging of cancer with positron emission tomography. *Nat Rev Cancer* **2**, 683-93 (2002).

182. Gatenby, R.A. & Gillies, R.J. Why do cancers have high aerobic glycolysis? *Nat Rev Cancer* **4**, 891-9 (2004).

183. Gatenby, R.A. & Gillies, R.J. Glycolysis in cancer: a potential target for therapy. *Int J Biochem Cell Biol* **39**, 1358-66 (2007).

184. Banka, C.L. et al. Estrogen induces lung metastasis through a host compartment-specific response. *Cancer Res* **66**, 3667-72 (2006).

185. Ong, S.E., Foster, L.J. & Mann, M. Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29**, 124-30 (2003).

186. Cox, J. & Mann, M. Is proteomics the new genomics? *Cell* **130**, 395-8 (2007).

187. Deutsch, E.W., Lam, H. & Aebersold, R. Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol Genomics* **33**, 18-25 (2008).

188. Hilario, M. & Kalousis, A. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform* **9**, 102-18 (2008).

189. Gulcicek, E.E. et al. Proteomics and the analysis of proteomic data: an overview of current protein-profiling technologies. *Curr Protoc Bioinformatics* **Chapter 13**, Unit 13 1 (2005).

190. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-42 (2001).

191. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-9 (1999).

192. Ong, S.E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc* **1**, 2650-60 (2006).

193. Boehm, A.M., Putz, S., Altenhofer, D., Sickmann, A. & Falk, M. Precise protein quantification based on peptide quantification using iTRAQ. *BMC Bioinformatics* **8**, 214 (2007).

194. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-69 (2004).

195. Hanash, S.M., Pitteri, S.J. & Faca, V.M. Mining the plasma proteome for cancer biomarkers. *Nature* **452**, 571-9 (2008).

196. Sun, Y. et al. Quantitative proteomic signature of liver cancer cells: tissue transglutaminase 2 could be a novel protein candidate of human hepatocellular carcinoma. *J Proteome Res* **7**, 3847-59 (2008).

197. Chen, R. et al. Quantitative proteomics analysis reveals that proteins differentially expressed in chronic pancreatitis are also frequently involved in pancreatic cancer. *Mol Cell Proteomics* **6**, 1331-42 (2007).

198. Wright, J.C. & Hubbard, S.J. Recent developments in proteome informatics for mass spectrometry analysis. *Comb Chem High Throughput Screen* **12**, 194-202 (2009).

199. Malmstrom, J., Lee, H. & Aebersold, R. Advances in proteomic workflows for systems biology. *Curr Opin Biotechnol* **18**, 378-84 (2007).

200. Gstaiger, M. & Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet* **10**, 617-27 (2009).

201. Wolters, D.A., Washburn, M.P. & Yates, J.R., 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* **73**, 5683-90 (2001).

202. Mason, C.J. et al. A method for automatically interpreting mass spectra of 18O-labeled isotopic clusters. *Mol Cell Proteomics* **6**, 305-18 (2007).

203. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-904 (2003).

204. Smyth, G.K. & Speed, T. Normalization of cDNA microarray data. *Methods* **31**, 265-73 (2003).

205. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).

206. Hyun, T.S. et al. HIP1 and HIP1r stabilize receptor tyrosine kinases and bind 3-phosphoinositides via epsin N-terminal homology domains. *J Biol Chem* **279**, 14294-306 (2004).

207. Jain, R.N. et al. Hip1r is expressed in gastric parietal cells and is required for tubulovesicle formation and cell survival in mice. *J Clin Invest* **118**, 2459-70 (2008).

208. Wienands, J. The B-cell antigen receptor: formation of signaling complexes and the function of adaptor proteins. *Curr Top Microbiol Immunol* **245**, 53-76 (2000).

209. Chiorazzi, N., Rai, K.R. & Ferrarini, M. Chronic lymphocytic leukemia. *N Engl J Med* **352**, 804-15 (2005).

210. Calin, G.A. et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* **353**, 1793-801 (2005).

211. Pekarsky, Y. et al. Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. *Cancer Res* **66**, 11590-3 (2006).

212. Binsky, I. et al. IL-8 secreted in a macrophage migration-inhibitory factor- and CD74-dependent manner regulates B cell chronic lymphocytic leukemia survival. *Proc Natl Acad Sci U S A* **104**, 13408-13 (2007).

213. Petricoin, E.F., 3rd et al. Mapping molecular networks using proteomics: a vision for patient-tailored combination therapy. *J Clin Oncol* **23**, 3614-21 (2005).

214. Kim, T.H. et al. A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-80 (2005).

215. Hill, E.G. et al. A statistical model for iTRAQ data analysis. *J Proteome Res* **7**, 3091-101 (2008).

216. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* **6**, 577-83 (2005).

217. Lange, V., Picotti, P., Domon, B. & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **4**, 222 (2008).