INTEGRATIVE ANALYSIS OF MULTI-MODALITY DATA IN CANCER

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Chao Wang

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2015

Dissertation Committee:

Dr. Kun Huang, Advisor

Dr. Raghu Machiraju, co-advisor

Dr. Umit V. Catalyurek

Dr. Charles L. Shapiro

Dr. Lori A. Dalton
Abstract

Gleaning insights of highly complex, heterogeneous cancer biology requires data collected from different levels - genetic, genomic and phenotypic. There is a high degree of diversity between individuals with a wide spectrum of clinical, pathologic, and molecular features. Traditionally in clinical settings, phenotypic data such as histopathological images are often used for diagnosis, subtyping, staging, prognosis and treatment. With the advent of new high-throughput biotechnologies, multi-modality of genomics and genetic data provide extremely valuable information for cancer research and clinical biomarker discovery. However, the challenge still remains towards the determination of causal relationship in these multi-modality data and effective integration to gain better understanding of cancer biology. In particular, molecular basis of cellular phenotypes manifest in histopathological images are unknown and remain inexplicable.

In this dissertation, I present a new analytic framework and accompanying computational methods to facilitate integrative analyses of multi-modality biomedical data. The first part of this volume describes the extraction of image features thus enabling quantitative analysis of the cellular structures. Our feature collections include texture features, previously discovered salient features and features designed to mimic the observations of a trained pathologist. In the next part, studies that establish the genotype–phenotype links using morphological features from histopathology are presented. Molecules and molecular events associated with breast cancer morphology are
discovered. In the third part, beyond pairwise correlations, I explore multivariate molecular basis of lung adenocarcinoma morphology. This study suggests that a cellular structure can be potential target in treatment of lung adenocarcinoma. Finally, the last part aims to develop computational methods that can jointly cluster cancer patient samples based on multi-modality data. These effective integrative cluster methods allow patient stratification based on both essential categorical attributes and multi-dimensional data from different sources. I demonstrate the application of these methods using datasets pertaining to breast cancer.

The proposed image processing workflows, the collection of morphological features, the analytical framework that links molecular expression to morphological measurements, and the integrative clustering methods show potential in revealing biological basis and new therapeutic targets of various types of cancer. The results from the studies indicate biologically interesting subtypes with potential biomarkers. The frameworks and methodologies presented in this dissertation can mine the large and complex collections of data to identify new comprehensive biomarkers generate new hypothesis.
Dedication

This document is dedicated to my family.
Acknowledgments

First and foremost, I would like to thank my research advisor Dr. Kun Huang for the support and guidance he has given me throughout the entire period of my work. He spent days revising my badly-written papers and provided me with scientific thinking that I am lack of as an engineering student. It has been a great pleasure to work with him. Without his help, it would be impossible for me to finish this research. I would also like to thank my co-advisor Dr. Raghu Machiraju for his unconditional help and suggestions. His intellectual and philosophical guidance always reminds me to keep thinking deep. He provided with many opportunities such as fellowships and internship during my study. He also spent many nights reviewing and revising my papers. I am grateful for his help. I thank Dr. Shapiro for his collaboration and providing wonderful resources for my study, including deep clinical knowledge. I'm grateful to Dr. Catalyurek and Dr. Dalton for being part of my dissertation committee, and for providing wonderful ideas and constructive comments. I thank Dr. Parag Mallick from Stanford for his helpful discussions and support for trip to Stanford. I would like thank Dr. Debra Zynger for her help on the pathological knowledge. I would like thank Dr. Lin Yang’s labs for collaboration with the lung adenocarcinoma project.

I'm very thankful to Dr. Yang Xiang for his willingness to help. I thank lab members Cenny Taslim, Hao Ding, Nan Meng, Qihang Li, Michael Sharpnack and Brian Arand
and all. The Ph.D. journey would be harder, if it was not support from them. The chalkboard discussions and exciting ideas are most unforgettable things during the days. I want to thank friends I know at OSU, Cong Wang, Dong Zhang, Xinyu Li and friends I know in Bay area. I would like to thank visiting scholars Drs. Zhi Han, Hongxing Yuan and Shengjun Xu for their great advice.

I'd like to acknowledge financial support from The Ohio State University and The Howard Hughes Medical Institute Grad-to-Med fellowship and Pelotonia fellowship. Most importantly, I’m very grateful to my parents for their kindness and endless support. I feel guilty for not accompanying them during my 8 years of graduate school and 5 years of undergraduate. Without their support, I would not complete my 13 years journey of study in universities. I would like to thank my fiancée Yue for standing by me through good times and bad times. I owe my family everything and I will spend the rest of life loving them and appreciating them.
Vita

2007.................................................................B.S Electrical and Computer Engineering,
Dalian University of Technology, China

2009.................................................................M.S Electrical and Computer Engineering,
Dalian University of Technology, China

2009-10 ...............................................................Software Engineer, Ericsson, Beijing, China

2013.................................................................M.S Electrical and Computer Engineering,
The Ohio State University

2010-12 ...............................................................Graduate Research Assistant, Electrical and
Computer Engineering, The Ohio State University

2010-12 ...............................................................OSU-HHMI Med-Into-Grad Fellow, The Ohio State University

2014 to present ...................................................Pelotonia Graduate Fellow, The Ohio State University
Publications


**Fields of Study**

**Major Field:** Electrical and Computer Engineering
# Table of Contents

Abstract.................................................................................................................................................. ii  
Dedication ............................................................................................................................................... iv  
Acknowledgments ................................................................................................................................... v  
Vita........................................................................................................................................................... vii  
Table of Contents .................................................................................................................................... x  
List of Tables ........................................................................................................................................... xvii  
List of Figures .......................................................................................................................................... xix  

Chapter 1 : Introduction .............................................................................................................................. 1  
  1.1 Motivation ........................................................................................................................................ 1  
  1.2 Background ..................................................................................................................................... 5  
  1.3 Thesis Statement ............................................................................................................................... 10  
  1.4 Outline of Solution ........................................................................................................................... 11  

Chapter 2 : Related Works .......................................................................................................................... 16  
  2.1 Analysis of Large Histopathological Images in Cancer Research ................................................. 16  
    2.1.1 Acquisition of Digital Histopathological Images ..................................................................... 16  
    2.1.2 Cell based Features for Histopathological Images ................................................................. 17
2.1.3 Tissue based Features for Histopathological Images .............................................. 19

2.1.4 Machine Learning Issues ............................................................................................ 20

2.1.5 Applications in Computer Aided Diagnosis ............................................................... 22

2.2 Integration with Other Data Modalities ......................................................................... 23

2.3 Resources and Tools ..................................................................................................... 27

Chapter 3 : Histopathological Image Processing and Morphological Feature Extraction 29

3.1 Overview of Image Processing and Feature Extraction Pipeline ................................. 29

3.2 Image Acquisition ......................................................................................................... 33

3.3 Preprocessing ................................................................................................................ 36

3.4 Cell Segmentation ......................................................................................................... 37

3.5 Superpixel Segmentation ............................................................................................... 39

3.6 Tissue Classification ...................................................................................................... 42

3.7 Feature Extraction ......................................................................................................... 46

3.7.1 Previously Discovered Salient Features .................................................................... 48

3.7.2 Texture Features ........................................................................................................ 57

3.7.3 Pathological Features ............................................................................................... 58

3.8 Correlations between Morphological Features .............................................................. 59

Chapter 4 : Identifying Salient Morphological Features of Triple Negative Breast Cancer using Integrative Methods ......................................................... 62
4.1 Introduction ..................................................................................................................... 63

4.2 Materials and Methods ................................................................................................. 65
  4.2.1 Datasets ................................................................................................................... 65
  4.2.2 Integrated Biomarker Identification Workflow ...................................................... 68
  4.2.3 TCGA and OSU Image Preprocessing and Segmentation ................................. 69
  4.2.4 Characterizing Cellular Morphological Features of TNBC Samples ............ 70
  4.2.5 Morphometric Analysis for Cell Nuclei ............................................................... 71
  4.2.6 Morphological Feature Extraction of Tumor and its Microenvironment ....... 72
  4.2.7 Correlations between Gene Expression and Tissue Morphology ................... 74
  4.2.8 Survival Analysis .................................................................................................... 74

4.3 Results and Discussion ................................................................................................. 75
  4.3.1 Translational Discover of Survival-related Morphological Features by Cancer-related Genes ........................................................ .. 75
  4.3.2 Metagenes with Strong Correlations to Morphological Features ....................... 76
  4.3.3 Survival of TNBC based on Morphological Biomarkers ................................. 77
  4.3.4 Survival Analysis of the Identified Biomarkers on Multiple Public Datasets.. 79

4.4 Summary ......................................................................................................................... 84

Chapter 5 : Integrative Genomic Analysis of Cancer Cell Morphology Regulatory Protein
........................................................................................................................................ 85
5.1 Introduction .................................................................................................................. 85

5.2 Methods and Materials .............................................................................................. 87

5.2.1 Image Processing and Feature Extraction ............................................................. 89

5.2.2 Protein Co-expression Network Analysis and Differentially Expressed Proteins ................................................................................................................................. 90

5.2.3 Survival Analysis of the Identified Biomarkers on Multiple Public Datasets .. 90

5.2.4 DNA Copy Number Variance Data Processing ..................................................... 91

5.3 Results .......................................................................................................................... 91

5.3.1 Patient Stratification based on Imaging Features .................................................. 91

5.3.2 Differential Expressed Proteins .............................................................................. 93

5.3.3 Survival Analysis of the Identified Proteins.......................................................... 95

5.3.4 Alterations of the DNA.......................................................................................... 96

5.3.5 STUB1 Protein Identified to be Associated with Morphology .............................. 97

5.4 Summary ...................................................................................................................... 98

Chapter 6 : Multivariate Model of Morphology and Transcriptome ............................... 100

6.1 Introduction ................................................................................................................ 100

6.1.1 Background ........................................................................................................... 100

6.1.2 Motivation ............................................................................................................. 102

6.2 Methods and Materials ............................................................................................. 104
6.2.1 Integrated Analysis Pipeline ................................................................. 104

6.2.2 Image and Genomic Data Collection .................................................... 105

6.2.3 Data Preprocessing and Imaging Feature Extraction ............................... 105

6.2.4 Associations between Morphology and Transcriptomes based on Lasso
Regression ........................................................................................................ 108

6.2.5 Identification of Survival-Related Image Features .................................... 108

6.3 Results .......................................................................................................... 109

6.3.1 Survival-related Image Features and Gene Cluster ................................. 109

6.3.2 Correlations between Image Features and Gene Clusters ....................... 110

6.3.3 Cellular Cilium Movement Genes Correlate with Morphology ............... 112

6.3.4 Prognostic Validation .............................................................................. 113

6.3.5 Alteration of the Cilium Genes ................................................................. 114

6.4 Summary ...................................................................................................... 116

Chapter 7 : Integrative Stratification via a Two-step Consensus Clustering ......... 117

7.1 Introduction .................................................................................................. 117

7.2 Methods and Materials .............................................................................. 119

7.2.1 Dataset .................................................................................................... 119

7.2.2 Workflow ................................................................................................ 120

7.2.3 Integrative Molecular Stratification using Canonical Correlation Analysis .. 122
8.3.1 Comparative Performance of MRCPS ................................................. 148
8.3.2 Clinical Evaluation based on Prognosis Prediction ............................. 149
8.3.3 Biological Evaluation ........................................................................ 153
8.4 Summary ............................................................................................. 156

Chapter 9: Conclusions and Future Works ................................................. 158

References ............................................................................................... 164

Appendix A: Tables .................................................................................. 186

Appendix B: Derivation of optimization in Algorithm 1 in Chapter 8 .......... 198

Appendix C: Proof of Convergence in Algorithm 1 in Chapter 8 .................. 200
List of Tables

Table 3.1: TCGA data portal statistics................................................................. 35
Table 3.2: Statistics of the image dataset and detailed training and testing cases. ........ 44
Table 4.1: Demographics summary of the OSU TNBC cohort. .............................. 67
Table 4.2: Examples of morphological features of cancer tissue images. ................. 73
Table 4.3: Examples of enriched Gene Ontology and human phenotype terms of the
metagenes strongly associated with morphology in Figure 4.3. ......................... 77
Table 5.1 The demographics of CPTAC breast cancer patients. ........................... 87
Table 5.2: Discovered protein networks that show differential expressions between the
two identified groups, and their characteristics. ............................................. 94
Table 6.1: Prognostic values of various image features in discover dataset............. 110
Table 6.2: Gene clusters showing strong correlation with texture image feature tfcm, and
their Gene Ontology terms and enriched cytobands. ..................................... 111
Table 6.3: Cilium genes are discovered to have high associations with top survival-
related image features – tfcm................................................................. 113
Table 7.1: Prognostic mRNAs and miRNAs used in the integrative molecular clustering.
....................................................................................................................... 120
Table A.1: Complete Epithelial Feature List. .................................................... 186
Table A.2: Complete Stromal Feature List. ...................................................... 190
Table A.3: Metagenes that identified to be associated with image features in Triple
Negative Breast Cancer........................................................................................................ 191

Table A.4: Subtype-Specific Genes for Subtype 1. ................................................................ 196

Table A.5: The Demographics of TCGA breast cancer subset used in Chapter 7 and
Chapter 8........................................................................................................................................... 197
List of Figures

Figure 1.1: Paradigm of translational research for integrative genomics research in cancer. ................................................................. 3
Figure 1.2: Biological systems multi-modality data .................................................. 7
Figure 3.1: Example images of the benign breast histology. .................................. 30
Figure 3.2: Imaging processing workflow in this study. ........................................ 32
Figure 3.3: Nuclei segmentation within each region of interest. ............................ 39
Figure 3.4: Two examples of superpixel segmentation of histological images. ....... 42
Figure 3.5: Local binary patterns (LBP) are defined in [20]. ................................. 45
Figure 3.6: Classified tissue superpixels .............................................................. 46
Figure 3.7: Feature hierarchy applied to this study .............................................. 48
Figure 3.8: Illustration of the definitions of inner border and outer border of an image object .................................................................................. 52
Figure 3.9: The surrounding area of and image object v defined by its bounding box with a distance of \(d = 1\). .......................................................... 53
Figure 3.10: Direct neighbors \(N_v(d)\) and the neighborhood within perimeter \(d\) of image object \(v\). ............................................................... 54
Figure 3.11: Asymmetry for a 2D image object .................................................... 56
Figure 3.12: Scatterplots of features demonstrating their correlations. ............... 60
Figure 3.13: Scatterplot of the numbers of fibroblast in stromal tissue (y-axis) and the numbers of cancer cells in epithelial tissue (x-axis).................................................................. 61

Figure 4.1: The workflow of cross-datasets feature discovery and validation. .................. 69

Figure 4.2: The workflow of the histopathological image analysis.................................. 71

Figure 4.3: Pairwise correlation heatmap between metagene expression and morphology of tissue in TCGA discover set. .......................................................................................... 77

Figure 4.4: Kaplan-Meier survival curves of prognostic model in OSU TNBC TMA. .... 79

Figure 4.5: The Kaplan-Meier curves for the metagene that shows the highest correlation with the image features. ........................................................................................................... 80

Figure 4.6: An example of the TCGA slide image with its extreme values of predictive image features. .......................................................................................................................... 82

Figure 5.1: The integrative analysis workflow used in this study. ................................. 89

Figure 5.2: Hierarchical clustering of breast cancer patients based on imaging features. Patient samples (row) in TCGA data subset were labels using their barcodes................. 91

Figure 5.3: Nuclei size in comparison. ............................................................................. 93

Figure 5.4: Kaplan–Meier survival curves of prognostic model on multiple public breast cancer datasets.. .................................................................................................................. 96

Figure 5.5: Scatter plots of proteins and the DNA copy number variance in protein cluster 1.................................................................................................................................................. 97

Figure 5.6: Characteristics of protein network 1. ........................................................... 98

Figure 6.1: Top 'high-correlated' pairwise correlations between eigengene expression and morphological features of LUAD.. ......................................................................................... 103
Figure 6.2: Schematic overview of constructing gene networks and analyzing correlations between gene networks and morphological features. .................................................. 105

Figure 6.3: Kaplan–Meier survival curves of prognostic model on two public lung adenocarcinoma datasets............................................................................................................. 114

Figure 6.4: Mutation plot of cilium-related genes in TCGA lung adenocarcinoma cases. ........................................................................................................................................... 115

Figure 7.1: Schematic plot for patient Stratification proposed in this chapter. .............. 122

Figure 7.2: The mutual information of the consensus clustering with original labels under different choices of k. ........................................................................................................ 126

Figure 7.3: Single type of molecular data clustering based stratification versus Integrative molecular subgrouping based on the 28 prognostic genes and 7 prognostic miRNAs using CCA. .................................................................................................................................................. 127

Figure 7.4: Kaplan–Meier plots for classification of the whole TCGA population. ..... 130

Figure 8.1: Integration of molecular expression data with clinically-defined patient stratification. ........................................................................................................................................ 139

Figure 8.2: Plotted values of NMI of MRCPS and other methods for different values of k. ........................................................................................................................................... 148

Figure 8.3: Prognostic power of different patient stratification methods.. ............... 151

Figure 8.4: Prognostic power of MRCPS.. .......................................................................................... 153

Figure 8.5: A gene network identified from the high-risk group......................... 155

Figure 8.6: Top diseases and functions identified from the subtype-specific genes using IPA. ........................................................................................................................................... 156
Chapter 1 : Introduction

With the rapid advances of acquisition technologies in biomedical research, such as DNA microarray, massive parallel sequencing and high-throughput imaging, large volumes of different types of biomedical data have been generated. These heterogeneous data provide a potential to enhance our understanding of diseases and to promote the development of new treatment schemes by many folds. However, the large variety of the data poses great challenges towards the manipulation, mining and visualization of the useful information hidden in the data. In particular, there is a need to develop new analytics framework to mine the large and complex collections of data to identify new comprehensive biomarkers, novel subtypes and therapeutic strategies. This dissertation presents a new analytical framework to address the problem of integrative analysis of multi-modality biomedical data. The proposed workflows and algorithms can be used to discover novel relationships between different data types, identify integrative subtypes of cancers, and generate new biomedical hypothesis.

1.1 Motivation

Cancer is a complex disease. There is a high degree of diversity among tumors from different individuals or even within the same patient. In many cancers, there are certain subtypes, which manifest different tumor aggressiveness, clinical outcomes and response to therapeutic treatments [1], [2]. In order to detect the cancer heterogeneity and provide
more accurate treatments to achieve personalized care, in clinical settings, clinicians often collect pathological images of tumor specimens and radiological images of the organ from cancer patients for purposes of diagnosis, subtyping, staging, prognosis and treatment. At the same time, basic scientists are leveraging multiple types of “omics” data to identify genetic and genomic markers for subtyping. However, there is a gap in translational research between the tissue/cellular level phenotypes observed in clinics (bed-side) and the molecular level genotype and phenotype data observed in labs (bench-side). To bridge the gap between the bed-side and the bench-side, I developed an integrative genomics framework in this dissertation. The translational paradigm for integrative genomics research is demonstrated in Figure 1.1.
Translational research in cancer aims to apply findings from the laboratory to improve patient diagnosis, prognosis and prediction of response to therapy. On the bench-side, multiple-omics data are being generated, while on the bedside clinicians are using clinical data and imaging data. However, the correlations between data being generated on both sides are unknown. And, integrative biomarkers need to be developed to comprehensively analyze cancer.

To understand the relationship between clinical phenotyping data and the molecular level “omics” data, a critical question is

“What are the relationships between clinical and omics data?”

If there are relationships between phenotypes observed in the clinics and molecular events observed in labs, a further question is how the multiple types of molecular data (mRNA expression, miRNA expression, etc.) correlate with cellular phenotypes such as
morphology, since cellular phenotypes are often regulated by multiple molecules. Furthermore, based on multi-modality data, can we develop computational models to identify integrative biomarkers, and thus novel subtypes of cancer? This dissertation is an attempt to explore answers to these following three fundamental but important questions in translational research:

Q1: How are pathological and genomics features related in cancers?

Q2: What are the genomic markers associated with an individual morphological feature?

Q3: How can the multi-modality data help to identify new subtypes and biomarkers in cancers?

In an effort to approach these questions, in this dissertation, a comprehensive histopathological image analysis pipeline is built to quantitatively measure cellular phenotypes of tumor. This computational imaging pipeline extracts multiple quantitative morphological features from histopathological images, allowing us to explore correlations between these cellular phenotypes and molecular phenotypes in breast cancer. This exploration is an effort to answer Q1. To address Q2, a multivariate model is constructed to further study the associations between multiple molecular data with cellular features. Finally, to address the integration of multi-modality data (Q3), integrative clustering algorithms are proposed and applied to identify cancer subtypes. While these methods are developed from various case studies for specific cancers such as
breast and lung cancers, the framework is nevertheless a foundation for general integrative genomic studies for many diseases.

1.2 Background

Cancer Heterogeneity

Both genetically and phenotypically, cancers are heterogeneous diseases. There is a high degree of diversity between and within tumors as well as individuals. Inter-tumor heterogeneity, which is defined by variation presents between individuals with the same tumor type [3], can lead to inaccurate diagnosis, prognosis and prediction of clinical outcomes. Research into understanding and characterizing this phenomenon can generate new insights of the mechanisms of disease as well as personalized treatment for the patients.

Cancer heterogeneity does not only exist between cancers, even within the same type of cancer, tumors show a variety of molecular characteristics. Taking breast cancer as an example, mutations in genes such as TP53 and BRCA1 as well as expression levels in ER, PR and HER2 have been used as biomarkers to classify and identify subtypes of breast cancer. High-throughput molecular profiling technologies enable further subtyping breast cancer into four major molecular subtypes: Luminal A, Luminal B, basal-like and HER2 positive type. Such classification of cancer subtype is determined using molecular and genetic information from tumors. However, even within subtype of cancer, there can be highly diverse groups. For example, triple negative breast cancer (TNBC) display six unique molecular patterns and oncologies, including 2 basal-like (BL1 and BL2), an
immunomodulatory (IM), a mesenchymal (M), a mesenchymal stem-like (MSL), and a luminal androgen receptor (LAR) subtype [4].

Cancer heterogeneity is also displayed at tissue and cellular level phenotypes. Traditionally, cellular heterogeneity in cancer samples is examined by pathologists who review the histopathological slide images. Diagnosis, subtyping, staging, and treatment strategies are often derived based on the pathologist’s evaluation of the tumor. At the same time, surgical strategies often rely on radiological images (e.g., MRI, mammogram), which have been studied intensively in quantitative analysis [5], [6]. Compared to radiological images, the quantitative characterization of cellular morphology in whole tissue slides and tissue microarrays (TMAs) is more challenging, computationally, given the large size of these images and large variation among them.

In clinically practice, different molecular and histological subtypes of cancer can display very distinct outcomes and response to therapies. New diagnostic methods and novel biomarkers could deliver a more detailed molecular analysis of cancer, giving clinicians key indicators for optimum treatment. Reproducible sub-classification of cancer based on novel biomarkers is biologically and therapeutically important. Eventually, implementing cancer genome sequencing in clinical trials and clinical oncology will greatly increase accuracy of diagnosis, prognosis and prediction of cancer.

Despite the fact that cancer heterogeneity exists on all levels: genetic, epigenetic, genomic, proteomic, morphological and clinical; comprehensive research on integrative data analysis is still lacking. The challenge remains on how to effectively integrate
multiple modalities of data into integrative computational models that can be used to gain new insights into cancer biology and mechanism, discover new prognosis biomarkers, and predict responses to treatments. Specifically, methods of systems biology that can identify the associations between cellular phenotypes and molecular events will lead to new insights and knowledge about cancer heterogeneity.

**Multi-Modality Cancer Data**

Advances in biomedical technologies have facilitated the generation of data at every level of biological systems - including DNA sequence data, RNA expression levels, methylation patterns, other epigenetic markers, proteomics and histopathological data, etc. Examples of these data are shown Figure 1.2.

![Multi-Modality Cancer Data Diagram](image)

**Figure 1.2:** Biological systems multi-modality data from the genome, epigenome, transcriptome, proteome to higher levels of phenotypes including morphology (Reproduced from [7]).

Currently many large-scale databases have been established towards the goal of the comprehensive genomic characterization of cancer. For example, the Cancer Genome Project, the International Cancer Genome Consortium (ICGC), and The Cancer Genome Atlas (TCGA) collect cancer data from various types and allow public access, enabling
comprehensive analyses of almost all common cancer types [8]–[11]. However, most of these studies combine data from a single source and biomarkers identified from difference types of data rarely display substantial overlap [12]. Therefore, there is an urgent need for effective computational methods that allow biomedical researchers to achieve reproducible biomarkers using multiple types of “omics” and clinical data.

**Morphology in Histopathology**

Imaging can also help elucidate cancer pathways and better understand cancer progressions because of its rich context of spatial information and morphological phenotypes. The recent advances in imaging acquisition, including high-resolution microscopy, have enabled automated analysis of cancer heterogeneity. The spatial information and application in clinics gives histopathology images great advantage to uncover the different molecular events happen in different compartments of the tumor. For example, in the study by Beck et al. [13], stromal features from microscopic images show stronger prognostic power than those extracted from tumor cells. However, imaging data are different from “omics” and clinical data – the original information is presented in images which cannot be directly used as quantitative features, instead morphological characteristics have to be extracted and quantified. Therefore, in this dissertation, morphological analysis of the histopathological images is emphasized and integrated with other types of data.

**Integration of Multi-Modality Data**
Genetic alterations, genomic changes and cellular morphological characterization, along with clinical phenotypes, are all pivotal information about cancer. Study of interplay within each level of data has been studied for decades [14]–[16]. For instance, large scale protein-protein interaction networks have been established to empower the current knowledge of disease pathogenesis. To gain more knowledge of cancer in term of understanding the biological mechanisms, we have to study the interactions between different levels of genotypes and phenotypes. There have been many efforts on developing computational models to integrate data from different levels, which has become one of the most rapid-growing fields in systems biology [12]. However, the high-dimensionality (possibly millions) of ‘omics’ data poses great challenges on building statistical and mathematical models. Many existing models usually fail when the dimensionality of the data becomes larger than the sample size [17].

By definition, integrative genomic analysis refers to the analysis of at least two different types of genomic data [12]. This problem is similar to information integration in the machine learning field. In fact, there have been methods in analyzing and integrating information from multi-modality data in machine learning. One of them is multi-information ensemble developed using information theory [18], where diversity and disagreement between data modalities are measured. Data fusion is another research area where data from multiple types of sensors have to been analyzed comprehensively [19]. In the context of multi-source data clustering, there are mainly two categories of methods. The first category is consensus clustering, where one desires to generate a
‘consensus’ from the multiple partitions (often called base clusterings or component clusterings) of a set of data points and each individual partition is drawn from an unknown feature space [20]. In each individual consensus clustering scheme, an optimal fit of data is obtained by solving an optimization problem with certain constraints. Another category of approaches that solve the problem of clustering multi-modality data is multi-view clustering, which uncorrelated mixture distributions of data are clustered [21]–[26].

However, biological data integration is different from other types of data. The complexity of interactions and the variation of driving molecular events require specific formulation of the problem and novel solutions. There have been many studies on integrative genomics in the recent years, but there are few efforts on integration of genomics data with histopathological images. The rich spatial information and phenotypes from histopathological images need to be well leveraged in the integrative models.

1.3 Thesis Statement

This dissertation is developed based on the following thesis statement:

**Thesis Statement:** Effective computational integration of multi-modality cancer data such as morphology from histopathology and molecular data can generate novel hypotheses in cancer biology and lead to new insight of personalized treatment schemes and novel biomarkers for cancer. A systematic computational framework for discovering multivariate relationships between molecular events and cellular morphology manifested
in histopathology is necessary. The computational predictions and case studies yielded by the framework will offer several novel hypotheses and suggest hitherto unknown biomarkers and subtypes of cancer.

Specifically, in this dissertation I present the methodological contributions on the following research topics in attempt to answer the three questions stated in the above section:

**Cellular Morphology:** While uncovering relationships between morphology and molecular expression profiles, new biomarkers can be identified for subtyping cancer patients.

**Multivariate Model:** By building multivariate model to explore the relationships between morphological, genomic and epigenetic characteristics of cancer, new hypothesis on therapeutic targets of cancer can be generated.

**Integration of Multi-modality data:** Beyond data analytic framework for discovering correlations between genomic data and morphological data, algorithms and computational methods that effectively and efficiently integrate multi-modality data can lead to integrative subtyping of cancers with distinct clinical outcomes and response to therapeutic treatments.

**1.4 Outline of Solution**

The dissertation provides a data analytic framework to investigate correlation between cellular and molecular characteristics of cancer, and to integratively cluster the cancer
patients based on multiple types of data. The key components in the framework include the following:

**Histopathological Image Processing Pipeline and Feature Extraction**

Histopathological images of breast cancer and lung cancer are acquired from TCGA and a local database. Images from the tissue sections are digitalized using a microscope. A subsequent image preprocessing pipeline implemented by me then segments both the cell nuclei and superpixels (which are relatively homogeneous regions in the image) and classifies the superpixels to tumor (epithelial) or stromal tissues. This process is inspired by the study by Beck et al. [13]. Finally, a variety of quantitative morphological features are extracted from the epithelial and stromal tissues, respectively.

**Correlative Studies between Histopathological Features and Omics Data**

Extracted image features, which characterize the cellular morphology of the tumor, are correlated with genomic and proteomic features. In one case study, the combined biomarkers of imaging features and gene expression profiles are tested to predict triple negative breast cancer outcomes. In another case study, salient morphological features are used to stratify the patients into two groups with distinct tumor burden. Proteins expressed differentially between these groups are discovered. These correlative studies help us understand the regulators of morphology better in breast cancer.

**Multivariate-based Studies between Histopathological Features and Omics Data**

By transforming gene expressions into eigengenes of a co-expression network, the task of predicting pairwise intra-relationships of morphology and genomic signatures can
be formulated as a feature selection problem. To systematically investigate the regulatory associations between lung cancer morphology and transcriptome, a large set of biological imaging features are extracted from representative lung tissue image slides. A multivariate model of regulatory morphology is built to explore the contributions of genes to morphology. A lung adenocarcinoma case study shows the importance of certain gene networks in lung cancer development.

**Integrative Cancer Patient Stratification**

Along with essential clinical information, numerical measurements such as molecular expression can be clustered using the proposed two-step consensus clustering method. Further, a more advanced approach is presented using a case study of breast cancer. A relatively aggressive subgroup of TCGA breast cancer population is identified to have specific gene expression pattern while simultaneously have poor prognosis. The proposed method is tested to perform robust patient stratification, which can be evaluated for prognostic power and biological relevance.

**Contributions:**

1. **Computational Workflow for Morphological Analysis**

   The majority part of this dissertation involves with studying cellular morphology from microscopic histopathological images. Thus, an image processing workflow for extracting a variety of image features from tumor cells and tumor microenvironment is developed. This image processing workflow generates more than 100 morphological features of histopathology, thus allows quantitative measurements of different
compartments of a tumor. The image processing pipeline and feature extraction details are described in Chapter 3.

2. Associations between Cellular Morphology and Molecular Profiles in TNBC

By employing the imaging workflow, survival-related imaging features are discovered for TNBC. In particular, nuclei-to-cytoplasm ratio related imaging features are identified to be salient in TNBC. Then, by finding the associations between imaging features and gene expression profiles, gene biomarkers are uncovered for TNBC. Specific gene biomarkers are identified to have strong prognostic power of differentiating TNBCs. The results and methods are demonstrated in Chapter 4.

3. Associations between Cellular Morphology and Proteomics in Breast Cancer

Through the integrative imaging and proteomics workflow, protein markers are discovered to be associated with cell morphology in breast tumors. These protein markers show significant association with the tumor burden and tumor aggressiveness. In particular, STUB1/CHIP gene is identified as a regulator of tumor burden. The case study is discussed in Chapter 5.

4. Multivariate-based study of Lung Adenocarcinoma Morphology and Transcriptome

Beyond pairwise correlations, a multivariate model is built to explore the associations between transcriptomics and morphology from pathological images in lung adenocarcinoma. The model and phenotypes find essential to lung adenocarcinoma are shown in Chapter 6. Importantly, cilium genes are found to have strong associations with
salient morphological features. Validation on public datasets shows the prognostic value of the cilium genes.

5. Patient Stratification based on Consensus Clustering method

Besides morphological study of cancers, a two-step algorithmic consensus clustering technique for integrative patient stratification and case study on breast cancer are presented in Chapter 7. This two-step method incorporates molecular expression data and clinical data for obtaining better prognostic subgroups of patients. Overcoming the limitations of the two-step consensus clustering method, a novel integrative clustering algorithm – Molecular Regularized Patient Stratification (MRPS), which allows more efficient patient stratification based on clinical attributes and molecular profiles, is proposed in Chapter 8.
Chapter 2 : Related Works

The analysis of multi-modality data is an emerging field in bioinformatics and machine learning. This chapter presents a review of existing methods and problems in this direction. It opens with a discussion of the main issues in analysis of histopathological images in cancer research in Section 2.1. After this, Section 2.2 describes the principles of integrating heterogeneous data. Finally, valuable existing resources and tools are listed in Section 2.3.

2.1 Analysis of Large Histopathological Images in Cancer Research

2.1.1 Acquisition of Digital Histopathological Images

Commonly available histopathological images are digital forms of the histopathological slides. There are in general two types of histopathological slides. The first is the typical diagnosis slide obtained from whole tissue (e.g., tumor sample after surgery) or biopsy. Such slides are usually large given the size of the tissue. Another type of slide is tissue microarray. This is obtained by cutting paraffin embedded samples from needle biopsies. These samples are usually densely aligned in a paraffin block, and a single tissue microarray slide can contain a few hundred samples. In both cases, the samples are embedded in paraffin block and sliced into thin sections (often 5 micron thick) and then stained. The most commonly used staining technique is H&E staining with H stands for hematoxylin which is a dark blue stain marking DNA/RNA in the
nucleus and RNA in ribosomes, while E stands for eosin which is a red stain marking acidophilic substances. Another type of staining is the immunohistochemical (IHC) staining which is to detect the presence and abundance of specifically selected protein by applying specific antibody targeting the protein of interest. Sometimes IHC is combined with other stains (e.g., hematoxylin) or multiple antibodies are applied using a multiplexing technology to stain for multiple proteins. Fluorescent staining is also used especially for in situ hybridization experiments. For the stained slides, while traditionally individual images can be taken using digital cameras connected to microscopes, whole slide image is often acquired using specialized digital scanners such as the Aperio ScanScope and Virtual Slide Microscopy (http://www.olympus-lifescience.com/en/microscopes/virtual/). The resolution of the images is determined by the selected magnification of the objective lens with ×20 and ×40 being the most commonly used ones.

2.1.2 Cell based Features for Histopathological Images

After the digital tumor tissue slides are digitalized using the scanners, computational pipelines are developed to automatically extract features from these images. Feature extraction method measures the cell morphology by quantifying cellular-level and tissue-level characteristics of the cancer.

Cells are basic structural and functional units in all tissues. Cells in most epithelial cancer tissue and its microenvironment are comprised of (not limited to) cancer epithelial cells, normal epithelial cells, fibroblasts, immune cells, and endothelial cells. Each type
presents different spatial and morphological phenotypes associated with distinct biological functions. The cell-based features are extracted based on shape, color intensity and textual patterns of individual cells. Prior to cell-based feature extraction, a step of cell segmentation for recognition of cell objects is often needed. Cell segmentation algorithms have been discussed in many literatures [27]–[31]. Shape features are usually computed based on the area, asymmetry, compactness, orientation, perimeter, solidity, eccentricity, equiv diameter, majoraxis length and minoraxis length of any cell. Increased size is a biomarker for cancer cells compared to normal cells. Along with increased size, abnormal shape of epithelial cells also indicates the severity of the cancer. Pixel intensities within the cells are also informative in revealing ribosome changes and DNA contents [32], [33]. Common intensity-based cell features include intensity mean, intensity standard deviation and contrast to neighbor pixels, and etc. The first, second or even third order statistics of the pixel intensities are also valuable markers for evaluating the malignance [34]. In comparison to shape and intensity based cell features, textural features are properties of the cells that are not visually accessible to pathologists. However, they gain great interest in the machine learning and computer vision field, due to their ability of discriminating different objects in a variety of images. Textural gray-level co-occurrence matrices (GLCM), wavelets and local binary patterns are common textures being used [35], [36]. A lot of the applications of these features in histopathological images are have been reported [37]–[41].
Since distinct cells function differently in cancer development and progression, in many cases researchers study cell types separately. Consequently, a cell classification step can be added before extracting the cell-based features. This allows us to compare the distributions of features from each individual cell type, and thus discover characteristics of cell types with distinct biological function in cancer.

2.1.3 Tissue based Features for Histopathological Images

Cells do not function alone; instead they interact with each other and the extracellular environment. In tumor, this cellular environment is called tumor microenvironment, which includes surrounding blood vessels, immune cells, fibroblasts, lymphocytes, signaling molecules, and the extracellular matrix (ECM) [42]. Cancer development depends on tumor microenvironment for sustained growth, invasion and metastasis. The properties of the cells and the composition along with arrangement of the tumor microenvironment elements vary tremendously, leading to tumor heterogeneity. Many studies have shown that tumor microenvironment and its stromal cells can be effective targets for treating cancers [13], [43], [44].

Quantitative analysis on the arrangement of tissue and compartments of tumor microenvironment relies on microscopic scanning of the tissue. Spatial and textural features are most common descriptors of tissues. Examples of spatial features include distances among tumor cells, distances between tumor cells and neighboring fibroblasts, cell density, and fiber arrangement. This information is essential to model tumor genesis. For instance, fibroblasts act in wound healing, angiogenesis and tissue remodeling by
releasing growth factors and proteases such as matrix metalloproteinases [45]. Therefore, their presence to cancer tissue may facilitate cancer cell growth and invasion. Cancer cell density is another important measurement used in pathological reviews in clinical practice. Studies using tissue-based image features often adopt the spatial information to model biological relevant information [13], [46], [47].

Another important set of features is texture features from local tissue. Texture features are widely used in the histopathological images due to their ability of characterizing tissues by a spatial organization of their cellular and connective tissues components. In an uncontrolled environment regarding staining conditions, intensity-based features can be noisy, making texture features most useful due to their invariance to these conditions [48]. As the most commonly calculated texture features, Haralick features are spatial descriptors that both measure the grey scale distribution in an image and consider the spatial interactions between pixels [49], [50]. Besides their popular use in computer vision and pattern recognition field, Haralick features are intensively utilized to histological image analysis [37], [38], [51], [52]. Another commonly-used types of texture features are local binary patterns (LBP), which are used to classify tissues [53], [54] and to predict patients’ survival [41].

2.1.4 Machine Learning Issues

In the above section, we discuss the cellular and tissue features of compartments in tumor. However, tumors are not homogenous tissues; instead they are composed of different compartments with distinct morphological and phenotypic profiles, including
cellular morphology, gene expression, and proliferation. To study these different compartments separately in the tissue slide images, we have to apply tissue segmentation methods to classify the smaller and homogeneous sub-image into tissue types. These methods are usually adopted from classic pattern recognition and machine learning methods, which are used to classify fingerprint and face images. In general, there are two steps: first, a set of image descriptors are extracted from each individual tissue; second, a classifier that is constructed from training samples is utilized to determine the membership of the remaining tissue.

The major challenge for tissue classification is the biological variations that dramatically limit the robustness of the techniques. To overcome this problem, recent studies have designed image features for tissue classification. For instance, Linder el al [40] used support vector machine (SVM) with LBP textures to classify epithelium and stromal patches from tumor tissue microarrays (TMAs). Mosaliganti el al [55] proposed the N-point correlation functions (N-pcfs) to construct an appropriate feature space for achieving tissue segmentation. Kothari et al [56] extracted shape-based features that capture the distribution of stain-enhanced cellular and tissue structures for classification of histological renal tumor images. Janowczyk et al [57] presented a morphologic scale method to characterize local heterogeneity as opposed to local homogeneity for the purpose of classifying tumoral tissue from the stromal regions. Prasath et al [58] proposed a tissue segmentation algorithm using local feature-based active contours in a
globally convex formulation. With the help of these discriminant features, the classifier can better differentiate the tissues from the other types.

The effectiveness of the tissue classification system also relies on the classification methodology. In almost all studies of tissue classification, supervised classification methods are used. In general pattern recognition, classifiers such as k-Nearest Neighbors (k-NN), decision trees, naïve Bayes and SVM can be applied [59], [60]. Multiple-classifiers [61] can also be potentially applied to histopathological images to combine the power of different methods by boosting each other. However, the high-dimensionality nature of the image data poses a great challenge to the task. Another issue of tissue classification is the heterogeneity of the image samples, which is a result of the segmentation of the original tissue slides. To fulfill the assumption of homogeneity in the to-be-classified image samples, the segmentation methods will need to guarantee that the segmented sub-image contains is homogeneous in tissue type.

2.1.5 Applications in Computer Aided Diagnosis

Many successful automated systems for diagnosis based on histopathological image have been developed. For example, Doyle et al [39] designed a computer-aided diagnosis system to assist pathologists by automatically detecting prostate cancer histological specimens. Textual and structural features from histopathological images were used by Kalkan et al to detect colorectal cancers [62]. Automated renal cell carcinoma subtyping was achieved by combining image morphological analysis, wavelet analysis and texture analysis, and a robust classification system based on a simple Bayesian classifier [63].
Wang et al suggested that integration of histopathological features from tumor nests could be a useful parameter to predict the prognosis of early stage breast invasive ductal carcinoma [64]. In the study by Beck et al [13], stromal morphological features were found to be strongly associated with survival time of breast cancer patients than the morphological features obtained from epithelial compartments. Morphologic variations of glioblastoma revealed prognostically significant morphology-driven subtypes [65]. Yuan et al identified the associations between higher number of infiltrating lymphocytes and better clinical outcomes of ER-negative breast cancer [66]. Novel histology markers were also found as better predictors for survival of non-small cell lung cancer patients [41].

2.2 Integration with Other Data Modalities

Histopathological images play an important role in characterizing cancer, due to the capacity of providing spatial information and discrimination of different compositions of the tumor. To date, remarkable progress has been made to integrate histopathological images with other data modalities. For example, using histopathological images, Yuan et al [66] developed an image processing method that automatically detect lymphocytes and found that the infiltrating of lymphocytes was associated with better ER-negative patient outcomes. Further, they quantitatively measured intratumour lymphocyte and found the measurement could provide a more powerful diagnosis of triple negative breast cancer [47]. Sung and Sohn discovered that 19 genes can classify endometrial cancers into the two major histological subtypes, endometrioid and serous [67]. Wang et al identified
integrative biomarkers of image features and genes that can predict triple negative breast cancer [68].

One major application of histopathological images is to stratify patients into cancer subtypes. Patient stratification based on single type of data is well-established. Kmeans [69], hierarchical clustering [70], consensus clustering [71] methods are usually used to cluster the patient samples into groups using multi-dimensional data.

In recent years, there have been multiple efforts on essential methodologies in the analysis of multi-modality biomedical data sets. Integrative patient stratification is a class of analysis that clusters the patients into subgroups based on multiple data types, so that the characteristics and features of the subgroups can be revealed and clinical-relevant subgroups can be identified. Recently, there are many methods that focus on achieve this goal. They can be categorized into four major classes:

1. **Latent variable methods.** Latent variable models are usually used when the data is in high-dimensionality. In integrative analysis of heterogeneous data, latent variable models advanced in finding associations between different data types. For example, Lê Cao at el proposed a new R package called integrOmics [72], which uses sPLS (sparse Partial Least Square) to study causal relationship between the two datasets and CCA (Canonical Correlation Analysis) for reciprocal relationship. Based on sPLS, Li et al. introduced a sparse Multi-Block Partial Least Squares (sMBPLS) regression method to identify multi-dimensional regulatory modules [73]. This method advanced sPLS by allowing groups of latent variables to be identified. Based on similar linear functions,
iCluster [74] utilizes a Gaussian latent variable model with lasso type penalty to estimate common latent labels among all data types. An upgraded iClusterPlus method was proposed to distinguish cancer genes that are reliable of a subtype from those that are less reliable [75]. The input of iCluster is limited to continuous measurements. These methods assume that there are common latent factor in the population. Yuan et al. [76] develop a model called Patient-Specific Data Fusion (PSDF), which assumes that associations between different data types occur, in each individual sample. The latent variable methods exploit the associations of biological entities (e.g., genes) to estimate the relationship between different levels. However, increasing the dimensionality of the features will increase the difficulty of estimation.

2. Penalized likelihood methods. These regression methods aim to select multimodal features (genes, copy number variance and etc.) to better predict patient outcomes. Based on the single data type model Lasso [77], Mankoo et al. develop a multivariate Cox Lasso model and apply it to four genomic data types [78]. Yuan et al propose a Lasso-based integrative approach to identify a sparse interaction network of DNA copy-number regions with their downstream targets [79]. Penalized likelihood methods essentially achieve patient stratification through features selection. These supervise methods usually require labels or outcomes, which are not available before subtype discovery, making these methods limited to classification into predefined subtypes.

3. Network based methods. Patient samples are represented as nodes in the networks and the edges represent their affinity. A single data type model is built in [80]. Recently,
Wang et al. [81] propose a Similarity Network Fusion (SNF) that solves integrative clustering by constructing networks of samples for each available data type and then fusing these into one network. Similarly, we have developed a visualization tool that accommodates clustering based on multiple molecular measurements using graphs iGPSe [82]. Network visualization tools help clinicians to visualize the population structures and explore subtypes, but the network stratification usually requires a threshold for sparsifying the input affinity network, which is difficult to define.

4. Consensus clustering methods. Consensus clustering methods aim to aggregate clusterings of measurements from different modules into a set of “consensus” clusters based on certain optimization criteria. Strehl and Ghosh [20] seek a consensus clustering by maximizing the mutual information. Topchy et al. [83], [84] consider a representation of multiple clusterings as a set of new attributes characterizing the data distributions, and then a mixture model (MM) offers a probabilistic model of consensus using a finite mixture of multinomial distributions in the space of base clusterings. A consensus result is found as the solution to the corresponding maximum likelihood problem using expectation maximization (EM) algorithm. However, this method cannot be directly applicable for integrating the numerical molecular data such as gene expression levels. Recently, we developed an integrative clustering algorithm called Regularized Patient Stratification [85], which utilizes the molecular density of subpopulations to regularize the clusterings from clinical classifications. RPS allows integrating the numerical molecular data such as gene expression levels with categorical data such as clinical
staging information given that the latter is highly important due to their wide clinical applications.

2.3 Resources and Tools

There are many software tools for reviewing digital slide images. For example, Cancer Digital Slide Archive [86] offers online tool for browsing TCGA digital cancer slides. As an offline tool, OpenSlide (http://openslide.org/) is an opensource library that provides interface to read whole-slide images. Leica Biosystems also offers Aperio ImageScope which is commonly used in clinical practice.

For histological image segmentation, superpixel segmentation methods are provided in C/Matlab package by [87]–[89]. CellProfiler is an opensource software designed to quantitatively measure phenotypes from thousands of images automatically. (http://www.cellprofiler.org). For cell images, CellProfiler can further automatically segment cells from other tissues.

To computationally extract morphology features from tumor histopathology, several commercial software and packages are available. Matlab offers feature extraction toolbox that includes features such as histogram of oriented gradients (hog), speeded up robust features (surf), local binary patterns (lbp) and Haar wavelets. Definiens Image Miner integrates data mining with image analysis (http://www.definiens.com/). In particular, it can generate more than 6000 features from pathological images.

Incorporation of bioimage measurements with genomic, epigenomic, transcriptomic and proteomic data can achieve integrative subtypes of cancers. This urges the
development of many integrative clustering algorithms that allows incorporation of multi-modalities data. The main goal of these methods is to improve accuracy and robustness of subtype identification. This problem can be addressed as consensus clustering problem [84]. In [85], a consensus clustering based method is proposed to regularize categorical clinical attributes using numerical measurements such as genomic and morphological features. iCluster [74] is another tool that allows achieving clustering while discovering latent variables from different data types. Patient-specific data fusion (PSDF) [76] constructed a hierarchy of Dirichlet processes that assists researchers to find concordance between datatypes. The similarity network fusion (SNF) [81] proposed by Wang et al is another tool which can be utilized to cluster patient samples based on bioimaging features and other data types. All the above algorithms and methods provide public-available code and tools.

Beyond exploring the multi-modality data of cancer patient, researchers and clinicians may also desire to visualize the data. The cBioPortal (http://www.cbioportal.org/) provides visualization, analysis and download of large-scale cancer genomic data sets such as TCGA. Two integrative visualization tools StratomeX [90] and iGPSE [82] were proposed to explore subtypes of cancer via different data types.
Chapter 3 : Histopathological Image Processing and Morphological Feature Extraction

In order to investigate the relationships between rich information in cellular morphology and molecular characteristics, histopathological images need to be quantitatively measured using image processing and feature extraction methods. In this chapter, I present the computational workflow of histopathological image processing and morphological feature extraction used in this study.

Section 3.1 provides a brief overview on the problem of image processing and feature extraction of histopathological images. This is followed by the image acquisition of our collected histopathology datasets in Section 3.2. Then, Section 3.3 describes the image processing workflow adapted in this dissertation. Section 3.4 and Section 3.5 discuss the issues of cell and superpixel segmentation, respectively. They are followed by tissue classification method in Section 3.6. Finally in Section 3.7 and Section 3.8, I shall describe the details of imaging features.

3.1 Overview of Image Processing and Feature Extraction Pipeline

Histological images are the key ingredients in medical diagnosis and prognosis in today’s medical field. In most cancers, a definitive diagnosis of the tumor can be provided based on a combination of histological evaluations that are visually perceived by the pathologists from tissues with hematoxylin and eosin (H&E) staining. A needle
core biopsy is a preferred method for most of the cancer diagnosis and prognosis because of its low morbidity, low cost and specific information about the tumor. Example images are shown in Figure 3.1 to demonstrate normal breast cancer tissue compared to cancer tissue in histology.

Figure 3.1: Example images of the benign breast histology (A) and the breast cancer tumor histology (B and C). On the upper panel (A), the normal breast tissue demonstrates clear duct lobular units and glands. Layers of epithelial cells can be observed. In comparison, the cancer histology presents intra-ductal proliferation (B) on margin of resection of an invasive ductal carcinoma. A higher-grade ductal carcinoma is shown on the panel (C). Tubular structures are poorly differentiated.
Histopathology also plays a crucial role in helping to improve our understanding of diseases and their treatment in biomedical science. For biologists, advances in high-throughput microscopy techniques enable them to study different compartments of the tumors, in order to ultimately understand the heterogeneity of cancer and how cancer cells are interacting with each other as well as the tumor microenvironment. Histopathological images provide important tumor features unavailable in omics and genetics data – the spatial information of the tumor cells and the tumor microenvironment. Integrative studies that involve with genetic, genomic and phenotypical features such as morphology from histopathological images are not only providing clinical implications, but also offering scientific hypothesis on cancer development and progression.
Figure 3.2: Imaging processing workflow in this study. The presentative patches are first manually selected from whole slide images. Image preprocessing steps remove the background and artifacts. Superpixel segmentation method further cuts the image patches into superpixels where tissues are homogeneous. Then, local binary pattern features are extracted for classification of the superpixels. Finally, a variety of features are extracted, some of which are based on cell segmentation.

However, histopathological data are not direct measurements such as DNA sequences. Pixels and structures in the images have to be quantified using computational feature extraction methods. Pathologists can visually assess the histology from tissue slides, but the perception may be subjective and not suitable for large datasets, such as TCGA. Consequently, computerized techniques for performing these analyses need to be developed. While there has been a long history of automated quantitative analysis of radiological images which provide organ level information, quantitative characterization of the morphology at cellular and tissue levels still remains a challenging problem even though the microscopic images are the most commonly used data in clinical pathology.
In this thesis, I develop a histopathological image processing workflow, which includes the state-of-art segmentation, tissue classification methods. Most importantly, it collects more than 100 biologically relevant imaging features. One subset of these images features are selected from previously published literatures and proved to be important in breast cancer. Although some texture features are included in previously discussed features, we also extracted the whole category of Haralick features as our second subset of features, aiming to fully capture the rich textual information in the tumor. The third subset of the features is designed to mimic the pathologist’s perception. The overall image computational pipeline is demonstrated in Figure 3.2.

3.2 Image Acquisition

In this research, we work with multiple datasets of both breast cancer and lung cancer histopathological images. Both cancers are leading causes of death worldwide [91] and have very high mortality rates. To discover clinical and biological relevance, patient cohorts in all datasets have clinical and genomic data available, along with histopathological images. The sample statistics of image datasets in TCGA is shown in Table 3.1. Here are details of the three histopathological image datasets:

1. TCGA Breast Cancer Dataset. The breast cancer data is obtained from The Cancer Genome Atlas (TCGA)[92]. The TCGA project collects high-quality breast tumor samples and makes available the clinical information, molecular/genomic profiling data, and histopathology slide images on its data portal. From the TCGA portal, we downloaded ×40 magnification whole slide
images in the SVS file (single-file pyramidal tiled TIFF) format. Tissue slides are available as thin slice of snapfrozen optimal cutting temperature embedded block of tissue for imaging. We used tissue slide images from frozen tissue sections instead of diagnostic slides, since their adjacent tissue samples were used to provide DNA and RNA material for generating genomic data. In total, the TCGA possesses about 1000 patient samples and the sample collection is ongoing. The initial SVS file for each slide is about 1GB in data size. Typical size of these images is about 100 000×300 000 pixels.

2. TCGA Lung Adenocarcinoma Dataset. The lung cancer adenocarcinoma (LUAD) data is also obtained from TCGA. It is supervised by National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) and freely available to researchers. Before publishing the data, a pathology review is conducted to determine whether the frozen tissue block (from which the slide originates) is qualified to make DNA/RNA for TCGA. From the TCGA portal, we downloaded ×40 magnification diagnosis images in the SVS file format. The tissue for producing a diagnostic slide should be geographically adjacent to the tissue for producing the corresponding tissue slide.
<table>
<thead>
<tr>
<th>Cancer Type</th>
<th>Image Type</th>
<th># Images</th>
<th># Patient Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA</td>
<td>Diagnostic slides</td>
<td>1110</td>
<td>1036</td>
</tr>
<tr>
<td>BRCA</td>
<td>Tissue slides</td>
<td>2306</td>
<td>1101</td>
</tr>
<tr>
<td>LUAD</td>
<td>Diagnostic slides</td>
<td>427</td>
<td>380</td>
</tr>
<tr>
<td>LUAD</td>
<td>Tissue slides</td>
<td>1022</td>
<td>487</td>
</tr>
</tbody>
</table>

Table 3.1: TCGA data portal statistics. The TCGA repository contains both diagnostic slides and tissue slides from breast cancer and lung adenocarcinoma patients samples. Note: The exact numbers of samples may alter, since the repository is being periodically updated.

3. OSU Triple Negative Breast Cancer Dataset. The other dataset used in this research is a local Triple Negative Breast Cancer (TNBC) dataset. The Ohio State University (OSU) Pathology Core Facility collects breast cancer biopsy specimens which are stored in the OSU Tissue Archive Service. The necessary clinical information is available from the Information Warehouse at the OSU Medical Center (OSUMC). A total of 365 TNBC patients were identified between the years 1995 and 2005. After pathology review of tumors with sufficient sample for study, 175 paraffin blocks for TNBCs were selected to generate tissue microarrays (TMAs) used in this study. The TMAs were stained using H&E and digitized by an Aperio ScanScope under ×20 magnifications.

In all the above cohorts, images are stored in TIFF uncompressed format. In the first two TCGA datasets, samples are matched to their data from other types using TCGA
barcodes. In the OSU TNBC cohort, a pathological core map from the data generation center is used to match the samples to their clinical and molecular data.

3.3 Preprocessing

Generally, an automated cancer computational pipeline consists of four main computational steps: preprocessing, segmentation, feature extraction, and diagnosis analysis [93]. The aim of the preprocessing step is to eliminate the background noise and to improve the image quality for the purpose of offering stable performance which is insensitive to imaging conditions and variations. In this research, according to the content in tissue slide images, we apply the following preprocessing steps:

1. Quality Control. First of all, the image slides are visually assessed for staining quality and adequate tissue information. Tissue slides with major missing tissues and artifacts are excluded for quality purpose. This quality-control step is to ensure that the images are presentative of the tissue morphology.

2. ROI (Region Of Interest) Selection. After quality control, four representative image patches for each whole breast cancer slide image were curated by manually selecting heterogeneous and informative regions containing both tumor and stromal tissues. These patches are considered as the ROIs. All patches varied between 3000 and 5000 pixels in width and height. This step is necessary because it is very difficult to process the entire image due to the large size and high computing cost. In addition, whole slide images contain redundant information and encompass artifacts such as folding and missing and broken tissue, which can
be removed from the following analyses. This process eliminated artifacts and low-quality regions from consideration. Magnification for digitization was enforced to be consistent so that all images within each cohort are of the same resolution.

3. Noise Reduction. During the ROI selection process, we ensured that both tumor epithelial and stromal compartments existed on chosen slides. However, background, red blood cells and white space are not eliminated. Therefore, in the third step of preprocessing, the most trivial method is applied which is thresholding the pixels of an image. This allows background and white space correction. Color images were then filtered to remove extreme values in the RED channel, which is used to delineate blood cells and spills.

4. Color normalization. One of the major steps in preprocessing of histopathological images is color normalization. This stage reduces the differences in tissue samples due to variation in staining and scanning process [94]. A simple approach to match the histogram of the target image to the one of the reference images is applied in this research.

3.4 Cell Segmentation

Cells are basic structural and functional units in all cancer tissues. Cells in cancer tissue and microenvironment are comprised of (not limited to) cancer epithelial cells, normal epithelial cells, fibroblasts, immune cells, and endothelial cells. The first two types of cells are generally called epithelial cells. And, the latter ones are called stromal
cells, which are dominantly found in stroma [95]. In both tumor and tumor microenvironment, cells are undoubtedly the most important structures in studying cancer.

Hence, segmentation of cells (nuclei in histopathology) is an important step towards successful automatic analysis of digitized microscopy images [96]. Automated image segmentation for cell analysis is generally a difficult problem due to the large variability [97]. In the context of automated segmentation of H&E images, there are five distinct categories of methods: histogram-based techniques (intensity thresholding), edge-detection methods, region-based algorithms, pattern recognition approaches, and active contour models [31]. Each category of methods has its advantages and drawbacks [31]. Intensity thresholding method is still one of the most predominant approaches to cell segmentation, because of its simplicity. However, the clumping of cells has to be solved after the initial thresholding. Active contour models have the advantage of resolve overlapping objects into independent shapes. For example, Hafiane et al. [98] present contour models based method using fuzzy c-means with spatial constraint algorithm to detect the potential regions of interest and multiphase vector-based level to refine the segmentation. Ali and Madabhushi [99] proposed a synergistic segmentation scheme, which combines shape priors with boundary and region-based active contours in a level set formulation with a watershed scheme for model initialization for identifying and resolving multiple object overlaps in an image scene. Kong et al. [100] provided another segmentation workflow, in which clumped object splitting using a watershed method. For
simplicity and low computational cost, we select otsu thresholding segmentation followed by morphological operations such as opening and closing. Otsu segmentation finds an optimal threshold based on distribution of gray values of pixels. Next we remove from the images all connected components that have fewer than 20 pixels. Morphological opening, closing and watershed segmentation operations are applied to separate the clumping cells. Figure 3.3 demonstrates the cell nuclei segmentation result.

![Figure 3.3: Nuclei segmentation within each region of interest. (A) An example of the original superpixel. (B) The result after Otsu cellular segmentation. Some of the nuclei overlap and form large clumps. (C) Boundary of the cell nuclei. (D) Final segmentation of the cell nuclei after using edge cut (example indicated by white arrow).](image)

### 3.5 Superpixel Segmentation

Tissue samples in histopathological images are heterogeneous with existence of multiple tissues (eg, tumor and stroma) and cells (eg, tumor epithelial cells, fibroblasts, endothelial cells, macrophages). In order to quantitatively measure different compartments of the tumor, it is also important to partition the whole slide images and
the tumor microarrays into relatively homogeneous non-overlapping regions, representing the corresponding compartments. The superpixel partitioning provides a decrease in data size of the whole image, while encoding potentially useful morphometric, geometric, and appearance information within the superpixels themselves. Generally, in histological image analysis, sub-images are often curated in two ways:

1. Gridding. The whole tissue images are cut into square grids of the certain size \([101]-[103]\). These operations aim to fill in the potential gaps between tissue fragments and to remove small artifacts from the image. However, in gridding techniques, the choice of square sizes is essential to the analysis. Larger size of the grid will result in more heterogeneity of the image tiles, and thus leading to miss-labeling and error. Smaller size of grids will increase and numbers of tiles and thus increase computational cost in the following steps, such as the tile classification stage. In addition, smaller size of the grids will also limit the number of cells permitted in a single grid. Due to great scale variation of compartments in cancer histology, the determination of the size of the grids is always a tough problem.

2. Superpixels [104]. By definition, superpixels are sub-images which are local, coherent, and preserve most of the structure necessary for segmentation at the scale of interest [105]. Superpixel algorithms group pixels into perceptually meaningful regions, which present more structural information than the rigid structure of the pixel grid. Superpixels have the advantages of capture image
redundancy, hence providing a convenient primitive from which to compute image features, and greatly reducing the complexity of subsequent image processing tasks [106]. In our study, superpixel technique is superior to gridding method in the way that the boundary of superpixel zigzags through cells and extracellular matrices. It also results in flexible sizes of the sub-images, generating more homogeneous regions.

Our choice of tissue segmentation was inspired by the study of Beck et al [13]. We adopt an entropy-based image segmentation algorithm [89] similar to that in Beck et al [13] to divide the images into small regions with relatively homogeneous cellular components and morphology - ‘histological superpixels’. Figure 3.4 demonstrates the superpixel segmentation results of breast cancer images of TCGA and OSU TNBC tissue microarrays. It can be observed that the cancer cells are clearly separated from the stromal tissues. The stromal superpixels are mostly pure fiber tissues with stromal cells, which are excluded from the cancer epithelial superpixels. It demonstrates the effectiveness and the robustness of the superpixel segmentation method.
3.6 Tissue Classification

For an automated image analysis pipeline, where tissue heterogeneity should be considered, differentiation of the tissues follows the tissue segmentation step, such as superpixel segmentation described in the previous section. In traditional clinical practice, a skilled pathologist easily segments the tissue into compartments and can report immunohistochemical (IHC) staining in tumor cells and stroma separately. However, in automated pipeline, the tissue classification requires extensive machine learning study.

It is noticeable that in some studies IHC staining is available for each H&E images, so labels of cancer compartments can be automated obtained from the IHC staining.
However, in most cases where IHC staining is not available, pattern classification methods are needed to classify the tissue slides into different compartments.

In our studies, cancer tissues possess epithelial compartment and stromal compartment. In the context of segmenting histological images into epithelial and stromal, there has been a substantial research focus on the problem \[53\], \[107\]. The common approach is to compute feature information from the images, and then train a classifier to distinguish cancer from non-cancer. A variety of features have been used to distinguish the tissues, from low-level feature and high-level features. Low-level features contain information about local distributions of pixel intensities, color and texture, intensity edges and their local orientations \[107\]. High-level features contain structural information about the content present in the image, such as tissue components (e.g., glands, lumina, stroma, nuclei), their shape, color, size, and geometric arrangement \[107\]. High-level features \[13\][108] correspond to information assessed by pathologists; however, they are sensitive to different cancer types and scales of images. In contrast, low-level features \[53\], \[109\], \[110\] such as Local Binary Patterns (LBP) are invariant to scales and types. For the task of classifying epithelial and stromal tissues, these texture features work particularly well due to the textural difference of the two classes. In this study, we adopt the LBP feature and supporting vector machine (SVM) pipeline, for its simplicity and reported high performance on classification rate \[53\].
This supervised tissue classification process requires training data. First, we randomly select certain numbers of test images. Then, for different datasets of histological images, a well-trained pathologist labels the superpixels as epithelial or stromal. The statistics of the datasets are shown in Table 3.2. The rest of the superpixels are considered as testing samples. The original color images are transformed to gray scale. LBP features are extracted from both of training and testing gray scale images. Figure 3.4 demonstrates the definition of LBP feature used in [53].

<table>
<thead>
<tr>
<th>Cohorts</th>
<th>TCGA/CPTAC</th>
<th>TCGA/TNBC</th>
</tr>
</thead>
<tbody>
<tr>
<td># samples</td>
<td>90</td>
<td>44</td>
</tr>
<tr>
<td># images per sample</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td># image total</td>
<td>360</td>
<td>176</td>
</tr>
<tr>
<td>data size</td>
<td>15.8GB</td>
<td>12.1GB</td>
</tr>
<tr>
<td># patches per image</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td># images training</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td># images testing</td>
<td>348</td>
<td>164</td>
</tr>
</tbody>
</table>

Table 3.2: Statistics of the image dataset and detailed training and testing cases.
Figure 3.5: Local binary patterns (LBP) are defined in [20]. (A) Gray-scale original image. For every $3 \times 3$ pixel neighborhood (B) within the image, an LBP code is generated by thresholding the surrounding pixels using the value of the central pixel (D). A histogram (E) of the all LBP codes within the analyzed image is formed to represent texture properties of the image.

Our implementation is slightly different, since our original images are superpixels which have irregular boundaries, unlike squared images shown in the figure. Hence, we abandon the pixels on the edge of each superpixel in calculating LBP features. The histogram of the LBP features from the training samples are then treated as discriminant descriptors and trained using quadratic SVM classifier. Finally, the testing images are classified into two classes (epithelial and stromal) using the trained classifier. Figure 3.6 shows examples of the classified tissue superpixels. We can observe that the right two superpixels are enriched with cancer cells, while the left two are enriched with fibers and elongated fibroblasts, other cells.
In order to test the performance of the tissue classification method, we first input the training samples to the classifier. The misclassification rate is $3/1200 = .25\%$. Further, we selected another 12 images from the TNBC testing group and label each superpixel as either epithelial or stromal. The misclassification rate is 3.08\%. The performance is similar to the agreement between the classifier and the human observer reported in [53].

3.7 Feature Extraction

In this section, we demonstrate the definition and implementation of histopathological image features extracted from the collected cancer image datasets. Inspired by Beck et al.’s study of breast cancer, where more than 6000 image features were extracted from the breast cancer tissue using commercial software – Definiens Developer XD, we re-implemented salient features discovered to be associated with breast cancer malignancy. This subset of salient features includes both stromal and epithelial features that were
found to be associated with samples from patients who died sooner. Hence, these features form our first category of our whole image feature set.

Another important set of features is texture features from local tissue. Texture is the most used feature in histopathology image analysis [111]. Among all the texture features, Haralick features are most common features of textures. Haralick features are statistical descriptors that both measure the grey scale distribution in an image and consider the spatial interactions between pixels [50]. In addition to popular use in computer vision and pattern recognition field, Haralick features are intensively applied to histopathological images [37], [38], [51], [52].

In many systems biology studies, the computational models are generally built to simulate the complex biological systems, including histopathological feature design. In our study, we desire to design features that can model crucial clinical and pathological information. Thus, we work with pathologists to understand the visual perceptive of their observation. Based on their guide, we design and implement the third category of image features, which we call structural features. These high-level features contain semantic information about the cellular structures in the images, such as nuclei and stroma, their shape, density, size and geometric arrangement.
Above all, Figure 3.7 shows the hierarchy of image features used in this study. From the epithelial and stromal tissue we classified, our feature extraction step provides measurements of cell nuclei shape, nuclei density, stromal textures and intra-tissue relationships. They instruct to find quantitative aspects of breast cancer tissue that can predict prognosis. These features are also breast cancer histology specific, and are examined by trained pathologists. In the follow descriptions, we demonstrate our implementation of the three classes of image features we extract from the breast cancer images.

### 3.7.1 Previously Discovered Salient Features

Definiens Developer XD is a comprehensive toolbox for the complete development process from rapid prototyping to deployment of automated image analysis [112]. It uses
image features as source of information, to measure the characteristics of image objects. Image objects have spectral, shape, and hierarchical characteristics. These characteristic attributes are called features. Features are used as source of information to define the inclusion-or-exclusion parameters used to classify image objects and quantitatively measure image objects.

According to Definiens’ definition (Defineins Software Manual obtained from personal communication to Andrew Beck), an image object is a set of pixels. The set of pixels of an image object \( v \) is denoted as \( P_v \). In the context of processing cancer histological image, superpixels are referred as image objects. An image layer \( k \) is referred to the channel of an image. For example, in our study, image layers are defined as the red, green and blue channels, since the histological images are all in RGB format. Sub-objects are sub-structures in the image objects, which in our definition are cell nuclei. In epithelial superpixels, sub-objects are epithelial cells, while in stromal superpixels cell nuclei are filtered based on eccentricity and those with high eccentricity are recognized as fibroblasts.

In Definiens, object features are grouped in the following categories:

**A. Layer Values**

Layer value features evaluate the first, second, and third statistical moment (mean, standard deviation, and skewness) of an image object’s pixel value and the object’s relations to other image object’s pixel values. The use of these is to describe image
objects with information derived from their spectral properties. The layer values features in our feature system consists of the following sub-categories:

**A.1. Mean Features**

Mean: Features in this group refer to the mean layer intensity value of an image object.

Brightness: 
\[
\bar{c}(v) = \frac{1}{w_B} \sum_{k=1}^{K_B} w_k^B \bar{c}_k(v),
\]
where \( w_k^B \) is the brightness weight of image layer \( k \) with \( w_k^B = \{0, 1\}; K \) is the number of image layer \( k \) used for calculation; \( w_B \) is the sum of brightness weights of all image layers \( k \) used for calculation with \( w_B = \sum_{k=1}^{K} w_k^B; \bar{c}_k(v) \) is mean intensity of image layer \( k \) of image object \( v \).

Max. Diff.: 
\[
\max_{i \in K_B} \left| \frac{c_i(v) - \bar{c}_i(v)}{\bar{c}(v)} \right|,
\]
where \( i, j \) are image layers, \( \bar{c}_i(v) \) is the mean intensity of image layer \( i \) of image object; \( K_B \) are image layers of positive brightness weight with \( K_B = \{k \in K : w_k = 1\} \), where \( w_k \) is the image layer weight.

**A.2. Standard Deviation Feature:**

Standard Deviation: The standard deviation is calculated from the image layer intensity values of all pixels forming an image object.

**A.3. Skewness Features:**

Skewness: The Skewness feature describes the distribution of all the image layer intensity values of all pixel/voxels that form an image object; this distribution is typically Gaussian. The value is calculated by the asymmetry of the distribution of image layer intensity values in an image object. A normal distribution has a skewness of zero. A
negative skewness value indicates that an image object has more pixel/voxels with an image layer intensity value smaller than the mean; a positive value indicates a value larger than the mean. Expression of the skewness feature is denoted by: $\gamma_k(v) = \sqrt{\frac{\sum_{(x,y) \in P_v}(c_k(x,y) - \bar{c}_k(v))^3}{\left(\sum_{(x,y) \in P_v}(c_k(x,y) - \bar{c}_k(v))^2\right)^{\frac{3}{2}}}}$, where $(x, y)$ are pixel co-ordinates, $\#P_v$ is the total number of pixels contained in $P_v$, $\gamma_k(v)$ is the skewness of intensity values of image layer $k$ of image object $v$.

A.4. Pixel Based Features

**Ratio:** The amount that a given image layer contributes to the total brightness. Expression of ratio is denoted by: if $w^B = 1$ and $\bar{c}(v) \neq 0$ then $\frac{\bar{c}_k(v)}{\sum_{k=1}^{m} \bar{c}_k(v)}$; if $w^B = 0$ or $\bar{c}(v) = 0$, then the ratio is defined as 0.

**Max. Pixel Value:** The value of the pixel with the maximum layer intensity value of the image object.

**Min. Pixel Value:** The value of the pixel with the minimum layer intensity value of the image object.

**Mean of Inner Border:** The mean layer intensity value of the pixels belonging to an image object, which shares its border with other image objects, thereby forming an inner border.

**Mean of Outer Border:** The mean layer intensity value of pixels not belonging to an image object of interest, which shares its border, thereby forming the outer border of the
image object. The outer border and the inner border in Difiniens is illustrated in Figure 3.8.

![Diagram of inner and outer borders](image)

Figure 3.8: Illustration of the definitions of inner border and outer border of an image object (Defined in Definiens Developer XD Manual obtained from personal communication to Andrew Beck).

**Contrast to Neighbor Pixels**: The mean difference in contrast compared to a surrounding volume of a given size. The feature is denoted as: \(1 - \frac{c_k(B_v(d) - P_v)}{1 + c_k(P_v)}\), where \(B_v(d)\) is extended bounding box of an image object \(v\) with distance \(d\) with \(B_v(d)\) equal to \(\{(x, y): x_{\text{min}}(v) - d \leq x \leq x_{\text{max}} + d, y_{\text{min}}(v) - d \leq y \leq y_{\text{max}} + d\}\). The definition of the bounding box is illustrated in Figure 3.9.
Figure 3.9: The surrounding area of an image object v defined by its bounding box with a distance of \( d = 1 \). (Defined in Definiens Developer XD Manual obtained from personal communication to Andrew Beck)

*Std Dev. to Neighbor Pixels:* The standard deviation of layer intensity values of pixels within the surrounding volume of a given size. The surrounding volume consists of the pixels located within the extended bounding box of an image object, but do not belong to this image object. It is denoted by \( \sigma_k(B_v(d) - P_v) \).

*Circular Mean:* the mean for all pixels within a ring around the center of an image object. The thickness of the ring is defined by two radius values \( R_1 \) and \( R_2 \), where \( R_1 \) is the inner radius of the ring; \( R_2 \) is the outer radius of the ring, \( c \) is the center of the image object.

*Quantile:* The pixel intensity value of the rank of \( r = \frac{p}{100} \times N \), where \( p \) is a number between 0 and 100; \( N \) is the number of pixels.

A.5. To Neighbors:
The Mean Diff. to Neighbors feature describes the difference between an image object and its neighbor image objects, in terms of mean layer intensity values. Expression: $\bar{\Delta}_k(v) = \sum_{u \in N_v(d)}(\bar{c}_k(v) - \bar{c}_k(u))$, where $u$ and $v$ are image objects; $\bar{c}_k$ is the mean intensity of image layer $k$; $N_v(d)$ is a neighbor to $v$ at a distance $d$ with $N_v(d) = \{u \in V_i : d(v, u) \leq d\}$. Figure 3.10 illustrates the Definiens’s definition of image neighbors.

Figure 3.10: Direct neighbors $N_v(d)$ and the neighborhood within perimeter $d$ of image object $v$. (Defined in Definiens Developer XD Manual obtained from personal communication to Andrew Beck)

The Mean Diff. to Neighbors (Abs): Expression: $\bar{\Delta}_k(v) = \sum_{u \in N_v(d)}|\bar{c}_k(v) - \bar{c}_k(u)|$.

Rel. Border to Brighter Neighbors: The Rel. Border to Brighter Neighbors feature describes the extent to which an image object is surrounded by brighter or darker direct neighbor image objects. In a given image layer, it is the ratio of the shared image border of an image object to the total border. A value of 1 that the image object is surrounded
completely by brighter neighbors; a value of 0 means it has no brighter direct neighbors.

The expression of the image feature is denoted by: \( \sum_{u \in N_v^B} \frac{b(v,u)}{b_v} \), where \( N_v^B \) is the darker direct neighbor to \( v \), with \( N_v^B = \{ u \in N_v : \bar{c}_k(u) < \bar{c}_k(v) \} \), \( b_v \) is the image border length; \( b(v,u) \) is the length of common border between \( v \) and \( u \).

A.6. To Scene:

**Ratio To Scene:** The ratio of the mean intensity of image object to the mean intensity if the whole image layer. Expression: \( \bar{c}_k(v) / \bar{c}_k \).

A.7. Hue, Saturation, Intensity:

**HSI transformation:**

\[
H = \begin{cases} 
60^\circ \times \frac{G-B}{\text{max}-\text{min}} & \text{if max} = R \\
60^\circ \times \frac{B-R}{\text{max}-\text{min}} + 120^\circ & \text{if max} = G \\
60^\circ \times \frac{R-G}{\text{max}-\text{min}} + 240^\circ & \text{if max} = B 
\end{cases}
\]

channel, blue channel and green channel respectively; \( \text{min} \) is the smallest of the RGB values; max is the greatest of the RGB values.

B. Geometry

Geometry features evaluate the image object’s shape. The basic geometry features are calculated based on the pixels forming an image object. If image objects of a certain class are represented by these according to their shape. The salient geometry features are classified to the following classes:

B.1. Extent

**Area:** The number of pixels forming an image object.
Border Length: The border length of an image object is defined as the sum of edges of the image object shared with other image objects.

Length: The length of a 2D image object is calculated using the length-to-width ratio.
Expression: $\sqrt{\#P_v \gamma_v}$, where $\#P_v$ is the total number of pixels contained in $P_v$; $\gamma_v$ is the length-width ratio of an image object $v$.

Width: The width of an image object is calculated using the length-to-width ratio. Expression: $\#P_v / \gamma_v$.

B.2. Shape

Asymmetry: The Asymmetry feature describes the relative length of an image object, compared to a regular polygon. An ellipse is approximated around a given image object, which can be expressed by the ratio of the lengths of its minor and the major axes. The feature value increases with this asymmetry. It is similar to the length to width ratio. Illustration of the feature is shown in Figure 3.11.

Figure 3.11: Asymmetry for a 2D image object. (Defined in Definiens Developer XD Manual obtained from personal communication to Andrew Beck)
**Compactness:** The Compactness feature describes how compact an image object is. The compactness of an image object is the product of the length and the width, divided by the number of pixels. Expression: \( \lambda_1 \times \lambda_2 / \#P_v \), where \( \lambda_1 \) and \( \lambda_2 \) are the length and width of an image object respectively.

**B.3. Polygons-based Features**

**Area (Including Inner Polygons):** The Area (Excluding Inner Polygons) feature calculates the area of a polygon based on Green’s Theorem in a plane. Different to the Area (Excluding Inner Polygons) feature, the feature value includes the areas of any existing inner polygons (for instance the single polygon formed in the center of a donut-shaped object).

**C. Texture features**

Texture features are used to evaluate the texture of image objects. They include texture features based on an analysis of sub-objects helpful for evaluating highly textured data. In addition, features based upon the gray level co-occurrence matrix after Haralick are available. The texture features in previous literature are Haralick features, which we will focus on in the next sub-section.

**3.7.2 Texture Features**

Texture features are relatively hidden from the observation of pathologists, despite of its importance in differentiating tissues and capacity of diagnosis. A total of 20 Haralick features corresponding to uniformity, entropy, dissimilarity, correlation and etc. are
extracted. Detailed implementation of the Haralick features can be found in [49], [50], [113].

3.7.3 Pathological Features

In fact, much of the pathologists’ characterization of the histopathological images can be considered a manual or semi-manual based quantitative phenotyping even though the large image data often prevent the pathologists from characterizing the entire images. We work closely with a pathologist, and encode the pathologist’s observation in the re-designed pathological features. Actually, our pathological features are designed to mimic the Nottingham grading system for breast cancer. For example, the uniformity of the cells, nuclear size and variation are encoded in the system.

A. Density

*Epithelial Cell Density*: The ratio of the number of epithelial cells to size of the image object (superpixels). Expression: $N_v/\#P_v$, where $N_v$ is the set of epithelial cell nuclei in the image object $v$; $\#P_v$, is area of the image object $v$.

*Fibroblast Density*: The ratio of the number of fibroblasts to size of the image object (superpixels). Expression: $F_v/\#P_v$, where $F_v$ is the set of fibroblast nuclei in the image object $v$.

B. Fraction

*Epithelial Cell Fraction*: The ratio of the area of epithelial cells to size of the image object (superpixels). Expression: $\sum_{i \in N} A_i/\#P_v$, where $N$ is the set of epithelial cell nuclei in the image object $v$; $A_i$ is area of the $i$-th cell nuclei.
Fibroblast Fraction: The ratio of the area of fibroblasts to size of the image object (superpixels). Expression: \( \sum_{i \in F} A_i / \#P_v \), where \( F \) is the set of fibroblast nuclei in the image object \( v \); \( A_i \) is area of the \( i \)-th cell nuclei.

C. Shape

Shape features are added to the previously discovered feature list because of their importance in pathological review. These extra features include nuclei orientation, nuclei perimeter, nuclei solidity, nuclei eccentricity, nuclei equiv diameter, nuclei majoraxis length and nuclei minoraxis length. The details of the implementation and expression of them are provided in Table A.1 in Appendix A. Although some of these shape features are included in previous salient features, we desire to measure a complete dimension.

The complete epithelial morphological feature list is given in Table A.1 Appendix A. The implementation of stromal features is also based on the above definitions. Fibroblasts in stromal superpixels are identified by cell/nuclei segmentation on the stromal tissue and then recognized by elongation. The complete stromal morphological feature list is provided in Table A.2 Appendix A.

3.8 Correlations between Morphological Features

Within the more than 100 morphological features the workflow extracted, they are not independent with each other. In fact, one may notice that the previous discovered features also include Haralick features, for example GLCM.Aug.2\textsuperscript{nd}.moment.layer.1 is the Haralick autocorrelation feature for red channel, while in our Haralick features collection, the Haralick autocorrelation of all channels is also included. The correlation between
these two features is shown in the scatter plot of Figure 3.12. However, there are features which have little dependency, particular between epithelial and stromal regions. For example, the GLCM autocorrelation features of epithelial superpixels are plotted against those of stromal superpixels in Figure 3.12. This indicates that the texture features of epithelial and stromal superpixels have significant difference.

Figure 3.12: Scatterplots of features demonstrating their correlations. On the left panel, Haralick autocorrelation features calculated from red layer and all layers have very high correlation (PCC = 0.98). On the right panel, Haralick autocorrelation features calculated from epithelial and stromal superpixels have very low correlation (PCC = 0.10).

Pathological features, such as density of cancer cells and area of cancer cells are proved to have better prognosis power in the following studies. In fact in Nottingham grading system of breast cancer, the density is also an important measurement by cognitive vision of pathologies. The benefit of automated quantitation of histopathological image presents when other cells such as fibroblasts are also measured.
In fact, we observe that the number of fibroblasts has moderate negative correlation with the number of cancer cell in the tested samples. The scatterplot is shown in Figure 3.13.

Figure 3.13: Scatterplot of the numbers of fibroblast in stromal tissue (y-axis) and the numbers of cancer cells in epithelial tissue (x-axis). Negative correlation is observed (PCC=−0.32).
Chapter 4: Identifying Salient Morphological Features of Triple Negative Breast Cancer using Integrative Methods

As imaging becomes an efficient tool in diagnosis and prognosis of cancer, tools and computational workflows that can integrate molecular profiles and imaging features to better predict patients’ outcomes, become more important. Such imaging-to-gene correlation studies enable to identify correlations between the image characteristics and the patients molecular traits [114], [115]. With the help of high-resolution imaging features, the prognostic and diagnostic power of these models can be significantly enhanced.

In addition to this, some cancer subtypes that have worse prognosis and resistance to commonly applied treatments are prone to be further divided into subgroups with have certain molecular or cellular characteristics using these multi-modality data. Upon these identified subgroups, novel comprehensive biomarker can be identified. Triple Negative Breast Cancer is one of these types of cancer.

To address the problem of subtyping triple negative breast cancer patients into more subtle differentiations, this chapter presents a workflow to identify correlation between imaging features, allowing the public datasets to validate the importance of the correlations. This method generates integrative biomarkers for TNBCs.
The organization of this chapter is as follows: motivation and introduction is provided in Section 4.1, while Section 4.2 covers the workflow details and datasets used in this study. In Section 4.3, I present results obtained by the workflow to explore subtypes of tnbc. Finally in Section 4.4, I conclude with remarks of the.

4.1 Introduction

Breast cancer is a highly heterogeneous disease [4], [116], [117]. During the past half century, several different subtypes of breast cancers have been discovered based on histological features, specific protein markers, and gene signatures obtained from high throughput and high-content experiments. These subtypes present diverse clinical outcomes including varying prognosis and response to treatments.

Histopathological images of tumor tissues play important roles in breast cancer diagnosis, staging and prognosis [118][119]. Typically, pathologists visually review stained slides of breast cancer biopsy samples and assign scores to the detected and prevailing tumors. During this inspection, cellular composition is often assessed semi-quantitatively. This process is costly in both time and labor and the results may differ across pathologists. Recently, computer-assisted quantitative analysis of stained histology images has received wide attention in the biomedical and bioimage informatics fields [13], [66], [120]–[122]. For instance, automated quantification of the levels of salient proteins has led to the discovery of new markers of malignant cells in cancers [123]. In the study by Beck and colleagues [13], stromal morphological features were found to be
strongly associated with survival time of breast cancer patients than the morphological features obtained from epithelial compartments.

Besides the clinical use of histopathological images, subtyping and stratification of breast cancer patients has been widely studied using high throughput gene expression data [124]. For instance, van Veer et al. identified a 70-gene signature for predicting prognosis of breast cancer patients [125], while Perou et al. [1] identified basal-like and non-basal like subtypes of breast cancers. In addition, in a recent study using The Cancer Genome Atlas (TCGA) breast cancer data, investigators led by Perou identified four major subtypes of breast cancers by combining five different genomic data including gene expression, exome-sequencing, copy number variance, DNA methylation and microRNA expression [92].

While both morphological features and genomic data are widely used for breast cancer subtyping and staging, the causal and inferential relationship between genomic data such as gene expression profiles and morphology in histopathological images from breast cancer patients is still not clear. In a recent study [120], the heterogeneity of the ER-negative breast cancer cells was explained using complementary DNA copy number variance information and consequently an improved prognostic biomarker was suggested. This study suggests that the prognostic power of these two types of features (histopathological and genomic) should be combined and is likely to lead to better biomarkers for classification of the breast cancer and other types of cancer as well [33], [126]. However, there is neither a study which associates salient gene expression
biomarkers with pertinent morphology of the tissue, nor a morphology-driven result whose scope is also extended to include datasets where gene expression profiles are solely available. In essence, there is a paucity of work that incorporates multiple measurements of disease from disparate sources to create biomarkers.

In this Section, we present a novel workflow for integrating histopathological images analysis with gene expression analysis for Triple Negative Breast Cancers (TNBC). Specifically, we develop a workflow for identifying the morphological features that correlate well with survival outcome of patients. Signature morphological features with strong associations to survival were then analyzed by correlating with specific gene expression profiles using another large publicly available dataset and subsequently cross-validated with multiple datasets. Our pipeline provides a novel platform of translating the morphological features to any gene expression data without histopathological images. Our results clearly demonstrate that specific morphological features are correlated with specific genetic features (post-transcription) with enriched biological functions pertaining to cancer development and/or tumor microenvironment structure. These results suggest that new integrative biomarkers can be developed via such integrative approach. An additional aspect of our work is that we are able to work across datasets that were either collected in the public realm or acquired in specific laboratories.

4.2 Materials and Methods

4.2.1 Datasets
The Cancer Genome Atlas (TCGA) project [92] collects high-quality breast tumor samples and makes available the clinical information, molecular/genomic profiling data and histopathology slide images on its Data Portal. Fifty one (51) triple negative breast tumor samples were available by selecting from subjects with reported ER (negative), PR (negative) and Her2/neu (0 and 1+) status (using data available till Aug.25, 2012). Forty four (44) out of these 51 TNBC subjects have accompanying histopathology images of adequate quality. No other stratification of these TNBC cases were available. Owing to the short median overall follow up (< 2 years) and scarcity of survival events (6 out of 51 currently marked as deceased), survival analyses for TCGA breast cancers will not be effective in the next few years.

From the TCGA portal, we downloaded ×40 magnification whole slide images in the SVS file (single-file pyramidal tiled TIFF) format. Tissue slides are available as thin slice of snap-frozen OCT (optimal cutting temperature) embedded block of tissue for imaging. We used tissue slide images from frozen tissue sections instead of diagnostic slides, since their adjacent tissue samples were used to provide DNA and RNA material for generating genomic data. Typical size of these images is about 100,000-by-300,000 pixels. It is thus very difficult to process the entire image due to the large size and high computing cost. In addition, whole slide images contain redundant information and encompass artifacts such as folding and missing and broken tissue. Thus, four representative image patches for each whole TNBC slide image were curated by manually selecting heterogeneous and informative regions containing both tumor and stroma tissues. All patches varied between
Demographic Characteristic | Complete Set 365 | Pruned Set 175
---|---|---
Median age (range) | 51 (20-84) | 51 (20-84)
Race(%;White:African American) | 91:8 | 81:9
Stage (%; I:II) | 35:54 | 31:54
Grade 3 (%) | 84 | 89
Basal Cancers (%) | 47 | 45
Adjuvant Chemo Therapy (%) | 73 | 84
Median follow-up (mo) | 74 (4-272) | 75 (4-272)

Table 4.1: Demographics summary of the OSU TNBC cohort.

The OSU Pathology Core Facility collects breast cancer biopsy specimens which are stored in the OSU Tissue Archive Service. The necessary clinical information is available from the Information Warehouse at The Ohio State University Medical Center (OSUMC). 365 TNBC patients were identified between the years 1995-2005. After pathology review of tumors with sufficient sample for study, 175 paraffin blocks for TNBCs were selected to generate tissue microarray (TMA) used in this study. The TMAs were stained using H&E and digitized by an Aperio ScanScope under \( \times 20 \) magnification. After filtering using measures of satisfactory image quality, TMA images for 143 patients

67
were finally selected. The overall demographic profiles of the cohort were not altered significantly after filtering (shown in Table 4.1).

There are several large public breast cancer gene expression datasets with adequate information on survival outcomes and subtyping. Identified biomarkers can be cross-validated using these datasets, of which, Perou [127] (NCBI GSE2741) and NKI [128] collections are among the most frequently used. To evaluate our discovered metagenes on the ER-negative samples, we tested them on these two breast cancer datasets to see if any of the gene signatures can predict ER-negative patient survival.

4.2.2 Integrated Biomarker Identification Workflow

In this study, we focus on the discovery of TNBC biomarkers by translating gene-morphology relationships across multiple datasets. Figure 4.1 shows our overall workflow. Algorithmic details of each are described in subsequent sections.

Using the TCGA dataset, in which both histopathological images and gene expression profiles are available, we computed correlations between morphological features and expression profiles of gene clusters (Figure 4.1(A)). Then, we selected the morphological features which can classify patients into higher and lower risk groups using the OSU TNBC cohort (Figure 4.1 (B)). Assuming that the selected morphological features glean similar relationships of survival in other datasets, gene clusters with strong correlations to these morphological features can potentially serve as biomarkers for survival. We test them using public breast cancer gene expression datasets without available histology images (Figure 4.1(C)).
4.2.3 TCGA and OSU Image Preprocessing and Segmentation

During preprocessing we ensured that both tumor epithelial and stromal compartments existed on chosen slides. Magnification for digitization was enforced to be consistent so that all images within each cohort are of the same resolution. Color images were filtered
to remove extreme values in the RED channel, which was used to delineate blood cells and spills. A mask was generated to separate the superpixels in each slide.

### 4.2.4 Characterizing Cellular Morphological Features of TNBC Samples

Each tissue sample is heterogeneous with existence of multiple tissues (e.g., tumor and stroma) and cells (e.g., tumor epithelial cells, fibroblasts, endothelial cells, macrophages). We first adopt an entropy-based image segmentation algorithm similar to that in Beck et al. [13] to divide the images into small regions with relatively homogeneous cellular components and morphology called “superpixels”.

This procedure removes artifacts in slide preparation [89]. It not only retains the homogeneity of each Region of Interest (ROI), but also better captures the local morphological structure of the tissue and the relationship between neighboring tissues. The workflow is shown in Figure 4.2. Usually in TMA image analysis, a preprocessing step of color normalization is often employed to mitigate bias; however, this step will also introduce artifacts in the final assessment of the morphological features. Additionally, the subsequent pixel segmentation step relies on the true texture of the images. Finally, since the selected OSU TMA images do not contain large color intensity variations, we did not include color normalization step in our current pipeline.
Figure 4.2: The workflow of the histopathological image analysis. First, after the removal of background and noise, each tissue slide (or TMA) image was segmented into “superpixels” delineating the tumor and the stromal compartments of the tissue (Green lines mark the boundary of the superpixels in slide). Then, each superpixel is represented by a series of quantitative morphological features.

4.2.5 Morphometric Analysis for Cell Nuclei

For breast cancer tissue, the abundance of tumor cells is crucial to the diagnosis and prognosis of the patients. Typically, the size of the tumor is a key factor to consider, when the tumor grade is given. Additionally, it is well known that tumor with higher grades often leads to shorter survival, early recurrence and metastasis [129], [130]. The density of tumor cells along with other types of cells, for example, lymphocytes and stromal cells are essential to the quantitative analysis of the histopathological features of
the cancer [131]. To evaluate these cellular characteristics, we first identify cell nuclei within each selected patch of the tissue slide images. The pipeline of cell nuclei quantification is illustrated [132] in Figure 3.3. After removing the background, artifacts and white space regions, a threshold-based segmentation step using the Otsu’s algorithm [133] is applied to a superpixel to obtain a coarse segmentation of the cell nuclei (Figure 3.3A-C). In histopathological images, nuclei often overlap with each other and form the clumps of cells during fixation and staining. The clumps are separated with an edge-cut set selection algorithm (Figure 3.3D) [133]. Objects without enough intrinsic nucleus area are considered artifacts and not counted for the analysis.

4.2.6 Morphological Feature Extraction of Tumor and its Microenvironment

Besides tumor density being an important characteristic of cancer tissue, morphology of tumor tissue is heavily influenced by stromal and immune cells as well as extracellular matrix (ECM). Here, we measured three classes of morphological features describing the distribution and spatial information of the tumor and its microenvironment. Figure 3.3 illustrates the extraction of nucleus from other molecular compartments. After we obtained the segmentation of the nuclei, three categories of morphological features are measured on the nuclei: signal intensity, texture and shape. Shape features include the area of the cell nuclei and eccentricity. These features are tested to be discriminatory and representative for survival of generic breast cancer population [13]. Examples of morphological features are described in Table 4.2. A full list and summarized descriptions of image features are shown in Appendix A. After features for each
individual superpixel in a slide are extracted, the mean estimate of these features is obtained to represent the feature vector for this whole slide.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLCM_Ang_2nd_Moment</td>
<td>Haralick Texture. grey-level co-occurrence matrix (GLCM) angular second moment (akin to variance)</td>
</tr>
<tr>
<td>Rel_Border_To_Image_Border</td>
<td>Rel. Border To determines the relative border length an object shares with the image border.</td>
</tr>
<tr>
<td>GLCM_Entropy</td>
<td>The value for entropy is high, if elements of GLCM are distributed equally</td>
</tr>
<tr>
<td>Rel_Area_Cell_Nuclei</td>
<td>Mean value of areas within cell nuclei.</td>
</tr>
<tr>
<td>Density_Cell_Nuclei_Stddev</td>
<td>Standard deviation of the densities of cell nuclei.</td>
</tr>
</tbody>
</table>

Table 4.2: Examples of morphological features of cancer tissue images.

In addition to the distribution and morphology of the nuclei, stromal components such as fibroblasts, ECM, cellular constituents of the vasculature, inflammatory/immune cells, and adipose tissue arrangement and interaction with the cancer cells are also essential to the tumor development and growth. Especially at the site of the primary tumor in the breast, the interaction between tumor cells and their surrounding milieu is reciprocal; tumor cells influence the stroma and vice versa, ultimately fueling tumor progression [130]. In our study, we also measure the structure of the stromal compartment of the microenvironment in a systematic way. Within each homogeneous region (Figure 4.1), we measured the spatial, texture, intensity and morphology of the tissue and build the
descriptive features of each of these patches, so that other components in the tissue were measured by these features. The complete list of features is given in Appendix A.

4.2.7 Correlations between Gene Expression and Tissue Morphology

mRNA expression profiles for the 44 selected TNBC tumors in TCGA were transformed from RPKM (Reads Per Kilobase per Million) normalized Illumina HiSeq 2000 RNA-seq readcounts. The mRNA data were preprocessed as follows: first, we selected genes with top 75% variance. Next, we clustered the mRNA expressions into K gene clusters (metagenes) using an iterative K-means clustering algorithm after 100 iterations. After examining the cluster homogeneity of these metagenes, we determined 50 (K=50) clusters that represented the measured gene expressions the best. Finally, each cluster was represented by its eigengene, which was defined as the first principal component from principal component analysis on the expression profiles of genes in this cluster [134].

Pearson Correlation Coefficients (PCC) between metagene expression and morphological features were calculated as $PCC(f_i, m_j) = \frac{\text{Cov}(f_i, m_j)}{\sigma_f \sigma_m}$, where $f_i$ is a feature vector for all samples and $m_j$ is the eigengene expression of metagene $j$ for the same samples. All pairs of correlations between features and eigengenes were obtained and then formed correlation matrix. Enrichment analysis on the selected metagenes were carried out using TOPPGene (http://toppgene.cchmc.org/).

4.2.8 Survival Analysis
Survival analysis was performed with OSU TNBC datasets, with median survival time of 75 months. Survival was calculated from the time of initial diagnosis of breast cancer to the time of death. Patients were divided into two groups as determined by feature values being greater or lower than the median value. Univariate Cox proportional hazards regression models were fitted to estimate the hazards of death among patients using each morphological feature. P-values were calculated based on univariate regression models to determine the significance of each covariate of interest, where p<0.05 was considered significant. Kaplan-Meier estimators were computed to plot the survival curves for covariates which were deemed to be significant. For survival analysis of metagenes on public datasets, a prognostic index of each patient is calculated by the sum of gene expression weighted by the hazard coefficients that are estimated by Cox proportional hazards model. After 10 times of 10-fold cross validation, the patients are divided by the 50 percentile of the tested prognostic index. The statistical significance is evaluated by log-rank test.

4.3 Results and Discussion

4.3.1 Translational Discover of Survival-related Morphological Features by Cancer-related Genes

In this study, we applied our proposed image analysis workflow on the more than 400 slide H&E images and extracted 37 previously-tested morphological features. We investigated the correlations between these features and the transcriptional expression profiles and found 23 significant positive statistically associations and 25 negative ones.
Analysis on the 143 OSU TNBC TMA images as the validation set unveils four morphological features that have strong correlations with survival. The corresponding gene clusters of these features were validated using two other independent datasets.

4.3.2 Metagenes with Strong Correlations to Morphological Features

Pearson Correlation Coefficients (PCC) between expression of the 50 metagenes and morphological features were calculated and shown in Figure 4.3(A). Forty eight (48) strong correlations (|PCC|>0.5) are highlighted in Figure 4.3(B). Examples of morphology-correlated metagenes are listed in Table 4.3, along with their major molecular functions and regulated human phenotypes obtained from enrichment analysis. Some of the metagenes strongly correlate with multiple morphological features. For instance, MetaGene_2 includes genes (e.g., MYOT, ACTA1) regulating molecular structural constituent of muscle motor activity and it controls the abnormality of protein fibers, which is the major component of the tumor microenvironment and is associated with fibroblast cells. It is noteworthy that MetaGene_2 negatively correlates with most morphological features.
Figure 4.3: Pairwise correlation heatmap between metagene expression and morphology of tissue in TCGA discover set. (A). Continuous correlation without threshold. The blue color demonstrates negative correlation; the red color demonstrates the positive correlation. (B). Thresholded correlation (|PCC| > 0.5).

<table>
<thead>
<tr>
<th>MetaGenes_ID</th>
<th>Top Molecular Functions</th>
<th>Top Human Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaGene_2</td>
<td>Structural constituent of muscle; motor activity</td>
<td>Myopathy</td>
</tr>
<tr>
<td></td>
<td>Cytoskeletal protein binding</td>
<td>Abnormality of muscle fibers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Muscle fiber cytoplasmatic inclusion bodies</td>
</tr>
<tr>
<td>MetaGene_13</td>
<td>Structural molecule activity; Structural constituent of cytoskeleton</td>
<td>Abnormal epidermal layer morphology</td>
</tr>
<tr>
<td>MetaGene_37</td>
<td>3',5'-cyclic-AMP phosphodiesterase activity</td>
<td>Smooth muscle contraction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Regulation of smooth muscle contraction</td>
</tr>
<tr>
<td>MetaGene_40</td>
<td>Zinc-finger transcription factor</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Examples of enriched Gene Ontology and human phenotype terms of the metagenes strongly associated with morphology in Figure 4.3.

4.3.3 Survival of TNBC based on Morphological Biomarkers
We conducted univariate survival analysis of the morphological features measuring variability of the TNBC tissue slides. Cox proportional hazard models were fitted based on patient’s survival time and morphological features. Survival tests of the top four predictive image features are shown in Figure 4.4. Feature ‘Area_Cell_Nuclei_stddev’ measures the standard deviation of the size of the nuclei. Higher values of this feature imply larger variations in nuclei sizes in the poorer prognosis group. This statistic was obtained by analyzing more than 30,000 cell nuclei of TNBCs in the validation cohort. Another marked characteristic of the poor survival group is the pixel density gradient among neighborhood of pixels. A high value indicates more dramatic deviations from the normal uniformly distributed tissue texture. The larger discontinuity, as noted in the survival curves among patients with poor prognosis, may result from larger proportion of tumor cells. Since we kept the tumor size bias as small as possible when the ROIs were selected and size measurements were normalized by the the size of the ROIs, the measurement bias is minimized.
Figure 4.4: Kaplan-Meier survival curves of prognostic model in OSU TNBC TMA. Higher values of the image features are plotted as blue lines, and lower values are plotted as red lines. (A) Survival on the two groups with distinct values of ‘Rel_Area_Cell_Nuclei’. (B) Survival of feature ‘Density_Cell_Nuclei_stddev’. (C) Survival of feature ‘Contrast_To_Neighbor_Layer_3’. (D) Survival of feature ‘Area_Cell_Nuclei_stddev’.

4.3.4 Survival Analysis of the Identified Biomarkers on Multiple Public Datasets

Most of the public breast cancer gene expression datasets do not possess histopathological images. In order to test the above morphological markers, we tested on these datasets using metagenes which are highly correlated with the above four predictive morphological features. In the TCGA data, three features (‘Area_Cell_Nuclei_stddev’,...
'StdDev_to_Neighbor_pixels, and 'GLCM_Correlation') out of the above four features have top correlations with one metagene (listed in the Table 3 in Appendix A). In Perou and NKI datasets, this metagene can separate the ER-negative patients into two cohorts with different outcomes. The Kaplan-Meier curves for this metagene was shown in Figure 4.5.

An interesting finding on these two validation datasets is that for the high risk cohort in Perou dataset, the survival drops dramatically after about 2 years (Figure 4.5B). These gene biomarkers are prognostic in both studies, with p-value 0.027 and 0.008, respectively.

Figure 4.5: The Kaplan-Meier curves for the metagene that shows the highest correlation with the image features. All time is represented in months. (A). Survival on NKI ER-negative patient subset. A higher risk group was revealed by this metagene. (B). Survival on Perou dataset ER-negative patient subset. This list of genes can separate the patients into groups with very significantly different outcomes.
Discussion: In this work, we present a workflow for correlating histopathological imaging features with gene expression profiles. Since TNBC is a subtype of breast cancer with poor prognosis and without clear predictive biomarkers, this study is part of our larger effort trying to establish more effective prognostics and predictive biomarkers for subtyping triple negative breast cancers by integrating morphological features with molecular/genomic profiles. By establishing the relationship between morphological features in histopathological images with gene expression profiles, not only can we derive novel insights on the molecular basis for different cell and tissue morphologies, but we also practically suggest that specific survival related molecular signatures can be manifest as morphological artifacts and features and thus avoid the cumbersome process of collecting gene expression profiles from patients. Moreover, among all the 37 previously-tested morphological features, the top prognostic ones are related to cell nuclei. Therefore, our study validated in a novel manner that nuclear features, which have been used in tumor grading in clinical practice, are prognostic in triple negative breast cancer.

To achieve this goal, we took further advantage of TCGA data, which proved to be an invaluable resource for such integrative genomic research. Even though the TCGA breast cancer data is relatively new and the follow-up time for patients is not long enough for effective survival analysis on TNBC patients, the matched histopathological images with gene expression profiles nevertheless provides the bridge for these two data modalities
allowing us to also use multiple modalities of data from different sources (e.g., OSU cohort as well as NCBI GEO).

Figure 4.6: An example of the TCGA slide image with its extreme values of predictive image features. (A) Boxplot of ‘Rel_Area_Cell_Nuclei’ feature and two patches with extreme values. (B) Boxplot of ‘Density_Cell_Nuclei_stddev’ feature and two patches with extreme values.

Specifically in this study we identified metagenes that are highly correlated with the four morphological features that can predict TNBC patient survival using the OSU cohort. We further determined that the features that best separated the better and worse survival groups were the cellularity of epithelial and the shape of the cancer cells. Essentially, the area of cancer cell nuclei and the diversity of the area of nuclei show great prognostic power for survival. To demonstrate a distinction between low-value
features and high-value features, the box plot of the area-based features and the patches with extreme values were shown in Figure 4.6.

While we currently do not have another independent datasets to validate these morphological features, we were able to test if the expression profiles of the associated metagenes have similar predictive power for multiple large datasets. We found that two of five gene clusters show predictive power in at least one of the test data. In particular, MetaGene_2 has strong predictive power in both public datasets. This metagene is enriched with cytoskeleton and fiber genes, which is not only consistent with its association with cell morphology, but also implies its potential roles in the development of tumor microenvironment including the stroma. The other metagene (MetaGene_13) is enriched with epidermal layer development indicating its association with tumor epithelial cells, which may explains its role in cancer development and relationship with tumor cell morphology. These observations strongly suggest that our approach effectively identified gene clusters that can partially explain the morphological characteristics and be used as predictive markers. Additionally, our methods to process whole tissue slides improve the current state-of-the-art pathological TMA image processing.

This study has several limitations: first, we only utilized the tissue slide images of TCGA to measure the patients’ phenotypes. The tissue slides were adjacent to tissues from which genomic data were derived. Thus, correlations between the morphology of these areas and genomic data are better reflected. However, these slide images are obtained from frozen sections and may possess larger artifacts than the diagnostic
images. Another limitation of this study is that we did not differentiate tumor epithelial, adipose, and stromal tissue in our measurement of the morphology. Classification of these cell types is an ongoing study and will be in future publication. Recently Yuan et al.[135] showed that immune cells can be used as an effective biomarker for prognosis suggesting tissue specific morphological features should be explored.

### 4.4 Summary

In this chapter, I present a novel workflow for discovering the associations between histopathological features with gene expression profiles. Our analysis reveals 48 pairs of strongly correlated morphological features and gene clusters. Four of the morphological features were identified as potential biomarkers separating TNBC patients into groups with different survival in a large validation cohort. Gene ontology analysis suggests that the high correlations are consistent with development and tumor related functions. Additionally, these morphological features on the tumor tissues can be extended as prognostic biomarkers for ER-negative breast cancers as the top gene cluster correlated with these morphological features was shown to be effective for predicting patient survival for ER-negative breast cancers in two independent public datasets. In addition to gene expression, other molecular data, such as proteomics data and microRNA expression have also shown to significantly improve the prognosis of TNBCs. In the following chapters, the correlations between imaging features and other types of molecular data are explored.
Chapter 5 : Integrative Genomic Analysis of Cancer Cell Morphology Regulatory Protein

The previous chapter identified biomarkers for TNBCs by associating morphological features with gene expression. In this chapter, I present another correlative study which extends the previous study by exploring proteomics data in breast cancer. The salient aspect of the study is that morphology of cancer cells and fibroblasts is correlated with proteomics data, thus enabling us to explore the protein markers related to cancer cellular morphology and possibly to the outcomes of the breast cancer patients. In order to link cellular phenotypes to the protein phenotypes, a comprehensive workflow is built and the difference between patient populations is modeled. Results from both clinical and biological evaluations are presented. The rest of this chapter is organized as follows. Section 5.1 provides motivations and related work. Section 5.2 describes the methods and data used in the studies. Section 5.3 provides the results of the study followed by summary in Section 5.4.

5.1 Introduction

Breast cancer is one of the most common cancers in the world with many subtypes [136]–[138]. Breast cancer represents the second leading cause of cancer death among women worldwide [139]. In addition, breast cancer is histologically diverse and has a large molecular heterogeneity. Therefore, its etiology is still poorly understood.
For comprehensive understanding of the breast cancer heterogeneity and seeking treatments, several levels of information including histological images and high-throughput genomics have been used to investigate breast cancers. Biomedical imaging has becoming crucial means to measure tumor and predict tumor response to drugs or other therapies [140][141]. Advanced quantitative imaging provides intrinsic and invisible biomarkers associated with treatment response. It has been reported as an effective and efficient biomarker in cancer studies, where the grade of the tumor has been examined based on only a small set of epithelial cells from 1920s [13]. On the other hand, functional proteomics represents a powerful approach to understand the pathophysiology and therapeutic treatment of cancer [142]–[144]. However, how proteo-transcriptomic biomarkers are translated to phenotypical level is yet unknown. Therefore, integrating quantitative imaging measurements and proteomics may potentially reveal interaction between proteins and morphology and the mechanism of how proteins regulate cancer development, and thus identify putative novel cancer associated pathways.

In this chapter, we used quantitative imaging features extracted from breast cancer tissue slides as a tool to measure the tumor burden. By identifying differentially expressed proteins between high-burden group and low-burden group, we aimed to reveal candidate morphology-associated protein regulated networks and chromosome regions. Finally, our study indicates that one of the genes in the networks - CHIP/Stub1 is associated with morphology besides involving in degradation of certain oncogenic proteins [145].
5.2 Methods and Materials

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n=85</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Diagnosis, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26-45 years</td>
<td>11</td>
<td>12.94</td>
</tr>
<tr>
<td>46-55 years</td>
<td>27</td>
<td>31.76</td>
</tr>
<tr>
<td>56-65 years</td>
<td>23</td>
<td>27.06</td>
</tr>
<tr>
<td>older than 66 years</td>
<td>24</td>
<td>28.24</td>
</tr>
<tr>
<td>Disease Stage, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>3</td>
<td>3.53</td>
</tr>
<tr>
<td>IA</td>
<td>5</td>
<td>5.88</td>
</tr>
<tr>
<td>IB</td>
<td>1</td>
<td>1.18</td>
</tr>
<tr>
<td>IIA</td>
<td>32</td>
<td>37.65</td>
</tr>
<tr>
<td>IIB</td>
<td>22</td>
<td>25.88</td>
</tr>
<tr>
<td>IIIA</td>
<td>11</td>
<td>12.94</td>
</tr>
<tr>
<td>IIIB</td>
<td>4</td>
<td>4.71</td>
</tr>
<tr>
<td>IIIC</td>
<td>5</td>
<td>5.88</td>
</tr>
<tr>
<td>IV</td>
<td>2</td>
<td>2.35</td>
</tr>
<tr>
<td>Tumor Grade, no (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>13</td>
<td>15.29</td>
</tr>
<tr>
<td>T2</td>
<td>54</td>
<td>63.53</td>
</tr>
<tr>
<td>T3</td>
<td>14</td>
<td>16.47</td>
</tr>
<tr>
<td>T4</td>
<td>4</td>
<td>4.71</td>
</tr>
<tr>
<td>ER, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>54</td>
<td>63.53</td>
</tr>
<tr>
<td>Negative</td>
<td>33</td>
<td>38.82</td>
</tr>
<tr>
<td>PR, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>46</td>
<td>54.12</td>
</tr>
<tr>
<td>Negative</td>
<td>39</td>
<td>45.88</td>
</tr>
<tr>
<td>HER2, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>23</td>
<td>27.06</td>
</tr>
<tr>
<td>Negative</td>
<td>42</td>
<td>49.41</td>
</tr>
<tr>
<td>Equivocal</td>
<td>14</td>
<td>16.47</td>
</tr>
</tbody>
</table>

Table 5.1: The demographics of CPTAC breast cancer patients. Clinical Proteomic Tumor Analysis Consortium (CPTAC) collects 105 breast cancer samples from TCGA, for robust, quantitative, proteomic technologies. Among these samples, ninety (90) of them contain tissue slides in TCGA repository. These samples were further matched with the samples in RNAseq repository. A total of 85 samples were found as intersection.

All human breast cancer samples (N=85) in the discovery cohort were collected from The Cancer Genome Atlas (TCGA) [92]. Proteomic data were collected from Clinical
Proteomic Tumor Analysis Consortium (CPTAC) [146], which contains TCGA Breast Cancer Mass spectrometry data using iTRAQ (isobaric Tags for Relative and Absolute Quantification) protein quantification methods. 105 TCGA tumor samples including 4 breast tumor subtypes: luminal A, luminal B, HER2E and basal-like were queried.

Relative quantification and protein identification were performed. Only peptides unique for a given protein were considered for relative quantitation. Proteins were identified on the basis of having at least one peptide with an ion score above 99% confidence. Under-expressed proteins and over-expressed proteins were identified based on their iTRAQ ratio. Tissue slide images were downloaded from the TCGA repository. DNA copy number variance data were extracted from SNP arrays. All datasets from the same patient samples were matched with their TCGA barcodes. The demographics of these patients are listed in Table 5.1.

In this study, we first stratified the breast cancer patients into a high malignant group and a low malignant group using the histopathological features we extracted from the tissue slides images. Then, a co-expression network algorithm was applied to find the protein networks. By comparing the eigen-proteins of the high and low malignant groups, we found protein networks that are highly differentially expressed between the two groups of patients. We further clinically tested the genes corresponding to these proteins in three public breast cancer datasets. Finally, we found the contribution of aberrations of DNA copy number variance to these proteins. The integrative analysis workflow is demonstrated in Figure. 5.1.
5.2.1 Image Processing and Feature Extraction

As stated in chapter 3, we downloaded ×40 magnification whole slide images from TCGA. The image files are in the SVS file (single-file pyramidal tiled TIFF) format. In general, the images have a resolution of 100,000×300,000 and they contain large amount of artifacts such as missing and folding tissues. Thus, to reduce the computational cost and to control the quality of the tissues, four representative image patches for each whole slide image were curated by manually selecting heterogeneous and informative regions. During the cropping process, a trained pathologist ensured the image patches contain
both tumor and tumor microenvironment. All patches varied between 3000 and 5000 pixels in width and height.

In our image processing pipeline, first, the tissue slides were segmented into patches called superpixels [89] with relatively homogeneous components in each superpixel. Second, for every superpixel, local binary pattern features were extracted, which were then used to classify the superpixels into stromal and epithelial tissues [40]. Epithelial tissues and stromal tissues were considered separately in the following feature extraction process. Epithelial features and stromal feature are listed in Appendix A. These features allow us to quantitatively evaluate the histology of the tissue samples.

5.2.2 Protein Co-expression Network Analysis and Differentially Expressed Proteins

Eigengene expressions [147] were calculated based on the co-expression networks. Then, eigengene expressions of low and high malignant groups of patients defined by image features were compared using student t-test. Significant differentially expressed protein co-expression modules were identified as p-value<0.05.

5.2.3 Survival Analysis of the Identified Biomarkers on Multiple Public Datasets

Due to the short follow-up time of TCGA breast cancer and relative smaller sample size of the CPTAC subset, survival analysis is not feasible on the discovery cohort. To evaluate the proteins we found in the analysis, we tested them on three most frequently used public breast cancer datasets. We matched the discovered proteins to gene symbols. Patients in Wang [148], NKI [149] and Perou [127] breast cancer datasets were divided into two groups based on the gene expression using KMeans algorithm [69]. Cox
proportional hazards regression models were fitted to estimate the hazards of death among patients between the groups. P-values < 0.05 was considered significant.

### 5.2.4 DNA Copy Number Variance Data Processing

Normalized level 3 DNA copy number variance data from SNP array were queried from TCGA. Locations of targeted genes were matched through hg19 reference genome, and then mapped to the hg19 version of the CNV data. The largest absolute segmented values were kept when multiple CNV alterations existed in one transcriptome location.

### 5.3 Results

#### 5.3.1 Patient Stratification based on Imaging Features

Patient groups identified by the malignancy of tumor, based on selected salient image features we computed for the patient sample tissue slides. These features are density of cancer cell nuclei (Nuclei_Density), mean area of cancer cell nuclei (Mean_Area_Nuclei) and fraction of fiber (Fraction_Fiber) on the tissue slides. These features were tested as related to breast cancer tumor malignance in our previous chapter [68].

![Hierarchical clustering of breast cancer patients based on imaging features. Patient samples (row) in TCGA data subset were labels using their barcodes.](image)

Figure 5.2: Hierarchical clustering of breast cancer patients based on imaging features. Patient samples (row) in TCGA data subset were labels using their barcodes.
A relatively high malignant group (N = 47, left part of heatmap in Figure 5.2) was identified, against the other relatively low malignant group (N = 38, right part of heatmap in Figure 5.2). It can been observed from the heatmap that cancer cell nuclei (Nuclei_Density) and mean area of cancer cell nuclei are consistent with each other, as larger nuclei area can be the result of denser nuclei density. Fraction of fiber is relatively proportion to the complement of cancer tissue, so it was selected as a control marker. Examples of slides in high malignant group and low malignant group are shown in Figure 5.3(A) and Figure 5.3(B), respectively. We can observe that the nuclei sizes in high malignant group are visibly larger than the low malignant group. Actually, the boxplot in Figure 5.3(C) shows the difference in nuclei size calculated based on features. As when we selected the regions, we keep the both tumor and microenvironment. Thus, the denser cancer is truly reprehensive to the tumor burden.
Figure 5.3: Nuclei size in comparison. (A) Example slides in high malignant group of breast cancer. (B) Example slides in low malignant group of breast cancer. (C). Boxplots of the mean area of cancer cell nuclei in low malignant and high malignant group.

5.3.2 Differential Expressed Proteins

One-hundred and twenty four (124) protein co-expression networks were identified on the 85 patient samples. Eigengene expressions of the protein co-expression networks from the high-malignancy group and the low-malignancy group were calculated and then compared to identify significant difference using student t-test. Four (4) gene networks differentially expressed between the high-malignancy group and the low-malignancy group were discovered with p-value less than 0.05. The discovered protein networks were shown in Table 5.2, along with their protein symbols, enriched cytoband (p-values) and gene ontology molecular functions. Distinct gene ontologies are associated with each
protein network. Protein network 1 contains MYOCD, MKL1, MKL2, FLYWCH2, CISD3, HN1L, STUB1 and SUCLG1, with most of the genes enriched in 17q and 16q13 regions. The top gene ontologies for the cluster are heavily enriched in smooth muscle cell differentiation, muscle cell differentiation and regulation of transforming growth factor beta receptor signaling pathway. Protein network 2 contains STK24, TPP2, TUBGCP3, DIS3, UCHL3, CARS2, IPO5, RANBP6, with most of the genes enriched in chromosome 13.

Table 5.2: Discovered protein networks that show differential expressions between the two identified groups, and their characteristics.

<table>
<thead>
<tr>
<th>Protein Networks</th>
<th>Proteins in the Networks</th>
<th>Cytoband</th>
<th>pValue</th>
<th>Enriched Cytobands</th>
<th>Top GO Molecular Functions (pValue)</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MYOCD</td>
<td>17p11.2</td>
<td>3.9E-2</td>
<td>17q, 16p13</td>
<td>smooth muscle cell differentiation(8.184E-7); muscle cell differentiation(1.889E-4); regulation of transforming growth factor beta receptor signaling pathway.</td>
<td>Enriched in 17q and 16q</td>
</tr>
<tr>
<td></td>
<td>MKL1</td>
<td>22q13</td>
<td>6.0E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MKL2</td>
<td>16p13.12</td>
<td>2.3E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FLYWCH2</td>
<td>16p13.3</td>
<td>2.2E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CISD3</td>
<td>17q12</td>
<td>1.4E-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HN1L</td>
<td>16p13.3</td>
<td>2.2E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STUB1</td>
<td>16p13.3</td>
<td>2.2E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SUCLG1</td>
<td>2p11.2</td>
<td>1.5E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>STK24</td>
<td>13q31.2-q32.3</td>
<td>2.3E-4</td>
<td>13q</td>
<td>Enriched in 13q</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TPP2</td>
<td>13q32-q33</td>
<td>4.6E-4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TUBGCP3</td>
<td>13q34</td>
<td>8.2E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DIS3</td>
<td>13q22.1</td>
<td>3.0E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UCHL3</td>
<td>13q22.2</td>
<td>1.8E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CARS2</td>
<td>13q34</td>
<td>8.2E-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IPO5</td>
<td>13q32.2</td>
<td>2.1E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RANBP6</td>
<td>9p24.1</td>
<td>8.1E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
Table 5.2 continued. Discovered protein networks that show differential expressions between the two identified groups, and their characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Protein</th>
<th>Chromosome</th>
<th>Fold Change</th>
<th>p value</th>
<th>Function</th>
<th>Role in Cancer Susceptibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>TBC1D15</td>
<td>12q21.1</td>
<td>2.3E-3</td>
<td>NA</td>
<td>iron incorporation into metallo-sulfur cluster (5.410E-4); metal incorporation into metallo-sulfur cluster (5.410E-4)</td>
<td>Mitochondrion</td>
</tr>
<tr>
<td></td>
<td>NUBPL</td>
<td>14q12</td>
<td>8.1E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NFS1</td>
<td>20q11.22</td>
<td>5.5E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LYRM4</td>
<td>6p25.1</td>
<td>2.9E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOCK5</td>
<td>8p21.2</td>
<td>3.9E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SLC25A35</td>
<td>17p13.1</td>
<td>1.7E-2</td>
<td>15q1-13</td>
<td>negative regulation of DNA metabolic process (2.023E-4); regulation of DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator (2.706E-4)</td>
<td>Role of BRCA1, BRCA2 and ATR in Cancer Susceptibility; Depletion of RAD17 sensitizes pancreatic cancer cells to gemcitabine;</td>
</tr>
<tr>
<td></td>
<td>ZNF385A</td>
<td>12q13.13</td>
<td>1.0E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RAD17</td>
<td>5q13</td>
<td>3.1E-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MSH3</td>
<td>5q11-q12</td>
<td>4.3E-4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KDM3B</td>
<td>5q31</td>
<td>1.6E-2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.3.3 Survival Analysis of the Identified Proteins

In order to test the prognostic power of the protein networks we identify, we performed survival analysis on three public-available datasets, since TCGA samples have inadequate follow-up time. We first matched the protein identifiers to the corresponding genes. Then, based on the four individual networks, we stratified the patients mRNA profiles into two groups on three public breast cancer datasets (Wang, Perou and NKI). Patient microarray profiles were clustered into two subgroups using Kmeans on the genes in cluster 3. All three datasets show significantly difference in survival between the two
groups (Figure 5.4). This result indicates that these morphology-associated genes have prognostic power in terms of overall survival.

Figure 5.4: Kaplan–Meier survival curves of prognostic model on multiple public breast cancer datasets. From left to right: Wang (A), Perou (B) and NKI (C) data respectively.

5.3.4 Alterations of the DNA

Among the 85 breast cancer patient samples, 84 of them have DNA copy number variance data. The correlations between DNA copy number variance and protein expression were calculated, with R-square values from linear fit. Figure 5.5 shows the scatter plots of proteins in protein clusters and their DNA copy number variance. Copy number changes contribute to 30% percent of protein expression on average. Copy number changes accounts for 30 percent of identified protein expression on average, compared to 13% average of the whole genome [150]. This suggests that the mechanisms resulting in nuclear aggressiveness have high associations with copy number variance.
5.3.5 STUB1 Protein Identified to be Associated with Morphology

We performed gene network analysis on the protein clusters identified. Interestingly, the first protein clusters contains a gene named STUB1, which is often recognized as a neurological disease gene. The boxplots of its protein expressions in low and high malignant groups were shown in Figure 5.6(A). This is indeed the case that the overexpression of STUB1 suppresses tumor progression, confirming the recent finding.
that CHIP targets Src-3, hypoxia inducible factor 1α, NF-κB, ErbB2 and c-Myc [145]. In addition, gene network analysis (cbioportal, http://www.cbioportal.org/) shows that it interacts with important breast cancer genes such as ERBB2 (HER2), ESR1 and TGFBR1 (Figure 5.6(B)). Specifically this is consistent with the discovery of Lv et al [151] that over-expression of STUB1 leads to apoptosis escaping.

Figure 5.6: Characteristics of protein network 1. (A). STUB1 protein expression in low and high malignant groups. (B). Gene network analysis identified using cbioportal. Proteins identified in cluster 1 are marked using darker circles.

5.4 Summary

In this chapter, we investigated protein biomarkers for prognostic and phenotypic of breast cancer. We proposed an integrative workflow first stratified the patient samples
into histological heterogeneous groups; and then identified 4 protein clusters that
differentially expressed between the groups. Finally, we verified their significant roles
using public dataset. Our results show that identified protein networks are enriched in
certain chromosome regions. DNA copy number variance analysis confirms contribution
of the copy number alteration to the protein expression. Protein network analysis suggests
that CHIP/Stub1 gene is an important gene that has potential link to the nuclei size of
cancer cells.
Chapter 6: Multivariate Model of Morphology and Transcriptome

In Chapter 4 and Chapter 5, histopathological strategies for associating image features with clusters of co-expressed genes and proteins are developed. Pairwise correlation map and differentially expressed genomic/proteomics features are utilized for building the association. However, the pairwise correlation adopted in this application assumes that one aspect of morphology is regularized by one molecule at a time, which is not always the reality, given the complexity of biology. In fact, many pathways and kinase families are found to be associated with cell morphology such as cell rounding, cell size, and tissue organization [152]. Cancer-specific cell morphology regulators are yet to be identified.

To address the multi-regulatory problem of morphology, in this chapter, a multivariate model is built to mine correlations between cellular morphology and molecular expressions. Prognostic significance of the predicted image features of lung adenocarcinoma is evaluated. Then, the associated gene signatures identified from the multivariate model are also validated using other datasets.

6.1 Introduction

6.1.1 Background

Being the most common histological type of lung cancer, lung adenocarcinoma links to smoking; however, an estimated 10-25% of lung adenocarcinomas are patients who never smoke, or smoke less than 100 cigarettes in their lifetime [153]. Despite the fact
that it is a sub-classification of lung cancer, Lung adenocarcinoma is a heterogeneous group of tumors with a highly variable prognosis [154].

The high-through-out sequencing technologies are making targeted therapies possible for lung adenocarcinoma. The once conceptual personalized medicine is no longer a practical challenge in treatment of patients with lung adenocarcinoma [155]. The advance of these technologies allows molecular diagnostic biomarkers for the detection of lung cancer in addition to computed tomography (CT) screening [9], [156]–[158]. For example, the utility of epidermal growth factor receptor (EGFR) mutation testing is strongly recommended [159] in clinical practice. V-Ki-ras2 Kirsten rat sarcoma oncogene homolog (KRAS) mutation accounts for 90% of RAS mutations in lung adenocarcinomas, with mutations detected in about 25% of all tumors. However, direct inhibition of KRAS mutations has been proven to be clinically challenging [160]. There is no successful therapeutic approach for these hypothesized mutations, so no effective treatment has benefited the patients. In addition, although EGFR-mutant lung cancers are sensitive to EGFR tyrosine kinase inhibitors (TKIs), they develop resistance [161]. Therefore, novel targets for lung adenocarcinoma are needed for enhanced personalized medicine.

Lung cancer diagnostic and classification have been traditionally based on elevation based on imaging approaches, such as CT and histopathology [162], [163]. For instance, five (5) distinct histologic subtypes and radiologic patterns have been reported recently. Traditionally, histopathology images serve as a golden standard for lung cancer
diagnosis. Cellular and inter-cellular level morphology are usually used by the pathologists for making diagnosis decisions. However, the current pathology diagnosis is commonly based on individual pathologists’ interpretations of the samples which are subject to large inter-observer variations and low throughput analysis. Unbiased quantitative pathology methods are showing promise by offering more cellular information[13], [41], [135]. Recently, pathology informatics study on lung cancer has attracted more interests. In one study [164], the diagnostic significance of nuclear features in differentiating small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) was investigated. Edwards et.al.[165] showed that adenocarcinoma diagnosis is more challenging compared to squamous carcinoma. An early automatic pathology analysis system was proposed in [166]. In the study of Mijović et al. [167], diagnostic values of seven Karyometric variables are examined for diagnosis of major histological types of lung carcinoma. In Zhang et al’s study [168], an image classification system is proposed to differentiate lung adenocarcinoma and squamous carcinoma. Compared to genomic biomarkers, advanced imaging may provide more clinically relevant information.

6.1.2 Motivation

To combine both the richness of histopathological information and molecular profiles, we developed an integrative computational pipeline that exploits diagnostic images and mRNA expression. The pipeline allowed us to discover the associations between cellular level and molecular level phenotypes, and thus novel biomarkers can be unveiled. In this
chapter, we extracted 282 histopathological features from Lung Adenocarcinoma (LUAD) tissue slides and initially attempted to identify co-expression gene clusters that have high correlation with these image features. The descriptions of the LUAD features and dataset are demonstrated in the following section. However, the pairwise correlations are below 0.3 for either Spearman or Pearson, regardless of the variation of parameters and the normalization techniques. Figure 6.1 shows the top two ‘highly-correlated’ pairs between the imaging features and gene clusters. The maximum correlations are under 0.3 for both Spearman and Pearson. It is skeptical that these lung adenocarcinoma morphological variances may not be regulated by any group of genes; instead they are resulted by multiple groups. Based on these quantitative experiments, we hypothesize that a multivariate model needs to be built.

Figure 6.1: Top 'high-correlated' pairwise correlations between eigengene expression and morphological features of LUAD. (A). Scatterplot of skew5 feature and eigengene cluster 191. (B). Scatterplot of std16 and eigengene 239.
6.2 Methods and Materials

Our analysis involves with molecular and histological analysis based on data from TCGA. The data modules of lung cancer adenocarcinoma we use are mRNA profiling, histological images and clinical data including survival information.

6.2.1 Integrated Analysis Pipeline

We collected matched diagnostic images and gene expression data for a discovery dataset of 202 lung adenocarcinoma patients. Our automatic imaging processing pipeline detected cell nuclei and extracted predefined features evaluating staining variations. Molecular expression of mRNAs was filtered and clustered using co-expression network algorithm. To select imaging features with clinical relevance, survival-related imaging features were filtered. Significant clusters of genes were represented by eigen-gene expressions. Then, we built a lasso regression model to select gene clusters that regulate the image features. By finding the co-expression patterns that correlate with imaging features, we can discover cellular processes associated with cancer cell morphology. The integrative analysis pipeline was shown in Figure 6.2.
6.2.2 Image and Genomic Data Collection

We focus on a data set downloaded from TCGA Data Portal (The Cancer Genome Atlas) that provides cancer specimens including clinical information, genetics information, and histopathologic whole slide images of the patients. 201 patients with adenocarcinoma are used for the experiments. For each patient, an image patch of size 1712 x 952 is cropped from the tumor region.

Expression profiles of 20,531 unique genes were investigated in 201 patients with lung cancer adenocarcinoma (LUAD) from TCGA. The TCGA RNA-seq data sets are publicly available from the TCGA website [92].

6.2.3 Data Preprocessing and Imaging Feature Extraction
We adopt the cell detection and segmentation methods proposed in [169]. In the cell detection stage, a radial voting scheme with Gaussian pyramid is employed. For each image, a Gaussian pyramid is created. A single-pass voting [170] is applied to each layer. The voting region receives scores weighted by a distance transform. Therefore, such weighted voting encourages the pixels closer to the cell center accumulates higher voting scores. The final voting score is obtained by summing up the voting scores from different layers. In the segmentation stage, a marker based active contour with a repulsive term is applied to the images using the detection results as the markers. An initial contour associated with each detected marker is created first. The contours evolve through an iterative procedure to reach the real boundaries of the cells. The repulsive term serves to prevent the contours from crossing and merging with each other. Specifically, our collaborators extracted the following features from lung adenocarcinoma histology:

**Group 1: Geometry Features.** Based on the segmentation results, five geometry features are calculated for each lung cancer cell to capture the cell shape information, including cell area, contour perimeter, circularity, major-minor axis ratio, and contour solidity. Contour solidity is defined as the ratio of the area of a cell region over the convex hull defined by the segmentation boundary.

**Group 2: Pixel Intensity Statistics.** Pixel intensity statistics features are used to capture the color of the segmented cells. This group of features are calculated based on the intensity of the pixels within the segmented cells, including intensity mean, standard
deviation, skewness, kurtosis, entropy, and energy. *Lab* color space is used for a better color representation.

**Group 3: Texture Features:** Texture is an important feature found to be closely related to cancer diagnosis in radiomics. This is rooted in the fact that texture patterns are linked to difference in protein expressions [171]. This group of features consists of co-occurrence matrix [113], center symmetric auto-correlation (CSAC) [172], local binary pattern (LBP) [173], texture feature coding method (TFCM) [174]. The co-occurrence matrix [113] computes an estimation of the joint probability distribution of the intensity of two neighboring pixels. CSAC is a measure of the local patterns with symmetrical structure. These patterns are characterized by a series of local auto-correlation and covariance introduced in [172], including symmetric texture covariance (SCOV), variance (SVR), and within-pair variance (WVAR), and between-pair variance (BVAR). 3×3 pixel unit of each channel is considered. LBP [173] feature measures the local textures by assigning a binary code to a pixel with respect to its intensity and those of its neighboring pixels. A histogram of the generated binary codes reveals the distribution of the present repeated local patterns. Similar to LBP, in TFCM [174], a texture feature number (TFN) is assigned to each pixel by comparing this pixel with its neighbors in four directions: 0°, 45°, 90°, and 135°. A histogram is calculated based on the TFNs of one image patch.

Gene Expression: The expression profiling study involved 201 samples with primary lung cancer adenocarcinoma from TCGA. Expression profiles of mRNA were converted
from mRNAseq with RPKM (read per kilobase per million) values downloaded from TCGA data portal. The mRNA expressions were log2 transformed and standardized.

### 6.2.4 Associations between Morphology and Transcriptomes based on Lasso Regression

We model the expression of imaging features as regulatory results of transcriptomes using lasso regression. Lasso regression model minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant.

Consider the linear regression model: we have \((x_i, y_i), i = 1, 2, ..., N\), where \(x_i = (x_{i1}, ..., x_{ip})^T\) and \(y_i\) are eigen-gene expression and image feature value for the \(i\)th observation (patient sample), respectively. With regular regression model, the least square estimates are obtained by minimizing the residual squared error. However, in feature selection models to predict biomarkers, only imperative transcriptomes contribute to biological functions and processes, requiring more stringent and interpretable features. With large number of features, we would like to determine a small subset of them that can predict strong correlations. Let \(\beta = (\beta_1, ..., \beta_p)^T\) and \(\beta_0\) to be a scalar. The lasso model estimate \((\beta, \beta_0)\) by solving the following problem

\[
\min_{\beta, \beta_0} \left( \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),
\]

where \(\lambda\) nonnegative regularization parameter. As \(\lambda\) increases, the number of nonzero components of \(\beta\) decreases, leading to smaller numbers of predictors.

### 6.2.5 Identification of Survival-Related Image Features
Univariate Cox Proportional Hazard models are used to identify morphological features and genes that have expression related significantly to survival. Morphological features that have p-values less than 0.05 are considered to be significant.

6.3 Results

6.3.1 Survival-related Image Features and Gene Cluster

Using a univariate Cox proportional hazards regression model, we assessed the image features related risk score in the prediction of the LUAD patient survival. Significant morphological features are listed in Table 6.1. Within the 6 types of imaging features, tfcm category shows the most significant prognostic power, indicating texture features in lung adenocarcinoma have strong potential to predict patients’ outcomes. In fact, top 6 survival-related imaging features are in tfcm category. Other features that capture prognosis by cellular behavior are contrast2, contrast, csac23, fourier15, csac5 and entry4.
<table>
<thead>
<tr>
<th>Feature Names</th>
<th>p-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfcm4</td>
<td>0.00456904</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm9</td>
<td>0.00532429</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm3</td>
<td>0.00563955</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm1</td>
<td>0.0064998</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm2</td>
<td>0.00657692</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm10</td>
<td>0.00685436</td>
<td>yes</td>
</tr>
<tr>
<td>contrast2</td>
<td>0.0082282</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm8</td>
<td>0.0093341</td>
<td>yes</td>
</tr>
<tr>
<td>contrast1</td>
<td>0.01210092</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm12</td>
<td>0.01247155</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm11</td>
<td>0.01361604</td>
<td>yes</td>
</tr>
<tr>
<td>csac23</td>
<td>0.01754474</td>
<td>yes</td>
</tr>
<tr>
<td>tfcm7</td>
<td>0.0178572</td>
<td>yes</td>
</tr>
<tr>
<td>fourier15</td>
<td>0.0178766</td>
<td>yes</td>
</tr>
<tr>
<td>csac5</td>
<td>0.01896244</td>
<td>yes</td>
</tr>
<tr>
<td>entry4</td>
<td>0.01995154</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 6.1: Prognostic values of various image features in discover dataset. The features are listed by their significance in the survival model.

6.3.2 Correlations between Image Features and Gene Clusters

The image-based pipeline allowed us to quantitatively analyze tumor characteristics on cellular level and associate these tumor characteristics to patient outcomes. On the other hand, gene co-expression network analysis allowed us to explore biological functions and processes on molecular level. To explore the relationships between the co-expression networks and morphological features, we built a lasso-based regression model. The lasso model summarizes pairwise correlations between images features and gene networks. We rank the importance of image features by their significance in survival analysis. For example, the most survival-associated feature is tfcm4, which correlates
with 12 gene networks. Table 6.2 lists the genes in this network and their coefficients in the lasso linear regression model.

<table>
<thead>
<tr>
<th>Gene Clusters</th>
<th>Image Feature Category Associated</th>
<th>GO Biological Process/p-values</th>
<th>Cytobands/p-values</th>
<th>Notes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>tfcm</td>
<td>microtubule-based movement, 5.484E-7; cillum movement, 1.428E-6</td>
<td>9p13.3, 4.976E-5</td>
<td>Genes up-regulated / down-regulated in nasopharyngeal carcinoma (NPC) compared to the normal tissue</td>
</tr>
<tr>
<td>10</td>
<td>tfcm</td>
<td>plasma lipoprotein particle remodeling, 5.696E-11; protein-lipid complex remodeling, 5.696E-11</td>
<td>3q27, 9.452E-6</td>
<td>extracellular space</td>
</tr>
<tr>
<td>28</td>
<td>tfcm</td>
<td>embryonic skeletal system morphogenesis, 1.016E-5; embryonic skeletal system development, 3.006E-5</td>
<td>7p15.2, 4.113E-8</td>
<td>Amplicon 7p15 identified in a copy number alterations study of 191 breast tumor samples</td>
</tr>
<tr>
<td>112</td>
<td>tfcm</td>
<td>ion transport, 1.486E-4;</td>
<td></td>
<td>Human Lung Hou10 187genes</td>
</tr>
</tbody>
</table>

Table 6.2: Gene clusters showing strong correlation with texture image feature tfcm, and their Gene Ontology terms and enriched cytobands.
Several morphological features have high correlations with multiple sets of gene clusters. Among them, several salient observations were noticed. For example, gene cluster 5 was enriched with microtubule-based movement and cilium movement; gene cluster 10 genes were linked to extracellular space; and gene cluster 28 was enriched in embryonic skeletal system morphogenesis. These particular associations suggest that these identified gene clusters might be related to texture morphology of the lung cancer adenocarcinoma.

6.3.3 Cellular Cilium Movement Genes Correlate with Morphology

Among the gene clusters discovered to have high correlation with texture image features, 24 genes are enriched with cilium. Detailed gene list is shown in Table 6.3. In fact, 3 out of the 24 genes show significant prognostic power in the discovery dataset (Table 6.3.). Cilia are microscopic, finger-like projections that stick out from the surface of cells. There are two types of cilia: primary cilia (non-motile), which are sensory organelles present in almost all cells; and motile cilia, which move liquid over the cell surface. Mutations of cilium genes are causes of primary ciliary dyskinesia, which result in defective cilia that move abnormally or are unable to move (immotile). The fact that these cilium genes are also associated with salient imaging features in histopathology suggests the cilium genes’ importance in lung adenocarcinoma.
<table>
<thead>
<tr>
<th>Gene Symbols</th>
<th>p-values</th>
<th>Significance</th>
<th>Matched in validation (GSE42127/GSE11969)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKAP14</td>
<td>0.350331</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>CATSPER1</td>
<td>0.985332</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>CCDC114</td>
<td>0.403194</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>CCDC39</td>
<td>0.450552</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>DNAH1</td>
<td>0.023958</td>
<td>yes</td>
<td>A/NA</td>
</tr>
<tr>
<td>DNAH12</td>
<td>0.043225</td>
<td>yes</td>
<td>NA/NA</td>
</tr>
<tr>
<td>DNAH6</td>
<td>0.304484</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>DNAH9</td>
<td>0.224714</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>DNAI1</td>
<td>0.858528</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>DNAI2</td>
<td>0.154144</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>ENKUR</td>
<td>0.955451</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>HYDIN</td>
<td>0.410639</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>PACRG</td>
<td>0.642716</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>PTCHD3</td>
<td>0.93942</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>RABL2A</td>
<td>0.354717</td>
<td>A/NA</td>
<td></td>
</tr>
<tr>
<td>RSPH1</td>
<td>0.970006</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>RSPH4A</td>
<td>0.568054</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>SNTN</td>
<td>0.055948</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>SPEF1</td>
<td>0.127674</td>
<td>NA/NA</td>
<td></td>
</tr>
<tr>
<td>TEKT1</td>
<td>0.305199</td>
<td>A/A</td>
<td></td>
</tr>
<tr>
<td>TEKT2</td>
<td>0.004016</td>
<td>yes</td>
<td>A/A</td>
</tr>
<tr>
<td>TLL10</td>
<td>0.198528</td>
<td>A/NA</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Cilium genes are discovered to have high associations with top survival-related image features – tfcm.

### 6.3.4 Prognostic Validation

Validation on heterogeneous external data sets allows for evaluation of the generalizability. To test the importance of cilium-related genes, we further performed survival analysis on two public-available datasets. Among the 24 genes correlated with
the image feature category tfcm, 13 of the gene symbols can be matched exactly to the external datasets. In the validation datasets, lung adenocarcinoma patients were stratified into two groups using Kmeans based on their expression levels of the 13/11 matched cilium genes (Table 6.3). In both datasets, a statistically significant group of patients with worse outcomes were differentiated (n = 22 and n = 63, respectively). Figure 6.3 shows the survival curves of the two patient cohorts.

![Survival Curves](image1.png)

Figure 6.3: Kaplan–Meier survival curves of prognostic model on two public lung adenocarcinoma datasets. Survival curves of two groups with distinct values of cilium gene expressions with p-values = 0.042 and 0.028 respectively.

### 6.3.5 Alteration of the Cilium Genes

In order to investigate the mutation of the cilium genes in lung adenocarcinoma, we examined the lung adenocarcinoma samples in the TCGA. Among the 230 samples, the
cilium genes altered in 100 (43%) patients. Figure 6.4 demonstrates the OncoPrint of these cilium genes obtained from cbioportal([http://www.cbioportal.org/](http://www.cbioportal.org/)).

Studies have shown that multiple components of the Shh (Sonic hedgehog) and PDGFRα (platelet-derived growth factor receptor-α signal transduction) pathways localize to primary cilia, and that cilia are essential for ligand-induced cell signaling by both pathways[175]. Mutual exclusivity shown in Figure 2 indicates the complexity of cilium genes. In particular, DNAH9 gene altered in 18% of the cohort in variety of patterns, including 1 case of amplification, 4 cases of homozygous deletion, 24 cases of missense mutation and 12 truncating mutation.

Figure 6.4: Mutation plot of cilium-related genes in TCGA lung adenocarcinoma cases.
6.4 Summary

Our integrative analysis pipeline allows us to find survival related textural features of lung adenocarcinoma. In addition to the image features, we also show that cilium genes have strong associations with these features, from our multivariate model of cellular and molecular phenotypes. Our finding indicates that the morphology may be the marker for monitoring cilium gene activation and related cancer pathways.
Chapter 7: Integrative Stratification via a Two-step Consensus Clustering

In recent decades, many studies have led to biomarkers for cancer outcome predictions, which assist clinicians on selecting the right treatment strategy. These biomarkers include both pathological attributes and various types of omics data. However, there is a lack of a unified means for patient stratification which can effectively integrate the heterogeneous types of molecular and clinical data and improve accuracy on patient outcome prediction.

In this chapter, we propose a novel two-step cancer patient stratification workflow, which aggregates clinical information and molecular expression profiles. First, stratification based on gene and miRNA expression profiles is achieved by an integrative clustering approach. Then, the task of comprehensive patient subtyping based on both clinical attributes and the molecular stratification is formulated and solved as a consensus-clustering problem. The results obtained from the TCGA breast cancer study suggest that this approach out-performs stratification based on any single clinical attributes or molecular data type.

7.1 Introduction

In recent years, the advanced researchers have brought personalized medicine into a brand new level by introducing technologies such as DNA microarray, mRNA
sequencing and miRNA sequencing [176], [177]. These genome-wide studies have heralded new era in clinical practice to account for variations across population. Risky groups classified by the status of the biomarkers are given more aggressive treatments, and prediction of cancer growth, proliferation and metastasis is also based on these ‘omic-generated’ biomarkers. Specifically, many types of cancer can be reclassified into distinct subtypes based on the genomic makeups and molecular profiles and these subtypes present different clinical outcomes including prognosis and response to treatments [1], [178], [179]. An example of success is the PAM50 gene panel being applied to breast cancer patient stratification[1].

Although the recent-developed high-through-put data have associated subtypes of cancers to these prognostic and predictive biomarkers, this data-driven approach of uncovering gene signatures is often hard to interpret due to the knowledge limitation about the biological functions of these biomarkers [180]. In addition, traditional stratifications of patients based on cancer staging, histological subtyping and other pathological clinical attributes are still the essential diagnosis means in clinical practice. However, these clinical knowledge-based classifications do not necessarily agree with the molecular stratifications. Furthermore, the multiple labels for each patient can be confusing and difficult to interpret. With all available patient stratification methods, there is a need for a consensus classification of patients implemented by both clinical information and molecular biomarkers.
Therefore, in this chapter we aim to develop a novel method that can integrate information of both clinical attributes and molecular data to provide a consensus clustering for patients. We focus on two major challenges on using ensemble of molecular expression data and clinical attributes: (1) Integrative clustering of the multiple molecular expression data; (2) Aggregation of all base-clustering labels. Specifically, we propose a novel two-step consensus clustering workflow integrating both the molecular clustering and clinical attributes. This workflow resolves the first challenge using a Canonical Correlation Analysis (CCA) based method to cluster the multi-types of data. Then it addresses the second challenge with a Bayesian consensus clustering method. The results obtained from applying this approach in The Cancer Genome Atlas (TCGA) breast cancer data demonstrated the power for this approach in differentiating the patient population with different outcome and its potential in integrative genomics studies.

7.2 Methods and Materials

7.2.1 Dataset

TCGA is a project aiming to demonstrate genetic mutations for various types of cancers using multiple genomic data [92]. mRNA and miRNA expression profiles were collected from 441 primary tumors of breast cancer patients who have all mRNA, miRNA and clinical data. The median follow-up is 3 years. Expression profiles of mRNA and miRNA were converted from level-3 mRNAseq and miRNAseq with RPKM (read per kilobase per million) values downloaded from TCGA data portal. The mRNA and miRNA expressions were log2 transformed and normalized. The clinical attributes
(disease stage, tumor grade and histological type) were discretized from 1 to the number of categories. Disease stage determines the extent to which a cancer has developed by spreading. Tumor grade measures cell appearance in tumors and other neoplasms. Histological type categorizes the heterogeneity found in breast cancer based on architectural features and growth patterns.

In order to eliminate the noise in the mRNA expression profiles, we selected 70 prognostic genes and 7 prognostic miRNAs from previous studies [128][181]. Out of the 70 genes, we found 28 genes exactly matched (Table 7.1) with genes in TCGA data.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Prognostic Biomarkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Expression</td>
<td>AKAP2, AP2B1, BBC3, CCNE2, CENPA, COL4A2, DCK, ECT2, ESM1, EXT1, FGF18, FLT1, GMPS, GNAZ, GSTM3, IGFBP5, MCM6, MMP9, NMU, OCR6L, PECI, PRC1, RAB6B, RFC4, SERF1A, SLC2A3, TGFBI3, WISP1</td>
</tr>
</tbody>
</table>

Table 7.1: Prognostic mRNAs and miRNAs used in the integrative molecular clustering.

### 7.2.2 Workflow

Instead of manually integrating separate patient clusters, our workflow achieves data integration through consensus clustering which combines all the cluster information as a unified stratification, which is a single partition of patient samples obtained by integrating multiple measurements. Prior to this study, to our knowledge there is no bioinformatics method that can integrate both numerical molecular measurements and categorical attributes. Current cluster ensemble algorithms, such as the cluster-based
similarity partitioning algorithm (CSPA) [20], hypergraph partitioning algorithm [20], or Bayesian based algorithms [182] are able to accomplish the integration of clusterings. However, none of them was designed to address the integration with numerical measurements such as gene expression. And, most of the integrative analytical methods require specific data type, but our method can be applied any type of omic data and other clinical attributes of the patients, such as distant metastasis status and lymph node infiltration.

Figure 1 provides an overview of the framework for the proposed patient classification approach. Specifically, in the first step, the salient microRNAs and mRNAs are selected and the patients are clustered based on the expression of these salient molecules. There are a few integrated unsupervised clustering algorithms developed to fulfill this task [23], [183], [184]. Considering the relative large sample size in our study, we adopt the CCA-based multiple-view clustering algorithm. Then, in the second step the cluster labels generated from the molecular expression profiles are jointly considered with clinical labels by consensus clustering, and a final label is constructed for each patient. Since it is not clear which measurement of similarity between two distinct clustering with both molecular and clinical data being used, a Bayesian Cluster Ensemble [182] method is selected.
7.2.3 Integrative Molecular Stratification using Canonical Correlation Analysis

In the first step, we have multiple measurements (molecular expressions in this study) from \( n \) modalities, with modality \((j = 1, 2, ..., n)\) from a mixture of \( k \) Gaussians \((D_{1j}^{j}, ..., D_{kj}^{j})\). The goal of integrative molecular clustering is to recover the \( k \) sub-distributions representing distinct molecular subgroups.

In this study, the molecular data are gene expression and miRNA expression, so there are two sources in this setup \((n = 2)\). In particular, we assume \( i \)-th sample can be represented as \( x_{i} = (x_{i}^{(1)}, x_{i}^{(2)}) \), where \( x_{i}^{(1)} \in \mathbb{R}^{d1} \), \( x_{i}^{(2)} \in \mathbb{R}^{d2} \). \( d1 \) and \( d2 \) are the dimensions of two types of data respectively. And, \( \mu_{i}^{(1)} \) and \( \mu_{i}^{(2)} \) are the means of distribution \( i \) in measurement 1 and measurement 2, respectively. In genomic data
preprocessing, the expression data is usually normalized and centered at zero mean, so 
\( \mu^{(1)} = \mu^{(2)} = 0 \). Hence, we can define the covariance matrix of the measurements 
\( x = (x^{(1)}, x^{(2)}) \) as

\[
\Sigma = \begin{bmatrix}
\Sigma_{11} & \Sigma_{21} \\
\Sigma_{12} & \Sigma_{22}
\end{bmatrix} = \begin{bmatrix}
E[x^{(1)}x^{(1)\top}] & E[x^{(2)}x^{(1)\top}] \\
E[x^{(1)}x^{(2)\top}] & E[x^{(2)}x^{(2)\top}]
\end{bmatrix}
\]

If we assume that the two measurements are independent to each other, which usually holds for miRNA expression and mRNA expression, then for each distribution \( r \) in the mixture Gaussian:

\[
E[x^{(1)}x^{(2)\top}] = E[x^{(1)}|r = i]E[x^{(2)\top}|r = i] = \sum_i \omega_i E[x^{(1)}] E[x^{(2)}] = \sum_i \omega_i \mu_i^{(1)} \cdot \mu_i^{(2)\top}.
\]

Previous publications[23] [185] showed that the column spans of matrix \( U \) and \( V \) from reduced SVD of \( E[x^{(1)}x^{(2)\top}] = UDV^* \) are subspaces spanned by the means in measurements 1 and 2, respectively. Therefore, following the method proposed by Chaudhuri et al.[23], the integrative clustering of mRNA and miRNA expression can be described as followed:

1. Partition the samples into two subsets \( X \) and \( Y \) randomly.

2. Compute the empirical covariance matrix \( \Sigma_{XY} \) between mRNA and miRNA expressions.

3. Compute the singular vectors of \( \Sigma_{12}(X) \) (\( \Sigma_{12}(Y) \) respectively) and project \( Y \) \( (X) \) on subspace spanned by the top \( k - 1 \) left singular vectors.

4. On the projected data, apply any single-linkage clustering algorithm.
7.2.4 Obtaining the Final Patient Stratification from Consensus Clustering

With all patients assigned to a certain clustering based on multi-source molecular data, the information of molecular heterogeneity is encoded into the labels of this population. In the second step, let us consider the method to integrate the molecular labels with the clinical attributes. Given $M$ sets of clustering $C_j = \{c_{ij}, [i]^N\}$ drawn from $N$ objects $D = \{d_i, [i]^N\}$, the cluster label vector for object $d_i$ is denoted by $x_i = \{x_{ij}, [j]^M\}$.

The goal is to construct a model that can find the consensus cluster integrating these $M$ sets of clustering. There are many algorithms to solve this problem. General speaking, these methods fall into two categories: probability model and similarity model. The former one postulates a probability model to determine the labels of the objects and solve via maximum likelihood optimization, while the latter one seeks the consensus clustering with the largest concordance (similarity) with original clusterings. The second approach requires a measurement of similarity between two distinct clustering, which is usually not trivial to find, thus we adopt a Bayesian model[186] in this study.

7.2.5 Assessing Parameterization of Consensus Clustering

To determine the parameter $k$, the number of final clusters in consensus clustering, we aim to optimize the similarity between the final clustering and the base clusterings. Here, normalized mutual information function is used to measure the similarity between clusterings. In other words, the parameter $k$ can be found, so that the similarity between the final cluster and all based clusterings is maximized.

7.2.6 Survival Analysis
For each set of patient classification, a low-risk group and a high-risk group were identified as the best separation from combinations of the resulting subgroups. Survival was analyzed according to the Kaplan-Meier method. The differences between survival distributions of the two groups were evaluated by log-rank test. A supervised analysis was performed using ANOVA (false-discovery rate of 5%) followed by Tukey post-hoc testing to identify genes with differential expression between pre-defined groups.

7.3 Results

We introduce this two-step model to address the challenge of integrative patient subtype identification based on both molecular expression and clinical attributes. In this section, the proposed workflow is applied to identify integrative patient stratification to breast cancer patients in TCGA. In this section, the identified breast cancer patient subgroups were evaluated computationally, clinically and biologically.

7.3.1 Selection of the Number of Clusters

It is desirable to partition the patient population into groups that can maximize the mutual information between the final partition and the base clusterings. After the integrative molecular partition was obtained, different choices of the parameter $k \in [2, 8]$ were tested by evaluating the normalized mutual information defined in the above section. We repeat the clustering for $r = 10$ times. In the case when $k = 4$, the final clustering has the largest mutual information with the base clusterings (Figure 7.2), so in the rest of the chapter, we use $k = 4$. 

125
7.3.2 Integrative Clustering of Multiple Types Molecular Expression Data

In this section, we will show the superiority of using multiple types of molecular expression data over the method of using just one. Based on the 28 genes previous literature has found, the whole TCGA population can be grouped into 4 groups using Kmeans and then finally assembled as two groups by selecting the most differentiating combination (Figure 7.3A) for comparison purposes. The same method is used to classify breast cancer patient using the existing prognostic miRNA biomarkers (Figure 7.3B). Although the gene and miRNA signatures selected in the clustering analysis are tested to be powerful in prognosis of breast cancer patient outcomes, neither of them shows
significant difference when applied to this data, with P-values of 0.129 and 0.071 respectively.

![Image of Figure 7.3: Single type of molecular data clustering based stratification versus Integrative molecular subgrouping based on the 28 prognostic genes and 7 prognostic miRNAs using CCA. (A). Kaplan-Meier curve for the two groups of patients stratified based only mRNA expression. (B). Patient stratification using only miRNA expression. (C). Heatmap for 28-mRNA expression profile after CCA multi-view clustering. (D). Heatmap for 7-miRNA expression profile after CCA multi-view clustering. Each row represents miRNA expression of one patient, which is aligned with mRNA profile in (C). (E). Molecular grouping labels (k=4). The 4 groups were further merged to two groups (labeled as blue and red) based on the best separation of survival curves. (F) Kaplan-Meier plot of the two molecular groups.

Integrative clustering takes both mRNA and miRNA inputs and conducts integrative clustering on both of these omic data. Therefore, unlike the single-linkage clustering method, it can measure multiple levels of the characteristics of the complex tumor.
Heatmaps in Figure 7.3C and Figure 7.3D show the resulting clustered profiles of 28 genes and 7 miRNAs, respectively. Visualized from the heatmaps in Figure 3, each cluster shows unique patterns in both of their mRNA expression and miRNA expression, indicating the efficacy of integrative clustering. The four molecular subpopulations (labeled by Figure 7.3E) were then combined into two groups, whose Kaplan Meier survival curves were shown in Figure 7.3F. Molecular group 1 (label in blue) has worse survival rate than molecular group 2 (labeled in red).

Non-coding small RNA is functioning as important molecular involving in many cellular processes. Our method shows that with the assistance of miRNA expression, molecular stratification of TCGA breast cancer patients is boosted upon mRNA expression. In addition, mRNA expression also boosts miRNAs in prognosis.

### 7.3.3 Integration of Molecular Expression and Clinical Attributes

Using integrative molecular clustering, we can cluster the cancer patients in two several molecular groups as labeled in Figure 7.3F. Upon the molecular stratification and the clinical annotations (tumor grades, disease stages and histology types) described earlier in this section, consensus clustering is applied to obtain a unified patient classification.

Disease Stage (DX stage) is a better indicator for patient stratification regarding to survival than clinical attributes, such as tumor grade and histological type. It is not surprising to find that classification based on disease stage (Figure 7.4A) is more powerful than others (Figure 7.4B and 7.4C). However, with all three clinical labels and
molecular subtype labels integrated in consensus clustering, the patient population is even better (p=0.00195) differentiated into a low-risk group and a high-risk group (Figure 7.4D).

### 7.3.4 Subtype-specific Genes of High-Risk Group

We observed significant expression differences of 93 genes between the subgroup with poor outcomes and the rest. Toppgene enrichment analysis was done for the list of 93 genes. The list of significantly enriched pathways includes regulation of gene expression in beta cells, cell cycle, DNA damage and Metabolism of proteins, etc. IPA gene network analysis was conducted on these 93 genes and the one with the top score was shown in Figure 4E. These genes include CDH1 and RUNX1, which have been tested for their role in breast cancer. We observed that somatic mutation incidence of RUNX1 in the poor outcome group (18%) is three times larger than that of the rest (5.4%). The somatic mutation incidence of CDH1 in the poor outcome group (29.4%) is significantly larger than that of the rest (18.1%). Our observation confirmed the mutation of CDH1 and RUNX1 to be important in this subgroup of breast cancer, indicating that they could be potential targets in this subtype.
Figure 7.4: Kaplan–Meier plots for classification of the whole TCGA population. (A) Stratification based on patient disease stage. (B) Tumor grade. (C) Histological type. (D) Consensus clustering. (E) A network identified from the high-risk group.
7.4 Summary

We propose a new integrative patient classification method which aggregates both essential clinical information and multiple omic data. A breast cancer case study shows that miRNA expressions and mRNA expression can better utilize the associations between mRNAs and miRNAs, thus improves the classification on molecular level. The result of this case study also demonstrates the effectiveness of this consensus clustering patient stratification in breast cancer patient prognosis. In addition, the high-risk group of patients distinguishes from the rest in expression of 93 specific genes and mutation of two well-studied genes. The platform can be extended to any other types of cancer with any selection of biomarkers and clinical signatures. The extended data integration can provide hypothesis to identify both clinical and biological interesting cancer subgroups.
Chapter 8 : Molecular Regularized Consensus Clustering Method

In this chapter, I present a mathematical formulation for integrative clustering of multiple-source data including both numerical and categorical data to resolve the above issue. Specifically, I formulate the problem as a novel consensus clustering method called Molecular Regularized Consensus Patient Stratification (MRCPS) based on an optimization process with regularization. Unlike the traditional consensus clustering methods, MRCPS can automatically and spontaneously cluster both numerical and categorical data with any option of similarity metrics. We apply this new method by applying it on the TCGA breast cancer datasets and evaluate using both statistical criteria and clinical relevance on predicting prognosis. The result demonstrates the superiority of this method in terms of effectiveness of aggregation and differentiating patient outcomes. Our method, while motivated by the breast cancer research, is nevertheless universal for integrative genomics studies.

Cancers are highly heterogeneous with different subtypes that lead to different clinical outcomes including prognosis, response to treatment and chances of recurrence and metastasis. An important task in personalized medicine is to determine the subtype for a breast cancer patient in order to provide the most effective treatment. In order to achieve this goal, integrative genomics approach has been developed recently with multiple modalities of large datasets ranging from genotypes to multiple levels of phenotypes. A
major challenge in integrative genomics is how to effectively integrate multiple modalities of data to stratify the breast cancer patients. Consensus clustering algorithms have often been adopted for this purpose. However, existing consensus clustering algorithms are not suitable for the situation of integrating clustering results obtained from a mixture of numerical data and categorical data.

8.1 Introduction

Breast cancers are highly heterogeneous with many different subtypes. These subtypes confer different outcomes including different prognosis, response to treatments, and chances of recurrence and metastasis. In addition, these subtypes are often associated with different genetic mutations, gene expression profiles, molecular signatures, tissue and organ morphologies as well as different clinical phenotypes. In order to effectively treat the patients, personalized characterization of the genetic, molecular and clinical biomarkers is essential. In order to achieve this goal, integrative genomics approaches are often adopted.

Currently one of the major integrative genomics effort for cancer research is The Cancer Genome Atlas (TCGA) project in which the genotype (SNPs, copy number variances and somatic mutations), epigenome (DNA methylation), transcriptome (mRNA and microRNA levels), proteome, morphology (histological and radiological images), and clinical records for hundreds of patients are made available for each selected cancer including breast cancer. One of the goals is to develop integrative biomarkers which can
effectively stratify the patients into subtypes with clearly different clinical outcomes. However, a major challenge is how to effectively integrate multiple modalities of data.

Since the beginning of the century, gene signatures such as PAM50 [1] and the well-known 70-gene signature [187] based on gene transcriptome data obtained from high throughput experiments such as microarray have been identified for subtyping breast cancer patients and prognosis prediction. However, different signatures often lead to different assignments of subtypes for the same cohort of breast cancer patients. Furthermore, subtyping based on other modalities such as microRNA, gene mutations and morphological features lead to even more discrepancies in the patient stratification [74], [76], [108], [188]. Most importantly, the molecular based subtypes are often not consistent with the clinically used staging diagnosis provided by pathologists. Therefore, there is an urgent need for an effective approach for deriving a consistent framework for stratifying patients based on multiple data modalities.

**Related Work**

Recently, a class of unsupervised machine learning methods called consensus learning has been frequently adopted in addressing the issue of patient stratification discrepancies from multiple datasets. In traditional consensus clustering, the patients are first clustered using individual data modality. These clusters are then aggregated into a set of “consensus” clusters based on certain optimization criteria. Generally speaking, there are two main approaches to achieve the consensus solution and evaluate its quality: (1) probabilistic approach, in which given the distributions of base labelings, a maximum
likelihood formulation is solved to return the consensus; (2) similarity approach, in which one can directly find the consensus clustering that agrees the most with the input base clusterings. For example, Topchy et al. [83] consider a representation of multiple clusterings as a set of new attributes characterizing the data distributions, then a mixture model (MM) offers a probabilistic model of consensus using a finite mixture of multinomial distributions in the space of base clusterings. A consensus result is found as the solution to the corresponding maximum likelihood problem using expectation maximization (EM) algorithm. Another probabilistic approach is Bayesian Consensus Ensemble, BCE [182]. Further, Lock and Dunson [184] adopt the first kind of approach, by extending the Dirichlet mixture model to accommodate data from multiple sources and apply it to multiple genomic data. In the second category, Strehl and Ghosh [20] seek a consensus clustering by maximizing the mutual information. However, currently the consensus clustering algorithms have often been applied to the studies when multiple molecular data types (e.g., genetic mutations with gene expression) are needed but they cannot be directly applicable for integrating the numerical molecular data such as gene expression levels with categorical data such as clinical staging information even though the latter is highly important due to their wide clinical applications.

In this work, I present a mathematical formulation and objective of integrative clustering of multiple-source data including both numerical and categorical data to resolve the above issue. Specifically, I formulate the problem as a novel consensus clustering method called Molecular Regularized Consensus Patient Stratification.
(MRCPS) based on an optimization process with regularization. Unlike the traditional consensus clustering method (Cluster-based Similarity Partitioning Algorithm, CSPA [20]; HyperGraph Partitioning Algorithm, HGPA[20] and Bayesian Consensus Ensemble, BCE [182], which can either take “hard” or “soft” base clusterings, this proposed method MRCPS can automatically and spontaneously cluster both numerical and categorical data with any option of similarity metrics. We apply this new method by applying it on the TCGA breast cancer datasets and evaluate using both statistical criteria and clinical relevance on predicting prognosis. The result demonstrates the superiority of this method in terms of effectiveness of aggregation and differentiating patient outcomes. Our method, while motivated by the breast cancer research, is nevertheless universal for integrative genomics studies.

This chapter is organized as following: in the next section (Section 8.2), the problem of integratively clustering multiple types of data is mathematically formulated, followed by a brief overview of the proposed method. Section 8.3 also describes the details of the proposed MRCPS algorithms with proof of convergence in the Appendix C. Then we apply this approach to a breast cancer subtype study and provide evaluation results in Section 8.3 and discuss its implications in Section 8.4.

8.2 Methods and Materials

8.2.1 Problem Statement

Notations: Consider multiple numerical genomic measurements \( \{X^{(m)}\}_{m=1}^{M} \) (each with dimension of \( p_m \)) collected from \( N \) cancer samples \( \{d_j\}_{j=1}^{N} \), such that \( X^{(m)} \) is a
$p_m \times N$ matrix from $M$ data types (gene expression (mRNA) and microRNA expression), with data types ($m = 1, 2, ..., M$) from a mixture of $k$ distributions ($D_1^m, ..., D_k^m$). Additionally we have at our disposal, various clinical attributes (histological types and tumor grades, etc.) represented by categorical vectors $\{ y^{(l)} \}_{l=1}^L$, all drawn from the same set of patient samples.

**Objective:** Our goal is to find a consensus partition or stratification of patients $C^*$ of $\{d_j\}_{j=1}^N$ from these $M$ sets of genomic measurements and $L$ sets of clinical attributes, such that this integration will reveal clinically and biologically relevant partition of the patient cohort based on the clustering. We propose a robust consensus clustering approach to achieve this aim.

**Rationale:** A natural means to cluster patient samples based on multiple types of data is to first cluster according to each individual data type and then find the consensus partition of the population from the multiple clustering results. This two-step approach carries out a series of clustering (labeled by $\{y^{(m)}\}_{m=1}^M$) from each type of molecular data, as well as each individual clinical attributes (labeled by categorical vectors $\{y^{(l)}\}_{l=1}^L$). With the clustering results from all datasets becoming available, consensus clustering methods such as CSPA [20] and BCE [182] can be used to integrate all the clusterings and provide the final consensus partition of the patients.

Current consensus clustering methods aggregate either the different clustering labels or the probability of each sample belonging to every cluster. For example, graph based methods like CSPA [20] signify a relationship between samples in the same cluster and
thus use the similarity matrix to partition the samples. The input of these methods must be “hard” partitions of the samples and the methods can only take categorical inputs. The same drawback exists in the probabilistic algorithms, such as BCE. Given the distributions of base labelings, they solve a maximum likelihood formulation to return the consensus. The probabilistic approaches also require categorical inputs. Although there are “soft versions” of consensus clustering, they seek consensus partition based on the probability of each sample belonging to every cluster. None of the consensus clustering methods can access to the original features of the samples. On the other hand, integrative clustering approaches such as iCluster \[74], \[189] can only take original features as input but not categorical labels such as clinical attributes.

**Taking advantage of the molecular data**

The above two-step approach clusters each data source separately, followed by a post hoc integration of these separate clusterings. However, the molecular measurements, designed for detecting subtle differences between samples, are converted to a binary pairwise similarity (i.e., if two samples are in the same cluster or not). This two-step method will reduce the accuracy of similarity measured by numerical molecular data and thus can miss these subtle differences in the final consensus.

Thus, instead of indirectly integrating multiple molecular data and clinical attributes, we propose to combine molecular data and clinical data automatically. Using numerical molecular expression data, we can define molecular subtypes and estimate density-based models as prescribed by the affinity of sample in the molecular features space. Next, we
develop a computational method to regularize the clinical classification using the molecular density function. Finally, the patient stratification can be evaluated from statistical, clinical and biological perspectives.

Figure 8.1: Integration of molecular expression data with clinically-defined patient stratification. Although sample $d_i$ and sample $d_j$ come from different clinical sub-types (I and II respectively), they come from the same, stable and dense molecular cluster, so they are desired to be combined in the consensus clustering.

A schematic diagram of this regularization method is shown in Figure 8.1. In the scenario where sample $d_i$ and sample $d_j$ are in the same molecular subgroup and this subgroup is manifested as a relatively dense cluster in the molecular feature space, they
might still be clustered in the same subgroup in the final clustering. In essence, the similarity of these two samples is not altered.

This regularization also works the opposite way when sample $d_i$ and sample $d_j$ are in the sparser molecular subgroup (such as the Molecular Subgroup 3 in Figure. 8.1), but they may belong to the same clinical cluster. In reality, they might have very different etiology for cancer. The regularization will lead to the separation of the two samples in the final consensus clustering. In this way, we can take advantage of molecular features, and at the same time integrate categorical clinical attributes to derive the final clustering.

### 8.2.2 Dataset

In this case study, the same patient samples are used as Chapter 7. In this repository, mRNA and miRNA ($M = 2$) (microRNA) expression profiles were collected from 441 ($N = 441$) primary tumors of breast cancer patients. Pertinent clinical data was also available for all of these patients. The median follow-up duration is 3 years long. The mRNA and miRNA expressions were converted from RNAseq RPKM (reads per kilobase per million) and miRNAseq FPMM (fragments per million miRNA) profiles downloaded from TCGA data portal. Gene and miRNA expressions were later log-transformed, standardized and compiled into matrices. The clinical attributes (e.g., tumor grade and histological type, $L = 3$) were discretized into a discrete number of categories. The demographic information of patient samples is listed in Table 5 in Appendix A.

### 8.2.3 Molecular Regularized Consensus Patient Stratification
With notations and objective stated in the previous section, we now discuss the details of our algorithm for patient stratification based on multiple types of data. To reiterate, our method is inspired by the current state-of-the-art in consensus clustering, whereby different clusterings are aggregated to obtain one robust clustering. However, we will use numerical molecular data distance between samples to tune the clustering defined by the clinical attributes. Therefore, it is necessary to define a distance metric to represent the molecular similarity between samples.

**Definition 1. Cluster density function** [190]: Based on the molecular features, a clustering algorithm such as Kmeans can be applied thus each sample \( x_i \) \((i = 1, 2, \ldots, N)\) is clustered in its molecular subgroup. Then, we can define a cluster density function \( f(i) \) of this sample. A classic choice of the density function is the Gaussian Kernel density function [191]

\[
 f(i) = \frac{1}{h_p N_i} \sum_{j=1}^{N_i} K_h (x_i - x_j) = \frac{1}{N_i (2\pi h^2)^{p/2}} \sum_{j=1}^{N_i} e^{-\frac{\|x_i - x_j\|^2}{2h^2}}, \tag{8.1}
\]

where \( K_h \) is a Gaussian Kernel function with parameter \( h \) and \( N_i \) is the number of samples in the same cluster with \( x_i \) in dimension \( p \). Users can also define their own the density function according to the distribution of data and the hypothesis.

**Definition 2. Categorical distance metric**: The categorical distance metric between two clinical clusters \( C^{(l)} \) and \( C^{(l)} \) can be defined as \( (C^{(l)}, C^{(l)}) = \sum_{i<j} s_{ij}^{(l)} - \).
\[ s_{ij}^{(l)}, \text{ where } s_{ij}^{(l)} = 1 \text{ if the data points } d_i \text{ and } d_j \text{ belong in the same cluster in } C^{(l)}, \text{ and } 0 \text{ otherwise.} \]

Now, let us consider how to take advantage of the density function provided by the molecular data to tune the clinical clusterings. In fact, we can weight patient samples based on the reliability of the molecular communities to which they belong. The underlying intuition is that, if two patient samples \(d_i\) and \(d_j\) are in a cluster of poor reliability in terms of molecular clustering, similarity between them can be deemed to be low and given a lower weight to leverage the high clinical similarity \(s_{ij}\). The reliability is measured by the density functions \(f(i)\) and \(f(j)\) of the samples \(d_i\) and \(d_j\). Specifically, given a set of \(L\) set of clinical partitions, \(C^{(1)}, ..., C^{(l)}, \) and \(dist(.\) the symmetric difference distance metric, we wish to find an overall partition \(C^*\) such that

\[
C^* = \arg\min_C \sum_{l=1}^{L} \text{dist}(C^{(l)}, C). \tag{8.2}
\]

Equipped with the cluster density function (Definition 1) for molecular affinity and with a distance measure to compare across different clinical clusters (Definition 2), we can rewrite the above optimization in an equivalent form:

\[
S^* = \frac{1}{L} \arg\min_{S^*} \sum_{l=1}^{L} \omega_l \sum_{i<j} \left[ s_{ij}^{(l)} - s_{ij}^{*}\right]^2, \tag{8.3}
\]

where \(S^{(l)}\) is a \(N\) by \(N\) coassociation matrix. And \(s_{ij}^{(l)}\) is the entry of \(S^{(l)}\); \(s_{ij}^{*}\) is the entry of \(S^*\). \(\omega_l\) is a weight for considering the contributions given by different base clusterings.
In this case, if we assume every clinical data type scores the same to the final stratification, then \( \omega_l = 1; \forall 1 \leq l \leq L \), changing Eq. (8.3) as follows:

\[
S^* = \frac{1}{L} \arg\min_{S^*} \sum_{l=1}^{L} \sum_{i<j} \left[ S_{ij}^{(l)} - s_{ij}^* \right]^2.
\]  

(8.4)

For samples \( d_i \) and \( d_j \), we have defined density estimators \( f(i) \) and \( f(j) \) respectively as the density of clusters they belong to. This density function denotes the ‘molecular affinity’ of the cluster sample that contains \( d_i \). The larger \( f(i) \) is, the more similar the subgroup is, in molecular level. The same intuition applies to sample \( d_j \). We can weight each pair of similarity \( s_{ij} \) with the \( \omega_{ij} = f(i) \times f(j) \) and for every \( d_i \) and \( d_j \). Hence, Eq. (8.4) becomes:

\[
S^* = \frac{1}{L} \arg\min_{S^*} \sum_{l=1}^{L} \sum_{i<j} \omega_{ij} \left[ s_{ij}^{(l)} - s_{ij}^* \right]^2.
\]  

(8.5)

where \( \omega_{ij} \) is the molecular density function defined as:

\[
\omega_{ij} = \begin{cases} 
  f(i) \times f(j) & \text{if } i \text{ and } j \text{ are in the same molecular cluster} \\
  0 & \text{if } i \text{ and } j \text{ are not in the same molecular cluster} \\
  1 & \text{if } i = j
\end{cases}
\]

We solve the above optimization problem as following:

\[
S^* = \frac{1}{L} \arg\min_{S^*} \| \tilde{S} - \tilde{S}^* \|_F^2,
\]  

(8.6)

where \( \tilde{S} = \frac{1}{L} \sum_{l=1}^{L} (S^{(l)} \circ W) \) and \( \tilde{S}^* = S^* \circ W \) are the Hadamard products with ‘molecular affinity matrix’ \( W \). \( \| \cdot \|_F \) denotes the matrix Frobenius Norm.
Since \( S^* \) represents the membership matrix of the final consensus clustering, we can write \( S^* = U U^T \), where \( U = \{0,1\}^{N \times k} \); \( k \) is the number of clusters in the consensus. The binary membership indicator matrix \( U \) should satisfy that each row in \( U \) can only have one ‘1’, \( \sum_{j=1}^{k} U_{ij} = 1, \forall i \in \{1,\ldots,N\} \). Then, the above optimization problem becomes:

\[
U^* = \arg\min_U \| \tilde{S} - W \circ U U^T \|_F \quad \text{s.t. } U = \{0,1\}^{N \times k}.
\]  

(8.7)

It is also required that \( U \) is diagonalizable, or

\[
U^T U = D = \text{diag}(U^T) = \text{diag}(N_1,N_2,\ldots,N_k), \text{ where } D \text{ is a diagonal matrix},
\]

\( \{N_1,N_2,\ldots,N_k\} \) are the numbers of samples in the clusters respectively, and further

\( N_1 + N_2 + \cdots + N_k = N \). The optimization of Eq. (8.7) also requires searching through the indicator domain, which is exhaustive. We can relax the optimization as:

\[
\bar{U}^* = \arg\min_{\bar{U}} \| \tilde{S} - W \circ \bar{U} \bar{D} \bar{U}^T \|_F^2 \quad \text{s.t. } \bar{U}^T \bar{U} = 1_{k \times k} ; D_{ij} > 0,
\]  

(8.8)

where \( \bar{U} = U(U^T U)^{-\frac{1}{2}} \). Given clinical similarity matrix \( S = \frac{1}{L} \sum_{l=1}^{L} S^{(l)} \) and molecular density matrix \( W \), similar to the classical Nonnegative Matrix Factorization (NMF) algorithm [192], this weighted Frobenius norm is non-increasing under the following updating rules:

\[
\bar{U}_{ij} \leftarrow \bar{U}_{ij} \sqrt{\frac{[(W \circ S) \bar{D} \bar{D}^T]_{ij}}{[(W \circ U \bar{D} U^T) \bar{D} \bar{D}^T]_{ij}}},
\]  

(8.9)

\[
D_{ij} \leftarrow D_{ij} \sqrt{\frac{[(\bar{U}^T S \circ W) \bar{U}]_{ij}}{[(\bar{U}^T \bar{U} \bar{D} \bar{D}^T \circ W) \bar{U}]_{ij}}},
\]  

(8.10)
The proof the convergence and derivations are provided in Appendix C and Appendix B. The detail of the MRCPS algorithm is given in Algorithm 1.

Algorithm 1. Molecular Regularized Consensus Patient Stratification

**Data:** Similarity Matrix $\tilde{S}$, Molecular Density Weight Matrix $W$, the number of clusters in final consensus $k$, MaxIter, precision $\epsilon$

**Result:** Cluster indicator matrix $U$.

Initialize $\tilde{U}^{(1)} > 0$, $t = 1$, $\Delta = +\infty$;

**while** $t <$ MaxIter and $\Delta > \epsilon$ do

Update $\tilde{U}^{(t+1)}_{ij} \leftarrow \tilde{U}^{(t)}_{ij} \sqrt{\frac{[W \circ S \tilde{U}D]_{ij}}{[W \circ \tilde{U}D\tilde{U}^T]_{ij}}}$;

Update $D^{(t+1)}_{ij} \leftarrow D^{(t)}_{ij} \sqrt{\frac{[\tilde{U}^T (S \circ W) \tilde{U}]_{ij}}{[\tilde{U}^T (\tilde{U}D\tilde{U}^T \circ W) \tilde{U}]_{ij}}}$;

Computer $\Delta = \| \tilde{S} - W \circ \tilde{U}D\tilde{U}^T \|_F^2$;

$t = t + 1$;

**end**

Discretize $\tilde{U}$ to binary membership matrix.

Here, the main computational cost is on calculating $\sqrt{\frac{[W \circ S \tilde{U}D]_{ij}}{[W \circ \tilde{U}D\tilde{U}^T]_{ij}}}$ and $\sqrt{\frac{[\tilde{U}^T (S \circ W) \tilde{U}]_{ij}}{[\tilde{U}^T (\tilde{U}D\tilde{U}^T \circ W) \tilde{U}]_{ij}}}$ in Eq. (8.9) and Eq. (8.10), each of which takes $O(kN^2)$ operations. So, the algorithm has complexity of #iteration $\times O(kN^2)$, the same complexity with its general NMF form [193]. In real applications, the algorithm converges within about 1000 iterations.

**8.2.4 Algorithm Usage**

The proposed MRCPS method integratively clusters multiple numerical genomic data and categorical clinical attributes from the same samples and seeks a consensus clustering
of the population. By this consensus clustering patient stratification, prognosis can be optimized and clinically and biologically interesting subtypes are identified.

Input: Numerical genomic measurements \( \{X^{(m)}\}_{m=1}^{M} \) collected from \( N \) cancer samples \( \{d_i\}_{i=1}^{N} \) and categorical vectors \( \{y^{(l)}\}_{l=1}^{L} \) drawn from the same set of patient samples. The number of final consensus clusters \( K \).

Output: The final consensus partition \( C^* \) of the \( N \) samples, where \( C^* \) is a \( N \)-by-1 label vector indicating which cluster each sample belongs to.

8.2.5 Selection of Molecular Features based on Prior Knowledge

High-throughput sequencing measures the activities of several thousand molecules simultaneously. However, the difficulty of partitioning over these data is intrinsically caused by the existence of many redundant features that do not contribute to patient stratification. Thus, we can select a subset of relevant features from previous literatures. In the breast cancer case study, 70 prognostic genes [187] and 7 prognostic miRNAs [181] were selected. Genes from TCGA breast cancer dataset were matched with 70 genes and 28 genes were precisely matched. (Table 7.1).

8.2.6 Choice of \( k \) and Statistical Evaluation

Traditional consensus clustering methods such as CSPA and HGPA [20] assess the effectiveness of cluster ensemble using mutual information function. In essence, the optimal consensus is expected to have the maximum mutual information with the base clusterings, meaning that it shares the most information. Here, we adopt the same statistical evaluation of using mutual information.
Let the entropy associated with \( u \)-th base clustering \( H(C_u) = -\sum_h \frac{N_h^{(u)}}{N} \log \frac{N_h^{(u)}}{N} \), where \( N \) is the number of samples and \( N_h^{(u)} \) is the number of samples with label \( h \) in cluster \( C_u \).

Similarly, the entropy arising from the final clustering label is \( H(C_f) = -\sum_i \frac{N_i^{(f)}}{N} \log \frac{N_i^{(f)}}{N} \), where \( N_q \) denotes the number of samples with label \( q \) in final cluster. Therefore, the final clustering number \( k \) can be found by maximizing the following Normalized Mutual Information (NMI) with the original clusterings \( C \):

\[
\phi^{(NMI)}(C_f, C) = \frac{\sum H(C_u) + H(C_f) - H(C_u, C_f)}{\sqrt{H(C_u)H(C_f)}},
\]

(8.11)

where \( H(C_u, C_f) \) is the mutual information between two clusterings \( C_u \) and \( C_f \).

Another important usage of the Normalized Mutual Information function is to determine the parameter \( k \), the number of final clusters in consensus clustering. With different choices of \( k \), we select the one which can maximize \( \phi^{(NMI)} \). In other words, the parameter \( k \) can be found, so that the similarity between the final cluster and all based clusterings is maximized.

8.3 Results

In this section, we compared the proposed MRCPS method with other consensus clustering methods that also follow the two-step model. A selected cohort of breast cancer patients in TCGA provides the necessary tested. The identified breast cancer patient subgroups were computationally tested for robustness, and then evaluated in clinical and biological contexts.
8.3.1 Comparative Performance of MRCPS

It is our hypothesis that MRCPS will maximize the normalized mutual information (NMI) measure in Eq. (8.11). We included the following consensus clustering algorithms in our study namely, CSPA, HGPA and BCE for comparison purposes. For these three traditional methods, mRNA and microRNA expression profiles were clustered separately using the K-means algorithm after completing the step of feature selection. Then, we partitioned the patient population into $k = 2, \ldots, 7$ clusters. The result shown in Figure 8.2 indicated that the BCE method performed worst in terms of combining clusterings.

![Figure 8.2: Plotted values of NMI of MRCPS and other methods for different values of k.](image-url)
Although $k = 7$ slightly outperformed smaller choices of $k$ for CSPA, HGPA and the proposed MRCPS methods. It is likely that over-fitting exists with $k$ being larger, since the observed numbers of clusters in clinical partitions are typically smaller than six. For example, the TCGA breast cancer data has four tumor grades and four disease stages. We did not provide NMI measurements beyond $k = 7$, since in this case study there were only 7 microRNA features. Larger $k$ (beyond 7) is likely to provide more subtle differences between patient groups, but also will generate more over-fitting. In the case study, we have also observed that with increasing $k$ beyond 7, the prognostic power of MRCPS decreases. In other application, users can determine the size of clusters by the sample size and separate the over-fitting from the actual improvements in clustering quality by examining not only the goodness of integrating, but also the prognostic power, which will not increase monotonically with $k$.

For the smaller number of clusters, MRCPS performs the best among the four methods. Additionally, the experiment also suggests that the choice of $k = 4$ is the best given the higher value of normalized mutual information (Eq. (8.11)).

**8.3.2 Clinical Evaluation based on Prognosis Prediction**

To determine the prognostic capabilities of MRCPS, we considered selected patients among the TCGA breast cancer datasets. For comparison, we provide $k = 4$ clusters to the each of the realizations of the CSPA, HGPA and BCE algorithm to obtain the consensus clusters eventually. These results, together with the induced clinical partitions obtained from using disease stage, tumor grade and histological type are shown in Figure
8.3. It is observed that infiltrating ductal carcinoma (IDC, n = 365), which starts in an epithelial duct of the breast, has relatively higher risk than infiltrating lobular type (n = 37) but is not distinct from mixed and other histological types in terms of overall survival (Figure 8.3(a)). Tumor size and disease stage are better prognostic markers than histological type (Figure 8.3(a)) and Figure 8.3(c)).
Figure 8.3: Prognostic power of different patient stratification methods. Kaplan–Meier survival curves of (a) Histology Type, (b) Tumor Grade, (c) Disease Stage, (d) BCE, (e) HGPA and (f) CSPA listed along with estimated p-values (log-rank test). The numbers of the patients in each stratification are also listed in the parentheses.
On the other hand, the use of BCE consensus clustering method will identify subtypes that do not imply significantly distinct patient survival trends (log-rank p = 0.789; Figure 8.3(d)). Graph-based methods, HGPA and CSPA, tend to cluster patient samples into subgroups with equal size Figure 8.3(d)) and Figure 8.3(f) respectively. Moreover, they exhibit better performance (log-rank p = 0.0419; Figure 8.3(f) and p = 0.0417; Figure 8.3(f), respectively) than the Bayesian method, BCE, in predicting patients overall survival. This outcome can arise from the unbalanced number of clusters in the initial clustering.

We further compared these results with the subtypes derived from the proposed MRCPS method for the same patient samples. It can be observed (Figure 8.4) that the identified subtypes are highly predictive with log-rank p-value of 1.53e-04, admittedly lower than any other method (Figure 8.3). Although the sizes of subtypes are not as balanced as the ones obtained from graph-based clustering methods, the proposed MRCPS method has superior ability to differentiate patient populations with distinct outcomes. Specifically, MRCPS identifies a subtype that has significantly poorer survival rate (red curve in Figure 8.4 (a)).

Furthermore, the identified subtypes display a good balance between early stage tumors and late stage tumors (Figure 8.4(b)). The association of tumor grade does not differ significantly across the subtypes (Figure 8.4 (c)). This indicates that MRCPS clustering does not dominantly depend on tumor grade or disease stage. The partition of the patient population is also regularized by molecular affinity and other clinical attribute.
Figure 8.4: Prognostic power of MRCPS. (a) Kaplan–Meier survival curves of MRCPS with its p-value (log-rank test). (b) Disease stage in each subtype; stages earlier than stage III are considered to be early, with the rest considered to be late. (c) Tumor grades for each subtype. Grades less than T3 are considered to be lower grades, with the rest considered to be higher.

8.3.3 Biological Evaluation

We identified genes that are differentially expressed in each subtype relatively to other subtypes derived from MRCPS. To achieve this, we carried out a supervised analysis using ANOVA (false-discovery rate of 0.05) followed by Tukey post hoc testing [194] to identify genes with differential expression between pre-defined groups. We focus on the group with the worst outcome (group 2; red curve in Fig. 4(a)). Ninety (90) genes were uncovered for this subgroup with poor outcome (Table 4 in Appendix A).

Among these genes, many have been reported to be significantly mutated, such as GATA3. Over-expression of CDH3 was also reported in esophageal, pancreatic, bladder and breast cancers [195]–[199].
Next, we conducted network and functional analysis using the Ingenuity Pathway Analysis software (IPA, fall release 2013, http://www.ingenuity.com) on the collected gene list. A network which contains 23 genes was identified (Fig. 5). Interestingly, GATA3 serves as a hub gene in this network and correlates strongly with estrogen receptor. Recently, it was discovered that GATA3 serves as a licensing factor for ESR1-mediated transcription and it might potentially explain a mutant-GATA3 subtype of ESR + breast cancer [200] and GATA3 is one of the most frequently mutated gene in breast cancers. Our observation suggests that ESR1 + patients (n = 37) in this poor-outcome subgroup might be this high-risk mutant-GATA3 subtype.
Figure 8.5: A gene network identified from the high-risk group.
We also analyzed the specific genes for the poor outcome group using IPA. Top diseases are Cancer and Reproductive System Disease, which are directly related to breast cancers (Fig. 6). In addition, the top enriched functions include gene expression, cellular growth/proliferation, and cell cycle, which are related to tumor development.

8.4 Summary

In this chapter, we propose a novel consensus clustering method enabling integrative patient stratification, MRCPS, which aggregates both essential clinical information and multiple molecular expression data. Our method solves the integrative clustering problem by regulating clinical partition of patients using the molecular affinity function, and then achieves the consensus clustering by a multiplicative update. We applied MRCPS to
stratify the large breast cancer patient cohort datasets collected from The Cancer Genome Atlas (TCGA). The result shows the proposed MRCPS method can robustly combine data from different sources into reliable and clinically relevant subtypes. Specifically, a subgroup of patients with extremely poor survival was identified based on the integrative analysis.

The application of MRCPS is not limited to breast cancer. Our case study combined mRNA and microRNA expression into viable molecular features. It can also be used to identify subtypes of other types of diseases. Nonetheless, MRCPS can also consider other kinds of genomic features which are genomic in nature. Our method can be also extended to applications in which integration of both categorical and numerical datasets for clustering is needed.

One of the limitations of our method is that the number of clusters k has to be set up by the users. Another future work will be the definition of genetic difference between samples, which will allow the integration with genetic variances in subtyping cancer patients.
Chapter 9 : Conclusions and Future Works

Histopathological images play an essential role in both basic research and clinical practices in cancers. New technologies in imaging help us localize cellular components in tissue. This localized information is critical in understanding organization and behaviors of cells and other components, as well as the underlying molecular events that regulate and support them in tumors. Integration of histopathological images with other data modalities such as genetic and genomic data is pivotal in biomedical research and clinical practice. There is a need for new digital pathology tools for linking the spatial information of cellular structures to genetic and phenotypic aberrations at molecular resolution. Nonetheless, rapidly accumulation of various types of cancer profiling data including histopathology images poses great challenge on mining the data to identify integrative biomarkers and subtypes.

This dissertation presents studies for three fundamental aspects of integrating multi-modality data in cancer. A data analytic framework is established, by contributing to answers to the following questions:

Q1: How are pathological and genomics features related in cancers?

Q2: What are the genomic markers associated with an individual morphological feature?
Q3: How can the multi-modality data help to identify new subtypes and biomarkers in cancers?

To address the first question, the correlation between cellular morphological features from histopathological images and genomic profiles of the same cancer tissue reveals several potential new biomarkers for triple negative breast cancer and lung adenocarcinoma. Then, beyond simple correlations, a multivariate model is established to explain individual morphological using multiple gene expression networks. Finally, novel algorithms are developed providing researchers and clinicians with computational tools to discover new subtypes of cancer based on multi-modality data.

Chapter 3 describes the computational histopathological image processing pipeline in this study. The pipeline consists of multiple stages from image acquisition to image preprocessing, from nuclei segmentation to tissue segmentation, from tissue classification to morphological feature extraction. This comprehensive workflow adopts the recent advance of methods in each stage and generates large amount of features that can capture the cellular information about the tissue.

Chapter 4 utilizes the morphological features and genomics features to measure the difference in triple negative breast cancer patients. The study suggests that salient imaging features from histopathology are potential biomarkers for tnbc. Correlations between the imaging features and gene expression are established using the proposed framework.
Chapter 5 generalizes the morphological features to identify proteomics biomarkers for breast cancer. Based on the morphological features, patient samples can be stratified into groups with distinct tumor aggressiveness. Then, the proteins that differentially express between the groups are uncovered. This supervised workflow generates a hypothesis that one of the protein/gene is an essential regulator for cancer cell density. Finally, we show the significance of these identified proteins in patient outcomes.

Chapter 6 introduces another model for exploring correlations between imaging features and genomics measurements. This model accounts for multiple molecules in regulation of cellular phenotypes and discovers that cilium related molecular events are important in lung adenocarcinoma. This study can also be extended to other types of cancers.

Chapter 7 and Chapter 8 present two algorithms for the goal of clustering patient samples based on multiple types of measurements. The problem of subtype identification is formulated as integrative clustering. Clinical attributes, which are essential information in clinical settings, can be combined with molecular data in the formulation to obtain integrative cancer patient subtypes of interest. The case studies show the performance of the algorithms by evaluating the patient subgroups clinical outcomes and biological processes.

The proposed framework also leaves considerable space for future research. I list some of the drawbacks and future directions for improvement:
1. **Whole slide image versus selected image patches.** In our image processing steps, patches of the whole slide images were curated based on visual observation of representation of the tissue. This step reduced the amount of tumor in the final representative features. However, the genomic and epigenomic data were extracted from the whole 3D stack, which means that the extracted cellular morphology is from a down-sampled tissue. Using the whole slide images will increase the proportion of sampled tissue in morphological measurement, though they are still 2D images. To overcome this problem, the advanced microscopic imaging technology can potentially generate 3D morphological measurements from the whole solid tissue. In that circumstance, the integration techniques presented in this dissertation can also be adopted. In addition, the pathological report data that include essential information about the tumor can be further integrated into the workflow.

2. **Full-fledged characterization of tumor microenvironment cell types.** In our cell segmentation stage, we classify two types of cells: epithelial cells and fibroblasts. However, there are also a number of other cells, such as macrophages and lymphocytes, which are also essential components of the tumor microenvironment. Since classification of all cell types itself would have been a very difficult task, our analysis was limited to the two major cell types. However, without systematic exploration of techniques and feature design, the automatic determination of all cell types using computational methods will generate
potentially large error. The high computational burden is also one of the considerations. In future work, it is conceivable that the differentiation of other cell types would improve the downstream modeling and patient stratification.

3. **More morphological features.** In morphological study, we extracted 74 epithelial features and 44 stromal features. Given the complexity of the tumor, more features may be needed to boost the classification of the tissues and to represent the cellular structures.

4. **Validation.** In exploring associations between morphological phenotypes and genomic phenotypes, we did not carry out extensive external validation due to the lack of cancer patient datasets that possess both omics data and tissue microscopic images. Most of the public available datasets, such as GEO, do not have corresponding tissue slides. Nevertheless, the lack of statistical measures of confidence can be addressed by validation datasets.

5. **Supervised integrative classification.** Our clustering algorithms are unsupervised methods. For biological systems, certain prior knowledge must be discovered and valuable in models. This requires the algorithms that can integrate supervised mechanisms. In our formulation, no supervised information or prior knowledge is involved. One of the future directions can be Bayesian integrative clustering with allowance of multi-modality of biomedical data.

This dissertation is one of the first efforts in computational biology to combine morphology in histopathology with genomic data for the goal of identifying subtypes.
The listed future directions can enhance the models and studies, leading to more accurate predictions.
References


181


### Appendix A: Tables

Table A.1: Complete Epithelial Feature List.

<table>
<thead>
<tr>
<th>Epithelial Feature Names</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previously Discovered Features</td>
<td></td>
</tr>
<tr>
<td>GLCM.Ang..2nd.moment..quick.8.11..Layer.1..all.dir..</td>
<td>GLCM, similar to haralick</td>
</tr>
<tr>
<td>Ratio.Layer.3</td>
<td>The amount that a given image layer contributes to the total brightness.</td>
</tr>
<tr>
<td>GLCM.Correlation..quick.8.11..Layer.1..all.dir..</td>
<td>GLCM, similar to haralick</td>
</tr>
<tr>
<td>Asymmetry.of.sub.objects..mean..1.</td>
<td></td>
</tr>
<tr>
<td>Rel..Border.to.Image.Border</td>
<td></td>
</tr>
<tr>
<td>GLCM.Entropy..quick.8.11..Layer.1..all.dir..</td>
<td></td>
</tr>
<tr>
<td>Ratio.to.scene.Layer.1</td>
<td>Ratio To Scene Red Channel</td>
</tr>
<tr>
<td>Skewness.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Rel..area.of.sub.objects..unclassified..1.</td>
<td>Area covered by sub-objects assigned to a given class divided by the total area of the image object concerned.</td>
</tr>
<tr>
<td>Rel..border.to.brighter.objects.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Density.of.sub.objects..stddev..1.</td>
<td>The density is calculated by the number of pixels forming the image object divided by its approximated radius.</td>
</tr>
<tr>
<td>Rel..area.of.sub.objects.Dark..1.</td>
<td></td>
</tr>
<tr>
<td>Skewness.Layer.3</td>
<td></td>
</tr>
<tr>
<td>Compactness..polygon.</td>
<td>The ratio of the area of a polygon to the area of a circle with the same perimeter.</td>
</tr>
<tr>
<td>Mean.Layer.3</td>
<td></td>
</tr>
<tr>
<td>StdDev..to.neighbor.pixels.Layer.2..3.</td>
<td></td>
</tr>
</tbody>
</table>

Continued
Table A.1 continued: Complete Epithelial Feature List.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>StdDev..to.neighbor.pixels.Layer.3..3.</td>
<td></td>
</tr>
<tr>
<td>Mean.of.outer.border.Layer.3</td>
<td></td>
</tr>
<tr>
<td>Direction.of.sub.objects..stddev..1.</td>
<td>The angle weighted by the eccentricity</td>
</tr>
<tr>
<td>Mean.of.sub.objects..stddev.Layer.1..1.</td>
<td></td>
</tr>
<tr>
<td>Min.10Quartile.Pixel.value.Layer.3</td>
<td></td>
</tr>
<tr>
<td>Mean.Diff..to.neighbors..abs..Layer.2..0.</td>
<td>Simplified as mean diff to the complement</td>
</tr>
<tr>
<td>Circular.Mean.Layer.1..R1..User..3..R2..Same..R1..border.</td>
<td></td>
</tr>
<tr>
<td>Area.of.sub.objects.Nucleus..1..Pxl.</td>
<td></td>
</tr>
<tr>
<td>Min.10Quartile.Pixel.value.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Area.of.sub.objects..stddev..1..Pxl.</td>
<td></td>
</tr>
<tr>
<td>Contrast.to.neighbor.pixels.Layer.3..3.</td>
<td></td>
</tr>
<tr>
<td>Area..including.inner.polygons..Pxl.</td>
<td></td>
</tr>
<tr>
<td>Length.µm.</td>
<td></td>
</tr>
<tr>
<td>Width.µm.</td>
<td></td>
</tr>
<tr>
<td>Mean.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Mean.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Standard.deviation.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Standard.deviation.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Standard.deviation.Layer.3</td>
<td></td>
</tr>
<tr>
<td>Skewness.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Ratio.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Ratio.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Min.10Quartile.Pixel.value.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Max.90Quartile.Pixel.value.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Max.90Quartile.Pixel.value.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Max.90Quartile.Pixel.value.Layer.3</td>
<td></td>
</tr>
<tr>
<td>Pathological Features</td>
<td></td>
</tr>
<tr>
<td>Nuclei_Density</td>
<td># of nuclei/size of image object</td>
</tr>
<tr>
<td>Fibroblast_Nuclei_Density</td>
<td># of Fibroblast nuclei/size of image object</td>
</tr>
<tr>
<td>Fraction_Fiber</td>
<td>size of Fiber/size of image</td>
</tr>
</tbody>
</table>

continued
Table A.1 continued: Complete Epithelial Feature List.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction_Nuclei</td>
<td>size of Nuclei/size of image</td>
</tr>
<tr>
<td>Mean_Distance_Fibroblast_Neigbour_Epi</td>
<td>If it is a cancer superpixel, the feature is defined as the distance to its neighbouring fibroblast</td>
</tr>
<tr>
<td>Nuclei_Orientation</td>
<td>Angle of the major axis. The angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.</td>
</tr>
<tr>
<td>Nuclei_Perimeter</td>
<td>Nuclei boundary</td>
</tr>
<tr>
<td>Nuclei_Solidity</td>
<td>Scalar specifying the proportion of the pixels in the convex hull that are also in the region. Computed as Area/ConvexArea.</td>
</tr>
<tr>
<td>Nuclei_Eccentricity</td>
<td>Specifies the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length.</td>
</tr>
<tr>
<td>Nuclei_EquivDiameter</td>
<td>Specifies the diameter of a circle with the same area as the region. Computed as sqrt(4*Area/pi)</td>
</tr>
</tbody>
</table>

continued
Table A.1 continued: Complete Epithelial Feature List.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclei_MajorAxisLength</td>
<td>Scalar specifying the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region.</td>
</tr>
<tr>
<td>Nuclei_MinorAxisLength</td>
<td>the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region.</td>
</tr>
<tr>
<td><strong>Texture Features (Haralick)</strong></td>
<td></td>
</tr>
<tr>
<td>Autocorrelation</td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
</tr>
<tr>
<td>Cluster Prominence</td>
<td></td>
</tr>
<tr>
<td>Cluster Shade</td>
<td></td>
</tr>
<tr>
<td>Dissimilarity</td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td></td>
</tr>
<tr>
<td>Entropy</td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td></td>
</tr>
<tr>
<td>Maximum probability</td>
<td></td>
</tr>
<tr>
<td>Sum of squares</td>
<td></td>
</tr>
<tr>
<td>Sum average</td>
<td></td>
</tr>
<tr>
<td>Sum variance</td>
<td></td>
</tr>
<tr>
<td>Sum entropy</td>
<td></td>
</tr>
<tr>
<td>Difference variance</td>
<td></td>
</tr>
<tr>
<td>Difference entropy</td>
<td></td>
</tr>
<tr>
<td>Information measure of correlation1</td>
<td></td>
</tr>
<tr>
<td>Information measure of correlation2</td>
<td></td>
</tr>
<tr>
<td>Inverse difference normalized (INN)</td>
<td></td>
</tr>
<tr>
<td>Inverse difference moment normalized</td>
<td></td>
</tr>
</tbody>
</table>
Table A.2: Complete Stromal Feature List.

<table>
<thead>
<tr>
<th>Stromal Feature Names</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previously Discovered Features</td>
<td></td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Std.Dev_.Mean.Diff.to.neighbors</td>
<td>Variability in the contrast of stromal matrix superpixels to neighboring objects</td>
</tr>
<tr>
<td>Stroma_Tissue.csv_Max.GLCM.Mean.Layer.1</td>
<td>Presence of stromal objects without nuclei</td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Min.Max.pixel.value.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Stroma_Tissue.csv_Sum.Min.pixel.value.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Stroma_Dark.csv_Mean.Rel.area.of.Stroma_Nucleus..0.</td>
<td>Average relative border of stromal spindle nuclei to stromal round nuclei</td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Min.Standard.deviation.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Stroma_Tissue.csv_Mean.Circular.Mean.Layer.2..R1..User..3..R2._Same..R1..border.</td>
<td></td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Min.GLCM.Ang.2nd.moment</td>
<td></td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Max.Skewness.Layer.1</td>
<td></td>
</tr>
<tr>
<td>Stroma_Matrix.csv_Min.Min.pixel.value.Layer.2</td>
<td></td>
</tr>
<tr>
<td>Texture Feature (Haralick)</td>
<td></td>
</tr>
<tr>
<td>Autocorrelation</td>
<td></td>
</tr>
<tr>
<td>Contrast</td>
<td></td>
</tr>
<tr>
<td>Correlation</td>
<td></td>
</tr>
<tr>
<td>Cluster Prominence</td>
<td></td>
</tr>
<tr>
<td>Cluster Shade</td>
<td></td>
</tr>
<tr>
<td>Dissimilarity</td>
<td></td>
</tr>
<tr>
<td>Energy</td>
<td></td>
</tr>
<tr>
<td>Entropy</td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td></td>
</tr>
<tr>
<td>Maximum probability</td>
<td></td>
</tr>
<tr>
<td>Sum of squares</td>
<td></td>
</tr>
</tbody>
</table>

continued
Table A.2 continued: Complete Stromal Feature List.

| Sum average |  |  |
| Sum variance |  |  |
| Sum entropy |  |  |
| Difference variance |  |  |
| Difference entropy |  |  |
| Information measure of correlation1 |  |  |
| Information measure of correlation2 |  |  |
| Inverse difference normalized (INN) |  |  |
| Inverse difference moment normalized |  |  |

Pathological Features
- Fibroblast_Orientation
- Fibroblast_Perimeter
- Fibroblast_Solidity
- Fibroblast_Eccentricity
- Fibroblast_EquivDiameter
- Fibroblast_MajorAxisLength
- Fibroblast_MinorAxisLength
- Fibroblast_Density
- Fibroblast_Area

Table A.3: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>Genes Features</th>
<th>Area_Sub_Object_s_stddev</th>
<th>StdDev_to_Neighbor_pixels_Layer_2</th>
<th>GLCM_Correlation</th>
<th>Standard_deviation_Layer_1</th>
<th>Density_Cell_Nuclei_stdd</th>
<th>Density_Cell_Nuclei_stdd' metagene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNC45 B</td>
<td>UNC45 B</td>
<td>UNC45 B</td>
<td>TNNT2</td>
<td>TNNT2</td>
<td>KLHL4</td>
<td>TARSL2</td>
</tr>
<tr>
<td>DUSP27</td>
<td>DUSP27</td>
<td>DUSP27</td>
<td>ADRA2 B</td>
<td>ADRA2 B</td>
<td>CACNA2D4</td>
<td>C4orf52</td>
</tr>
<tr>
<td>SMTNL2</td>
<td>SMTNL2</td>
<td>SMTNL2</td>
<td>SLURP1</td>
<td>SLURP1</td>
<td>PRDM5</td>
<td>PPP3CC</td>
</tr>
<tr>
<td>DUSP26</td>
<td>DUSP26</td>
<td>DUSP26</td>
<td>GRAMD1B</td>
<td>GRAMD1B</td>
<td>GABRA2</td>
<td>FHL5</td>
</tr>
<tr>
<td>AMPD1</td>
<td>AMPD1</td>
<td>AMPD1</td>
<td>POPDC3</td>
<td>POPDC3</td>
<td>SMAD9</td>
<td>SOBP</td>
</tr>
<tr>
<td>ASB5</td>
<td>ASB5</td>
<td>ASB5</td>
<td>VMO1</td>
<td>VMO1</td>
<td>CMA1</td>
<td>GANC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>TNFSF1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CXorf36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>KLF10</td>
</tr>
</tbody>
</table>

continued
Table A.3 continued: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>ACTN3</th>
<th>ACTN3</th>
<th>ACTN3</th>
<th>ADAM2</th>
<th>ADAM2</th>
<th>C18orf3</th>
<th>FAM43</th>
<th>CDC14B</th>
<th>DSEL</th>
<th>TMEM47</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOH12C</td>
<td>LOH12C</td>
<td>R1</td>
<td>R1</td>
<td>CHST7</td>
<td>C17orf9</td>
<td>FIGF</td>
<td>FLJ1035</td>
<td>7</td>
<td>ADD3</td>
</tr>
<tr>
<td>HRC</td>
<td>HRC</td>
<td>HRC</td>
<td>COL23A1</td>
<td>COL23A1</td>
<td>SAA4</td>
<td>CORO2B</td>
<td>SELP</td>
<td>CEBPD</td>
<td>SH3D19</td>
</tr>
<tr>
<td>MYH3</td>
<td>MYH3</td>
<td>MYH3</td>
<td>LSAMP</td>
<td>LSAMP</td>
<td>LRRC2</td>
<td>KCNAB1</td>
<td>TMCC3</td>
<td>PLSCR4</td>
<td>EPB41L</td>
</tr>
<tr>
<td>DUSP13</td>
<td>DUSP13</td>
<td>DUSP13</td>
<td>SFTPD</td>
<td>SFTPD</td>
<td>TMEM88</td>
<td>FAM13</td>
<td>C</td>
<td>PLAC9</td>
<td>THBD</td>
</tr>
<tr>
<td>MYH6</td>
<td>MYH6</td>
<td>MYH6</td>
<td>DIRAS2</td>
<td>DIRAS2</td>
<td>DOC2B</td>
<td>MUSTN1</td>
<td>SHE</td>
<td>YPEL5</td>
<td>KLF4</td>
</tr>
<tr>
<td>CHRNA1</td>
<td>CHRNA1</td>
<td>CHRNA1</td>
<td>FOXQ1</td>
<td>FOXQ1</td>
<td>AVPR2</td>
<td>ABCA1</td>
<td>0</td>
<td>CARD8</td>
<td>KLF2</td>
</tr>
<tr>
<td>ARP21</td>
<td>ARP21</td>
<td>ARP21</td>
<td>VAT1L</td>
<td>VAT1L</td>
<td>FAM82A</td>
<td>FBXO33</td>
<td>STEAP1</td>
<td>MPDZ</td>
<td>METTL7A</td>
</tr>
<tr>
<td>LMOD3</td>
<td>LMOD3</td>
<td>LMOD3</td>
<td>ZNF488</td>
<td>ZNF488</td>
<td>FRMPD1</td>
<td>ADRB2</td>
<td>ABCA6</td>
<td>DENND5A</td>
<td>ARL6IP5</td>
</tr>
<tr>
<td>FAM25A</td>
<td>FAM25</td>
<td>FAM25A</td>
<td>ADORA2B</td>
<td>ADORA2B</td>
<td>SSTR1</td>
<td>CACNA2D1</td>
<td>RIMKL</td>
<td>DUSP3</td>
<td>MAP1B</td>
</tr>
<tr>
<td>ADRA1B</td>
<td>ADRA1B</td>
<td>ADRA1B</td>
<td>ASB9</td>
<td>ASB9</td>
<td>CCBE1</td>
<td>TTC33</td>
<td>C2orf194</td>
<td>AK3</td>
<td>LIMCH1</td>
</tr>
<tr>
<td>MYPN</td>
<td>MYPN</td>
<td>MYPN</td>
<td>ERMN</td>
<td>ERMN</td>
<td>C2orf86</td>
<td>RASL11A</td>
<td>DNAJB4</td>
<td>CHL1</td>
<td>PPAP2B</td>
</tr>
<tr>
<td>MYOM3</td>
<td>MYOM3</td>
<td>MYOM3</td>
<td>JPH3</td>
<td>JPH3</td>
<td>CBLN4</td>
<td>PKN3</td>
<td>C6orf72</td>
<td>PDGFD</td>
<td>LJFR</td>
</tr>
<tr>
<td>SRL</td>
<td>SRL</td>
<td>SRL</td>
<td>ANKRD37</td>
<td>ANKRD37</td>
<td>CUBN</td>
<td>LRRN3</td>
<td>KAT2B</td>
<td>SELE</td>
<td>FMO2</td>
</tr>
<tr>
<td>ABLIM2</td>
<td>ABLIM2</td>
<td>ABLIM2</td>
<td>EYA4</td>
<td>EYA4</td>
<td>FOXP2</td>
<td>SLCO9A1</td>
<td>AG2</td>
<td>MMRN1</td>
<td>DPYSL2</td>
</tr>
<tr>
<td>IP6K3</td>
<td>IP6K3</td>
<td>IP6K3</td>
<td>CSMD1</td>
<td>CSMD1</td>
<td>ZNF366</td>
<td>TRIM35</td>
<td>FGF7</td>
<td>ESD</td>
<td>ENPP2</td>
</tr>
<tr>
<td>ITIH4</td>
<td>ITIH4</td>
<td>ITIH4</td>
<td>FAM150A</td>
<td>FAM150B</td>
<td>MAP1C3C</td>
<td>ATR8B4</td>
<td>DUSP22</td>
<td>FAM107A</td>
<td>GNG12</td>
</tr>
<tr>
<td>MYO18B</td>
<td>MYO18B</td>
<td>MYO18B</td>
<td>CALCA</td>
<td>CALCA</td>
<td>NLGN1</td>
<td>TLR3</td>
<td>ID1</td>
<td>CRY2</td>
<td>FBLN5</td>
</tr>
<tr>
<td>PGAM2</td>
<td>PGAM2</td>
<td>PGAM2</td>
<td>PCDHG1</td>
<td>PCDHG1</td>
<td>LOH3C</td>
<td>ADAM18</td>
<td>MECOM</td>
<td>PROS1</td>
<td>ILIR1</td>
</tr>
<tr>
<td>RRAD</td>
<td>RRAD</td>
<td>RRAD</td>
<td>GCNT3</td>
<td>GCNT3</td>
<td>LRCH2</td>
<td>SYT15</td>
<td>LOC339</td>
<td>290</td>
<td>AKAP2</td>
</tr>
<tr>
<td>MYLPF</td>
<td>MYLPF</td>
<td>MYLPF</td>
<td>USH1C</td>
<td>USH1C</td>
<td>CSF3</td>
<td>MCTP1</td>
<td>CDH13</td>
<td>IGFBP6</td>
<td>SVEP1</td>
</tr>
<tr>
<td>CACNA1S</td>
<td>CACNA1S</td>
<td>CACNA1S</td>
<td>USP43</td>
<td>USP43</td>
<td>EBF2</td>
<td>C4orf31</td>
<td>ZDHHC2</td>
<td>SRPX</td>
<td>UTRN</td>
</tr>
<tr>
<td>MYOT</td>
<td>MYOT</td>
<td>MYOT</td>
<td>VTI1A</td>
<td>VTI1A</td>
<td>SCN9A</td>
<td>PRKD1</td>
<td>MAP3K8</td>
<td>MAP1L3C</td>
<td>JAK1</td>
</tr>
<tr>
<td>SLPN</td>
<td>SLPN</td>
<td>SLPN</td>
<td>STYK1</td>
<td>STYK1</td>
<td>RGS6</td>
<td>PREX2</td>
<td>NOV</td>
<td>####### #</td>
<td>OGN</td>
</tr>
<tr>
<td>TNNI2</td>
<td>TNNI2</td>
<td>TNNI2</td>
<td>KCNIP3</td>
<td>KCNIP3</td>
<td>ARHGA20</td>
<td>FAT4</td>
<td>APCDD1</td>
<td>CRTC4</td>
<td></td>
</tr>
<tr>
<td>STAC3</td>
<td>STAC3</td>
<td>STAC3</td>
<td>C6orf176</td>
<td>C6orf176</td>
<td>ITIH3</td>
<td>LILRB5</td>
<td>OSTM1</td>
<td>GALNTL2</td>
<td>ATF3</td>
</tr>
</tbody>
</table>

continued
Table A.3 continued: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>ANKRD1</th>
<th>ANKRD1</th>
<th>ANKRD1</th>
<th>PRSS21</th>
<th>PRSS21</th>
<th>RUNDC3B</th>
<th>NLRX1</th>
<th>STON1</th>
<th>FOXO1</th>
<th>KCTD12</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASQ1</td>
<td>CASQ1</td>
<td>CASQ1</td>
<td>C9orf110</td>
<td>C9orf110</td>
<td>WHAMML1</td>
<td>CDC35L1</td>
<td>RAPGEF3</td>
<td>IL33</td>
<td>SYNPO2</td>
</tr>
<tr>
<td>SPTB</td>
<td>SPTB</td>
<td>SPTB</td>
<td>C12orf5</td>
<td>C12orf5</td>
<td>C9orf125</td>
<td>CLN8</td>
<td>FLII</td>
<td>KDR</td>
<td>ABLIM1</td>
</tr>
<tr>
<td>PCDH20</td>
<td>PCDH20</td>
<td>PCDH20</td>
<td>FDE3A</td>
<td>PDE3A</td>
<td>KIAA1614</td>
<td>MRGPRF</td>
<td>PTPN21</td>
<td>STX7</td>
<td>TSC22D3</td>
</tr>
<tr>
<td>DDIT4L</td>
<td>DDIT4L</td>
<td>DDIT4L</td>
<td>RNFI82</td>
<td>RNFI82</td>
<td>C14orf28</td>
<td>MOBKL2C</td>
<td>ERG</td>
<td>VWA5A</td>
<td>CAST</td>
</tr>
<tr>
<td>XIRP1</td>
<td>XIRP1</td>
<td>XIRP1</td>
<td>FAM108B1</td>
<td>FAM108B1</td>
<td>RASSF9</td>
<td>LRRN4CL</td>
<td>IL6</td>
<td>LEPR</td>
<td>ADAMT51</td>
</tr>
<tr>
<td>TRDN</td>
<td>TRDN</td>
<td>TRDN</td>
<td>LOC154822</td>
<td>LOC154822</td>
<td>NUDT7</td>
<td>MFNG</td>
<td>TEK</td>
<td>NR4A2</td>
<td>JUNB</td>
</tr>
<tr>
<td>ANKRD23</td>
<td>ANKRD23</td>
<td>ANKRD23</td>
<td>TDRD9</td>
<td>TDRD9</td>
<td>MYOCD</td>
<td>DENND2A</td>
<td>ARHGEF15</td>
<td>ELTD1</td>
<td>PRELP</td>
</tr>
<tr>
<td>ENO3</td>
<td>ENO3</td>
<td>ENO3</td>
<td>PCP4L1</td>
<td>PCP4L1</td>
<td>P2RY14</td>
<td>KBTBD11</td>
<td>FOLR2</td>
<td>S1PR1</td>
<td>TGFBR3</td>
</tr>
<tr>
<td>LDB3</td>
<td>LDB3</td>
<td>LDB3</td>
<td>AVP1I</td>
<td>AVP1I</td>
<td>USHBPI</td>
<td>RHOJ</td>
<td>NDN</td>
<td>USP53</td>
<td>TRIB1</td>
</tr>
<tr>
<td>MYBPC2</td>
<td>MYBPC2</td>
<td>MYBPC2</td>
<td>ITFG2</td>
<td>ITFG2</td>
<td>P2RY12</td>
<td>RASA4</td>
<td>SAMD4A</td>
<td>GEM</td>
<td>MFAP4</td>
</tr>
<tr>
<td>VSIG10L</td>
<td>VSIG10L</td>
<td>VSIG10L</td>
<td>RASL10B</td>
<td>RASL10B</td>
<td>CLDN15</td>
<td>LPHN3</td>
<td>VIT</td>
<td>SPG20</td>
<td>MME</td>
</tr>
<tr>
<td>MYOZ1</td>
<td>MYOZ1</td>
<td>MYOZ1</td>
<td>INHA</td>
<td>INHA</td>
<td>TMEM31</td>
<td>LOC654342</td>
<td>ARHGEF10</td>
<td>PCDH18</td>
<td>LGALS3</td>
</tr>
<tr>
<td>CSRFP3</td>
<td>CSRFP3</td>
<td>CSRFP3</td>
<td>CDH19</td>
<td>CDH19</td>
<td>THSD7B</td>
<td>LPCAT4</td>
<td>GIMAP8</td>
<td>ZEB2</td>
<td>CRIM1</td>
</tr>
<tr>
<td>ARC</td>
<td>ARC</td>
<td>ARC</td>
<td>AMY1A</td>
<td>AMY1A</td>
<td>PARK2</td>
<td>CLCN6</td>
<td>EDNRB</td>
<td>VAMP3</td>
<td>SOCS3</td>
</tr>
<tr>
<td>TBLNB</td>
<td>TBLNB</td>
<td>TBLNB</td>
<td>MFSD4</td>
<td>MFSD4</td>
<td>CLEC1A</td>
<td>TMEM55A</td>
<td>FEZ2</td>
<td>SSPN</td>
<td>TACC1</td>
</tr>
<tr>
<td>LMOD2</td>
<td>LMOD2</td>
<td>LMOD2</td>
<td>MCTP2</td>
<td>MCTP2</td>
<td>ODZ1</td>
<td>ARHGA524</td>
<td>ADAMTSL4</td>
<td>GNG11</td>
<td>MYADM</td>
</tr>
<tr>
<td>TNC1</td>
<td>TNC1</td>
<td>TNC1</td>
<td>MOCOS</td>
<td>MOCOS</td>
<td>EFHA2</td>
<td>OTUD1</td>
<td>KIAA0355</td>
<td>GABAR</td>
<td>APL1</td>
</tr>
<tr>
<td>CHRN1</td>
<td>CHRN1</td>
<td>CHRN1</td>
<td>UST</td>
<td>UST</td>
<td>GRRP1</td>
<td>BMP6</td>
<td>CLDN5</td>
<td>RBPMS</td>
<td>CXCL12</td>
</tr>
<tr>
<td>ACTC1</td>
<td>ACTC1</td>
<td>ACTC1</td>
<td>GEMIN4</td>
<td>GEMIN4</td>
<td>AVPR1A</td>
<td>PCDHG4B7</td>
<td>PGM5</td>
<td>SWAP70</td>
<td>CPE</td>
</tr>
<tr>
<td>PYGM</td>
<td>PYGM</td>
<td>PYGM</td>
<td>FGFBP1</td>
<td>FGFBP1</td>
<td>PPP1R14A</td>
<td>LHX6</td>
<td>RAPGEF2</td>
<td>PPAP2A</td>
<td>TNXB</td>
</tr>
<tr>
<td>TNC1</td>
<td>TNC1</td>
<td>TNC1</td>
<td>C4BPA</td>
<td>C4BPA</td>
<td>PCDHGA9</td>
<td>PLCL2</td>
<td>LRRK2</td>
<td>CCL2</td>
<td>NR4A1</td>
</tr>
<tr>
<td>ATP2A1</td>
<td>ATP2A1</td>
<td>ATP2A1</td>
<td>ARNTL2</td>
<td>ARNTL2</td>
<td>ABCB5</td>
<td>ZNF330</td>
<td>LEPROT</td>
<td>ARRD4</td>
<td>EFEMP1</td>
</tr>
<tr>
<td>TCAP</td>
<td>TCAP</td>
<td>TCAP</td>
<td>CIGALT1</td>
<td>CIGALT1</td>
<td>ANKDD1A</td>
<td>CYP2U1</td>
<td>LDB2</td>
<td>PTPRM</td>
<td>EPAS1</td>
</tr>
<tr>
<td>ENHO</td>
<td>ENHO</td>
<td>ENHO</td>
<td>MGAT3</td>
<td>MGAT3</td>
<td>CDADC1</td>
<td>NOD1</td>
<td>CRYBG3</td>
<td>EP58</td>
<td>TGFBR2</td>
</tr>
<tr>
<td>TNNT3</td>
<td>TNNT3</td>
<td>TNNT3</td>
<td>CRTAC1</td>
<td>CRTAC1</td>
<td>BMX</td>
<td>RCBTB2</td>
<td>RECK</td>
<td>CCL21</td>
<td>ZFP36L</td>
</tr>
</tbody>
</table>

continued
Table A.3 continued: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>MS4A1 4</th>
<th>MS4A14</th>
<th>MS4A14</th>
<th>AEBP2</th>
<th>AEBP2</th>
<th>ATP1B2</th>
<th>FXYD1</th>
<th>TCN2</th>
<th>CD302</th>
<th>ITM2B</th>
</tr>
</thead>
<tbody>
<tr>
<td>RYR1</td>
<td>RYR1</td>
<td>RYR1</td>
<td>FOXJ2</td>
<td>FOXJ2</td>
<td>PHYHIP</td>
<td>TTC28</td>
<td>ZCCHC24</td>
<td>EMP1</td>
<td></td>
</tr>
<tr>
<td>TNNI1</td>
<td>TNNI1</td>
<td>TNNI1</td>
<td>GABBR2</td>
<td>GABBR2</td>
<td>CCDC152</td>
<td>GPR133</td>
<td>SNX18</td>
<td>TPTRB</td>
<td>AQP1</td>
</tr>
<tr>
<td>NRAP</td>
<td>NRAP</td>
<td>NRAP</td>
<td>TULP3</td>
<td>TULP3</td>
<td>RUNX1T1</td>
<td>ALDH1A2</td>
<td>CFL2</td>
<td>EPB41L3</td>
<td>DUSP1</td>
</tr>
<tr>
<td>MYL1</td>
<td>MYL1</td>
<td>MYL1</td>
<td>PDE10A</td>
<td>PDE10A</td>
<td>RFTN2</td>
<td>NPR2</td>
<td>HSD17B11</td>
<td>CELF2</td>
<td>ANXA1</td>
</tr>
<tr>
<td>MYL2</td>
<td>MYL2</td>
<td>MYL2</td>
<td>RNF128</td>
<td>RNF128</td>
<td>MAB21L1</td>
<td>ATOH8</td>
<td>SHANK3</td>
<td>FGL2</td>
<td>VWF</td>
</tr>
<tr>
<td>KBTBD10</td>
<td>KBTBD10</td>
<td>KBTBD10</td>
<td>PDE4D</td>
<td>PDE4D</td>
<td>EDA</td>
<td>ANKRD29</td>
<td>PPP1R16B</td>
<td>NFE2L2</td>
<td>CYR61</td>
</tr>
<tr>
<td>ACTN2</td>
<td>ACTN2</td>
<td>ACTN2</td>
<td>KSR1</td>
<td>KSR1</td>
<td>ST8GALNAC3</td>
<td>GRK5</td>
<td>NOTCH4</td>
<td>DDX24</td>
<td>BTG2</td>
</tr>
<tr>
<td>CSF3R</td>
<td>CSF3R</td>
<td>CSF3R</td>
<td>MFF</td>
<td>MFF</td>
<td>RNLS</td>
<td>SOX7</td>
<td>EGFL7</td>
<td>LYVE1</td>
<td>SPARCL1</td>
</tr>
<tr>
<td>XIRP2</td>
<td>XIRP2</td>
<td>XIRP2</td>
<td>CDK18</td>
<td>CDK18</td>
<td>SCN3B</td>
<td>TACR1</td>
<td>FLRT2</td>
<td>ITM2A</td>
<td>FOSB</td>
</tr>
<tr>
<td>MYH1</td>
<td>MYH1</td>
<td>MYH1</td>
<td>GDF10</td>
<td>GDF10</td>
<td>WSCD1</td>
<td>PTGER4</td>
<td>IGFS10</td>
<td>ABCA8</td>
<td>ZFP36</td>
</tr>
<tr>
<td>CA3</td>
<td>CA3</td>
<td>RHCG</td>
<td>RHCG</td>
<td>LLRC70</td>
<td>GPR34</td>
<td>C13orf15</td>
<td>PTEN</td>
<td>SEPP1</td>
<td></td>
</tr>
<tr>
<td>FLNC</td>
<td>FLNC</td>
<td>SLNC</td>
<td>SLC45A4</td>
<td>SLC45A4</td>
<td>GIPC2</td>
<td>INMT</td>
<td>F8</td>
<td>ADAMTS5</td>
<td>FOS</td>
</tr>
<tr>
<td>CCM</td>
<td>CCM</td>
<td>CCM</td>
<td>UPK3BL</td>
<td>UPK3BL</td>
<td>LOC644538</td>
<td>INPP5A</td>
<td>DPYD</td>
<td>SCARA5</td>
<td>EGR1</td>
</tr>
<tr>
<td>MB</td>
<td>MB</td>
<td>MB</td>
<td>MAPK4</td>
<td>MAPK4</td>
<td>MASP1</td>
<td>PLEKHG5</td>
<td>ISCU</td>
<td>MMRN2</td>
<td>TXNIP</td>
</tr>
<tr>
<td>TNNT1</td>
<td>TNNT1</td>
<td>TNNT1</td>
<td>B4GALNT3</td>
<td>B4GALNT3</td>
<td>SEMA3A</td>
<td>PEAR1</td>
<td>PIK3C2B</td>
<td>BCL6</td>
<td>GSN</td>
</tr>
<tr>
<td>MYH2</td>
<td>MYH2</td>
<td>MYH2</td>
<td>MAP2K1</td>
<td>MAP2K1</td>
<td>C20orf60</td>
<td>PTGFR</td>
<td>SLC31A2</td>
<td>CCL14</td>
<td>DCN</td>
</tr>
<tr>
<td>ACTA1</td>
<td>ACTA1</td>
<td>ACTA1</td>
<td>DDX47</td>
<td>DDX47</td>
<td>RBMS3</td>
<td>CNRIP1</td>
<td>TLR4</td>
<td>Ccdc50</td>
<td>AHNK</td>
</tr>
<tr>
<td>NEB</td>
<td>NEB</td>
<td>NEB</td>
<td>MOXD1</td>
<td>MOXD1</td>
<td>DNASE1L3</td>
<td>DPH3</td>
<td>SNTB2</td>
<td>SH3BP5</td>
<td></td>
</tr>
<tr>
<td>MYH7</td>
<td>MYH7</td>
<td>MYH7</td>
<td>PKP2</td>
<td>PKP2</td>
<td>CH25H</td>
<td>PLK3</td>
<td>NR4A3</td>
<td>LHFP</td>
<td></td>
</tr>
<tr>
<td>TTN</td>
<td>TTN</td>
<td>TTN</td>
<td>TAPBPL</td>
<td>TAPBPL</td>
<td>HYAL1</td>
<td>CD209</td>
<td>ZNF503</td>
<td>SYNE1</td>
<td></td>
</tr>
<tr>
<td>DES</td>
<td>DES</td>
<td>DES</td>
<td>PLA2G4A</td>
<td>PLA2G4A</td>
<td>TAL1</td>
<td>ADAM3</td>
<td>ZH3</td>
<td>PRG4</td>
<td></td>
</tr>
<tr>
<td>PEG3</td>
<td>PEG3</td>
<td>PEG3</td>
<td>PEX5</td>
<td>PEX5</td>
<td>PDE7B</td>
<td>ARAP3</td>
<td>RAMP2</td>
<td>KLF13</td>
<td></td>
</tr>
<tr>
<td>RGD1</td>
<td>RGD1</td>
<td>RGD1</td>
<td>SCN7A</td>
<td>JAM2</td>
<td>CEP68</td>
<td>SPRY1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRLF1</td>
<td>CRLF1</td>
<td>PVRL3</td>
<td>ARL15</td>
<td>VGLL3</td>
<td>PPP1R15A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POLR2J3</td>
<td>POLR2J3</td>
<td>TRPC1</td>
<td>CLIC5</td>
<td>EMCN</td>
<td>MSRB3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EYA2</td>
<td>EYA2</td>
<td>EYD2</td>
<td>PCDHG8B6</td>
<td>LMO2</td>
<td>LPAR6</td>
<td>TFF1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

continued
Table A.3 continued: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>MYO1E</th>
<th>MYO1E</th>
<th>CTSG</th>
<th>RASIP1</th>
<th>MEF2C</th>
<th>C4orf3</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMOC1</td>
<td>SMOC1</td>
<td>EMX2O</td>
<td>FAM69A</td>
<td>PKD2</td>
<td>GPR116</td>
</tr>
<tr>
<td>PITPN</td>
<td>PITPN</td>
<td>GRASP</td>
<td>FLT4</td>
<td>ZNF436</td>
<td>PTGIS</td>
</tr>
<tr>
<td>ST3GAL4</td>
<td>ST3GAL4</td>
<td>GFRA2</td>
<td>HBEGF</td>
<td>LIMS2</td>
<td>ETS2</td>
</tr>
<tr>
<td>PITX1</td>
<td>PITX1</td>
<td>MT1M</td>
<td>CA5B</td>
<td>PHF17</td>
<td>C7</td>
</tr>
<tr>
<td>RHOV</td>
<td>RHOV</td>
<td>KANK3</td>
<td>ZBTB16</td>
<td>ABCA9</td>
<td>ERAP1</td>
</tr>
<tr>
<td>MANS C1</td>
<td>MANS C1</td>
<td>XPNPEP</td>
<td>PID1</td>
<td>STX12</td>
<td>CDH5</td>
</tr>
<tr>
<td>PTGS1</td>
<td>PTGS1</td>
<td>HAS1</td>
<td>MEOX2</td>
<td>GIMAP6</td>
<td>STAT5B</td>
</tr>
<tr>
<td>PLBD1</td>
<td>PLBD1</td>
<td>F10</td>
<td>KLF8</td>
<td>SKAP2</td>
<td>C1orf11</td>
</tr>
<tr>
<td>NDUFA9</td>
<td>NDUFA9</td>
<td>WSCD2</td>
<td>RBPJ</td>
<td>SNRK</td>
<td>DARC</td>
</tr>
<tr>
<td>SLC6A14</td>
<td>SLC6A14</td>
<td>KL</td>
<td>CREB5</td>
<td>P1K3P1</td>
<td>LAMA2</td>
</tr>
<tr>
<td>MAP7</td>
<td>MAP7</td>
<td>TMEM220</td>
<td>TMOD1</td>
<td>RGL1</td>
<td>IRS2</td>
</tr>
<tr>
<td>PTHLH</td>
<td>PTHLH</td>
<td>FCER1A</td>
<td>FREM1</td>
<td>C1orf3</td>
<td>ABI3BP</td>
</tr>
<tr>
<td>TM7SF3</td>
<td>TM7SF3</td>
<td>LG14</td>
<td>ARHGA10</td>
<td>RBMS2</td>
<td>IGF1</td>
</tr>
<tr>
<td>M6PR</td>
<td>M6PR</td>
<td>HOXD9</td>
<td>EIF1B</td>
<td>TPRG1L</td>
<td>CLEC3B</td>
</tr>
<tr>
<td>GALNT2</td>
<td>GALNT2</td>
<td>MSRA</td>
<td>SHROOM4</td>
<td>EIF4E3</td>
<td>C5orf4</td>
</tr>
<tr>
<td>TMPRS3</td>
<td>TMPRS3</td>
<td>THSD1</td>
<td>RSPO3</td>
<td>ROBO4</td>
<td>ALDH1A1</td>
</tr>
<tr>
<td>A2ML1</td>
<td>A2ML1</td>
<td>CCL13</td>
<td>GULP1</td>
<td>IRAK3</td>
<td>PECAM1</td>
</tr>
<tr>
<td>FRZB</td>
<td>FRZB</td>
<td>ECSCR</td>
<td>FBXO31</td>
<td>KCNJ8</td>
<td>FOXN3</td>
</tr>
<tr>
<td>MTCH1</td>
<td>MTCH1</td>
<td>KIF26A</td>
<td>RNF125</td>
<td>TEF</td>
<td>AOX1</td>
</tr>
<tr>
<td>SIK1</td>
<td>SIK1</td>
<td>ARHGA6</td>
<td>PLEKHM1</td>
<td>EZH1</td>
<td>CXCR7</td>
</tr>
<tr>
<td>PHB2</td>
<td>PHB2</td>
<td>CUTC</td>
<td>LATNA1L1</td>
<td>SH3BG2</td>
<td>H6PD</td>
</tr>
<tr>
<td>CLSTN3</td>
<td>CLSTN3</td>
<td>NR5A2</td>
<td>ICAM2</td>
<td>PDE2A</td>
<td>DLC1</td>
</tr>
<tr>
<td>ATN1</td>
<td>ATN1</td>
<td>C2CD4B</td>
<td>PLEKHH2</td>
<td>NYNRIN</td>
<td>TMTC1</td>
</tr>
<tr>
<td>CCND2</td>
<td>CCND2</td>
<td>RNF180</td>
<td>EGFLAM</td>
<td>SLC25A25</td>
<td>PARM1</td>
</tr>
<tr>
<td>PLA2G2A</td>
<td>PLA2G2A</td>
<td>C7orf23</td>
<td>ANGPT1L1</td>
<td>COL4A3BP</td>
<td>PI16</td>
</tr>
<tr>
<td>PTP4A1</td>
<td>PTP4A1</td>
<td>TMCO7</td>
<td>ISC1A</td>
<td>PRICKL2E2</td>
<td>LTBP4</td>
</tr>
<tr>
<td>ECEL1</td>
<td>ECEL1</td>
<td>WBP4</td>
<td>CLIC2</td>
<td>TSPAN7</td>
<td>NR3C1</td>
</tr>
</tbody>
</table>

continued
Table A.3 continued: Metagenes that identified to be associated with image features in Triple Negative Breast Cancer.

<table>
<thead>
<tr>
<th>MSLN</th>
<th>MSLN</th>
<th>RERGL</th>
<th>AASS</th>
<th>THRB</th>
<th>KLF9</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGA</td>
<td>CGA</td>
<td>CCDC46</td>
<td>PLEKHM3</td>
<td>PDZRN3</td>
<td>SDPR</td>
</tr>
<tr>
<td>NDRG1</td>
<td>NDRG1</td>
<td>SYN2</td>
<td>GGTAA1</td>
<td>HNMT</td>
<td>SLIT3</td>
</tr>
<tr>
<td>GAPDH</td>
<td>GAPDH</td>
<td>KLHL3</td>
<td>TBRG1</td>
<td>CYYR1</td>
<td>RNASE4</td>
</tr>
</tbody>
</table>

Table A.4: Subtype-Specific Genes for Subtype 1.

<table>
<thead>
<tr>
<th>ABCD1</th>
<th>ATG7</th>
<th>C2orf24</th>
<th>CTSL1</th>
<th>FCGR1C</th>
<th>IL4I1</th>
<th>MYSM1</th>
<th>ROR1</th>
<th>SS18</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABP1</td>
<td>ATP5A1</td>
<td>C2</td>
<td>DCAF16</td>
<td>FGGY</td>
<td>JAKMIP1</td>
<td>NGLY1</td>
<td>RTTN</td>
<td>TLE4</td>
</tr>
<tr>
<td>ADORA3</td>
<td>ATP6V0D2</td>
<td>C3orf38</td>
<td>DCX</td>
<td>FLVCR2</td>
<td>KCTD1</td>
<td>OSCAR</td>
<td>SEMA6A</td>
<td>TMEM150B</td>
</tr>
<tr>
<td>AKR1B10</td>
<td>ATP9B</td>
<td>C5AR1</td>
<td>DNAJC5B</td>
<td>FUC2A</td>
<td>KIAA1199</td>
<td>PAK7</td>
<td>SETBP1</td>
<td>TMEM200C</td>
</tr>
<tr>
<td>AKT3</td>
<td>BACH2</td>
<td>CCR8</td>
<td>EIF2S2</td>
<td>GOT2</td>
<td>KIAA1328</td>
<td>PCTP</td>
<td>SIGLEC1</td>
<td>VSTM2A</td>
</tr>
<tr>
<td>ALKBH6</td>
<td>BHLHE22</td>
<td>CCR5</td>
<td>EPYC</td>
<td>GRIN2D</td>
<td>LDB1</td>
<td>PHLPP1</td>
<td>SIRPD</td>
<td>WDT1</td>
</tr>
<tr>
<td>AMOTL1</td>
<td>BIRC7</td>
<td>CHRNA1</td>
<td>ETS1</td>
<td>HP1BP3</td>
<td>LPHN2</td>
<td>PLEKHG1</td>
<td>SLC16A3</td>
<td>ZNF33A</td>
</tr>
<tr>
<td>ANKRD58</td>
<td>BMS1P5</td>
<td>CPNE8</td>
<td>FAM126A</td>
<td>HS3ST2</td>
<td>MAML2</td>
<td>PRICKLE1</td>
<td>SLC34A3</td>
<td>ZNF397OS</td>
</tr>
<tr>
<td>ANKRD6</td>
<td>BTNL8</td>
<td>CSGALNACT1</td>
<td>FAM171A</td>
<td>HTRA4</td>
<td>MAPRE1</td>
<td>RBBP8</td>
<td>SLC6A12</td>
<td>ZNF397</td>
</tr>
<tr>
<td>APOC1</td>
<td>C10orf10</td>
<td>CTSA</td>
<td>FCGR1B</td>
<td>IGFN1</td>
<td>MBD1</td>
<td>RENBP</td>
<td>SPRR3</td>
<td>ZNF521</td>
</tr>
</tbody>
</table>
Table A.5: The Demographics of TCGA breast cancer subset used in Chapter 7 and Chapter 8.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n=441</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Diagnosis, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26-45 years</td>
<td>93</td>
<td>21.08844</td>
</tr>
<tr>
<td>46-55 years</td>
<td>108</td>
<td>24.4898</td>
</tr>
<tr>
<td>56-65 years</td>
<td>122</td>
<td>27.6644</td>
</tr>
<tr>
<td>older than 66 years</td>
<td>118</td>
<td>26.75737</td>
</tr>
<tr>
<td>Disease Stage, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>43</td>
<td>9.750567</td>
</tr>
<tr>
<td>IA</td>
<td>33</td>
<td>7.482993</td>
</tr>
<tr>
<td>IB</td>
<td>3</td>
<td>0.680272</td>
</tr>
<tr>
<td>IIA</td>
<td>151</td>
<td>34.24036</td>
</tr>
<tr>
<td>IIB</td>
<td>95</td>
<td>21.54195</td>
</tr>
<tr>
<td>IIIA</td>
<td>55</td>
<td>12.47166</td>
</tr>
<tr>
<td>IIIB</td>
<td>13</td>
<td>2.947846</td>
</tr>
<tr>
<td>IIC</td>
<td>18</td>
<td>4.081633</td>
</tr>
<tr>
<td>IV</td>
<td>13</td>
<td>2.947846</td>
</tr>
<tr>
<td>Tumor Grade, no (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1</td>
<td>124</td>
<td>28.11791</td>
</tr>
<tr>
<td>T2</td>
<td>243</td>
<td>55.10204</td>
</tr>
<tr>
<td>T3</td>
<td>51</td>
<td>11.56463</td>
</tr>
<tr>
<td>T4</td>
<td>19</td>
<td>4.30839</td>
</tr>
<tr>
<td>ER, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>342</td>
<td>77.55102</td>
</tr>
<tr>
<td>Negative</td>
<td>91</td>
<td>20.63492</td>
</tr>
<tr>
<td>PR, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>295</td>
<td>66.89342</td>
</tr>
<tr>
<td>Negative</td>
<td>138</td>
<td>31.29252</td>
</tr>
<tr>
<td>HER2, no (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>112</td>
<td>25.39683</td>
</tr>
<tr>
<td>Negative</td>
<td>153</td>
<td>34.69388</td>
</tr>
<tr>
<td>N/A</td>
<td>176</td>
<td>39.9093</td>
</tr>
</tbody>
</table>
Appendix B : Derivation of optimization in Algorithm 1 in Chapter 8

Based on the definitions and derivation in Chapter 8, to achieve the integration of molecular patient classification with clinical classification, we propose a Molecular Regularized Consensus Patient Stratification (MRCPS), which minimizes the objective function in Eq. (8.8) in the Section 8.2.3, as follows:

\[
J(\bar{U}, D) = \|\bar{S} - W \circ \bar{U}D\bar{U}^T\|_F \text{ s.t. } \bar{U}^T\bar{U} = 1_{k \times k}; D_{ij} > 0. \tag{D.1}
\]

\[
J(\bar{U}, D) = \|\bar{S} - W \circ \bar{U}D\bar{U}^T\|_F^2
= \text{tr} \left[ (\bar{S} - W \circ \bar{U}D\bar{U}^T)(\bar{S} - W \circ \bar{U}D\bar{U}^T)^T \right]
= \text{tr} \left[ S^2 - 2(W \circ S)\bar{U}D\bar{U}^T + (W \circ \bar{U}D\bar{U}^T)\bar{U}D\bar{U}^T \right] \tag{D.2}
\]

where \(\text{tr}(\cdot)\) denotes the trace of a matrix and \(\bar{S}\) and \(W\) are symmetric matrices.

The derivative of \(J(\bar{U}, D)\) with respect to \(\bar{U}\) is

\[
\frac{\partial J(\bar{U}, D)}{\partial \bar{U}} = -4(W \circ S)\bar{U}D + 4(W \circ \bar{U}D\bar{U}^T)\bar{U}D. \tag{D.3}
\]

Using the Karush-Kuhn-Tucker complementary condition (KKT condition) for the non-negativity of \(\bar{U}\), we have

\[
(-4(W \circ S)\bar{U}D + 4(W \circ \bar{U}D\bar{U}^T)\bar{U}D)_{ij}\bar{U}_{ij} = 0 \tag{D.4}
\]

which leads to the following update rule for \(\bar{U}\):

\[
\bar{U}_{ij} \leftarrow \bar{U}_{ij} \frac{[(W \circ S)\bar{U}D]_{ij}}{[(W \circ \bar{U}D\bar{U}^T)\bar{U}D]_{ij}} \tag{D.5}
\]
Similarly, the computation of $D$ can be achieve by the derivative of $J(\bar{U}, D)$ with respect to $D$ is

$$
\frac{\partial J(\bar{U}, D)}{\partial U} = -2\bar{U}^T (W \circ S)\bar{U} + 2\bar{U}^T (W \circ \bar{U}D\bar{U}^T)\bar{U}.
$$

(D.6)

Using the Karush-Kuhn-Tucker complementary condition (KKT condition) for the non-negativity of $D$, we have

$$
(-2\bar{U}^T (W \circ S)\bar{U} + 2\bar{U}^T (W \circ \bar{U}D\bar{U}^T)\bar{U})_{ij} D_{ij} = 0,
$$

(D.7)

which leads to the following update rule for $D$:

$$
D_{ij} \leftarrow D_{ij} \frac{[\bar{U}^T (S \circ W)]_{ij}}{[\bar{U}^T (\bar{U}D\bar{U}^T \circ W)]_{ij}}.
$$

(D.8)
Appendix C: Proof of Convergence in Algorithm 1 in Chapter 8

The objective of this proof is to show that the cost function in Eq. (8.8) (Chapter 8) is non-increasing under the update steps of Eq. (8.9) and Eq. (8.10). The proof follows the similar procedures used in the Expectation-Maximization algorithms.

**Definition 1.1 - Auxiliary Function:** $G(v, v')$ is an auxiliary function for $F(v)$, if the conditions

$$G(v, v') \geq F(v), G(v, v) = F(v) \quad (C.1)$$

are satisfied.

**Lemma 1.1** If $G(v, v')$ is an auxiliary function of $F(v)$, then $F(v)$ is non-increasing under the following updates:

$$v^{(t+1)} = \text{argmin}_v G(v, v^{(t)}) \quad (C.2)$$

**Proof.** $F(v^{(t+1)}) \leq G(v^{(t+1)}, v^{(t)}) \leq G(v^{(t)}, v^{(t)}) = F(v^{(t)}) \quad (C.3)$

**Proposition 1.1.** For any nonnegative matrices $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^{k \times k}$, $U \in \mathbb{R}^{N \times k}$, $U' \in \mathbb{U}^{N \times k}$, and $A$, $B$ are symmetric, then the following inequality holds:

$$\sum_{ij} \frac{(AU' B)_{ij} u'_{ij}}{u_{ij}'} \geq \text{tr}(U^T A U B). \quad (C.4)$$

For the update of $\bar{U}$, the objective function can be rewritten as:

$$L(\bar{U}) = \text{tr}\left[-2(W \circ S)\bar{U}D\bar{U}^T\right] + \text{tr}\left[(W \circ \bar{U}D\bar{U}^T)\bar{U}D\bar{U}^T\right]. \quad (C.5)$$

**Lemma 1.2** Function
\[ G(\bar{U}, \bar{U}') = -2 \sum_{ij} [(W \circ S)\bar{U}D]_{ij} \bar{U}'_{ij} \left( 1 + \log \frac{\bar{u}_{ij}}{\bar{u}'_{ij}} \right) + \sum_{ij} \frac{[(W \circ \bar{D}\bar{U}^T)\bar{U}D]_{ij} \bar{u}'_{ij}^2}{\bar{u}'_{ij}} \]  

(C.6)
is an auxiliary function for \( L(\bar{U}) \).

Proof.

1. \( G(\bar{U}, \bar{U}) = L(\bar{U}) \) is obvious.

2. Proof of \( G(\bar{U}, \bar{U}') \geq L(\bar{U}) \): Let \( z = \frac{\bar{u}_{ij}}{\bar{u}'_{ij}} \). Because \( \bar{u}_{ij} \) is guaranteed due to its nonnegative property, \( z > 0 \). And consider the inequality \( z \geq 1 + \log z \), we have

\[ \frac{\bar{u}_{ij}}{\bar{u}'_{ij}} \geq 1 + \log \left( \frac{\bar{u}_{ij}}{\bar{u}'_{ij}} \right). \]  

(C.7)

So, from Eq. (C.5), the first term is bounded by

\[ \text{tr} \left\{ -2(W \circ S)\bar{U}D\bar{U}^T \right\} = -2 \sum_{ij} [(W \circ S)\bar{U}D]_{ij} \bar{U}_{ij} \]

\[ \leq -2 \sum_{ij} [(W \circ S)\bar{U}D]_{ij} \bar{U}'_{ij} \left( 1 + \log \frac{\bar{u}_{ij}}{\bar{u}'_{ij}} \right) \]  

(C.8)

Using Proposition 1.1, and setting \( A = W \circ \bar{U}D\bar{U}^T \) and \( B = D \), the second term in Eq. (C.5) is bounded by

\[ \text{tr} \left\{ (W \circ \bar{U}D\bar{U}^T)\bar{U}D\bar{U}^T \right\} = \text{tr} \left\{ \bar{U}^T(W \circ \bar{U}D\bar{U}^T)\bar{U}D \right\} \leq \sum_{ij} \frac{[(W \circ \bar{U}D\bar{U}^T)\bar{U}D]_{ij} \bar{u}'_{ij}^2}{\bar{u}'_{ij}}. \]  

(C.9)

Based on both bounds from Eq. (C.8) and Eq. (C.9), we can prove that

\[ G(\bar{U}, \bar{U}') \geq L(\bar{U}) \]. Thus, Lemma 1.2 holds.

Since Eq. (C.6) is an auxiliary function, \( L(\bar{U}) \) is non-increasing under the update rule.
For the update of $D$, the objective function can be rewritten as:

$$L(D) = \text{tr}[-2\bar{U}^T(W \circ S)\bar{U}D] + \text{tr}[(\bar{U}^T(W \circ D\bar{U}^T)\bar{U}D].$$ (C.10)

**Lemma 1.3** Function

$$G(D, D') = -2 \sum_{i,j} \left[\bar{U}^T(W \circ S)\bar{U}\right]_{ij} D_{ij}' \left(1 + \log \frac{D_{ij}}{D_{ij}'}\right) + \sum_{i,j} \frac{[\partial (W \circ S) D \partial D]}{D_{ij}'} D_{ij}'^2.$$ (C.11)

is an auxiliary function for $L(D)$.

**Proof.** Similarly to the Proof of $\bar{U}$, we can also prove Lemma 1.3. So, under the update of $D_{ij}$, the cost function in Eq. (8.8) in section 8.2.3 is non-increasing, hence it converges.