January 2020

Methods To Integrate Genetic And Clinical Data For Disease Subtyping

Diana Mabel Diaz-Herrera
Wayne State University

METHODS TO INTEGRATE GENETIC AND CLINICAL DATA FOR DISEASE SUBTYPING

by

DIANA M. DIAZ HERRERA

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2020

MAJOR: COMPUTER SCIENCE

Approved By:

_____

Advisor Date

_____

_____

_____

_____

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Enormous efforts have been made to collect genetic and clinical data from cancer patients to advance the understanding of disease development and progression. Processing and analyzing these flows of data is challenging. Many computational methods have been proposed to help different fronts of biology and medicine. The personalized medicine[1] (PM) paradigm recognizes that many factors influence the diagnosis, treatment, and prognosis of patients, factors that could vary among patients. The integration of clinical and genetic data using computational methods towards personalized medicine is considered the future for oncology studies [9, 64], and this thesis contributes in this direction. This thesis mainly presents data integration approaches to identify granular and meaningful disease sub-types from heterogeneous genetic and clinical data, which is an essential step towards PM implementation. Although research has shown that integrating different data sources for disease subtyping[2] increases the analytic power [46, 43, 48, 44, 45], integrative approaches are in their early stages due to many computational challenges, such as high-dimensionality, data collection noise, and heterogeneity of data sources [45, 76, 363, 340, 47, 158, 283]. Here, we present three new methods to integrate high-dimensional genetic data and clinical variables to elucidate disease-subgroups and their different biological mechanisms.

Disease subtyping is an important research topic in health-care informatics. For example, it is well known that "breast cancer" (BC) encompasses several profoundly different BC sub-types, such as ER+, PR+, HER2+, and triple-negative. The specific sub-type of cancer highly conditions diagnosis, treatment plan, treatment success, prognosis, and response to treatment. Identifying subgroups of patients that can lead to disease sub-types incorporating clinical and genetic data is still a challenge. Many attempts to disease subtyping based solely on genetic characteristics of patients have been undertaken but returned only moderate success so far (for example, rarely gene expression tests have been FDA approved).

[1] screening, diagnostic, therapeutic, and prognostic procedures that take into account individual variability of patients [9, 64].

[2] partitioning patients into outcome related cohorts

This thesis is organized as follows. Chapter 2 presents the biological terminology and concepts that are studied with our proposed computational methods. Chapter 3 stages a comprehensive literature review of computational methods for disease subtyping. Here we examine many tools and technologies that have been applied to cancer patient stratification, also known as disease subtyping. Some of these tools focus on one of the many types of patient data, while others use a combination of various data types. Here we approach disease subtyping as a computational problem to bridge the gap between oncology and computer science. We also study and compare the different techniques that have been used to tackle this problem, as well as highlight opportunities for new methods. Chapter 4 presents our integrative disease subtyping based on microRNA, gene expression, and gene networks. We demonstrate the feasibility of the algorithm using real data patients from publicly available datasets. We show that data integration is critical to identify sub-types of diseases and their underlined biological mechanisms. In Chapter 5, we describe the current limitations of integrative approaches and propose the integration of clinical data to genetic subtyping to find relations between the clinical and genetic components, in addition to identifying meaningful sub-types of patients. In Chapter 6, we extended our integrative pipeline by incorporating gene expression to somatic mutation and clinical data for subtyping. We tested our method with data from six different cancer types and identified meaningful sub-types of patients. These results can potentially have important impacts on diagnosis, prognosis, and treatment of cancer. We conclude this thesis by proposing future works.

1.1 Integrating high-throughput biomolecular data

The advent of high-throughput genomics technologies has resulted in massive amounts of diverse genome-scale data. These technologies measure molecular expression at different levels, such as gene expression, microRNA expression, protein abundance, DNA methylation, and copy number variation, across the whole genome. Simultaneously analyzing multiple data types allows us to gain a more comprehensive view and a deeper understanding of complex diseases that any single data type analysis is unable to provide. Pathway analysis

and disease subtyping are often the first steps needed to interpret the diverse data types and to gain insights into biological processes. Despite the efforts, the integration of massive amounts of high-dimensional and diverse data types presents significant algorithmic and computational challenges. These challenges include the curse of dimensionality of the high-throughput data, contradicting and redundant signals from heterogeneous types of data, poor reproducibility of independent studies designed for the same disease, and algorithmic complexity. In this introduction, we focus on approaches that integrate multi-omics data for disease subtyping and show the differences, advantages, and disadvantages compared to the existing methods.

The Centers for Disease Control and Prevention (CDC) expects the number of new cancer cases in the United States to go up about 24% in men to more than 1 million cases per year, and by about 21% in women to more than 900,000 cases per year until 2020[3]. These alarming figures motivate to accelerate the research on cancer, which underlines complicated cellular processes that are not fully understood. These processes encompass thousands of chemical compounds and physical reactions. High-throughput molecular biological methods perform thousands of simultaneous measurements of biological molecules to read a particular state of cells. Recent technologies have extended the broadness of available high-throughput molecular biological data. Nowadays, most of the molecular data types are analyzed separately. Single data type studies have provided essential discoveries like biomarkers for some diseases. However, analyzing various data types together can potentially lead to a more coherent understanding of cellular processes [124].

The term *high-throughput data* is referred to here as large measures of genetic data taken in a short time. Different technologies generate these data, commonly referred to as "omics technologies", which are the foundation for systems biology [268]. Omics seek to quantify, describe, and identify all of the components on cellular systems with spatial and temporal dimensions [285]. There are several data types of high-throughput measurements from which

[3] "Expected New Cancer Cases and Deaths in 2020", The Centers for Disease Control and Prevention (CDC), viewed Feb. 10, 2020, https://www.cdc.gov/

four categories are the most important: proteomics, transcriptomics, metabolomics, and genomics [362]. Proteomics is the large-scale study of proteins present in cells. Transcriptomics measures all gene expression values. Metabolomics aims for the quantification and identification of metabolites. Furthermore, genomics includes the large-scale genotyping of SNP (single nucleotide polymorphisms) and epigenomics. Each of these data types is unique and provides different perspectives on cellular processes.

Proteomic measurements strive to determine the presence and quantity of the proteins that have been translated into a sample. The process of protein identification is performed serially and rapidly using antibodies or Mass Spectrometry (MS). The general process for an MS study starts by preparing the sample. Then proteins are detached using chromatography, which typically consists of several protein gels that can be: 1-dimensional protein gels which detach proteins based on size, or 2-Dimensional protein gels which detach proteins based on size and electrical charge. Next, these gels are digested and drive through mass spectroscopy, which identifies the volume of the peptides. Finally, each type of protein can be recognized by querying on protein databases the volume of the peptides, which retrieves the corresponding protein. Proteomic measures are extensively used for early identification of diseases and disease subtyping.

Transcriptomic measurements or gene expression microarrays are the best established of the high-throughput technologies. The most common arrays consist of hundreds of thousands of probes. A broad pipeline of these technologies starts by extracting RNA from cells. Then perform purification of the sample to perform cDNA coupling then. Then, the cDNA is hybridized to the array. Finally, the probes are scanned to assess the expression level of approximately 30,000 different mRNAs. There are several statistical methods developed to perform each of these steps. Gene expression is largely used for early disease identification and disease subtyping.

Metabolomics target the measure of metabolites present in cells. Similar to proteomics, metabolites are measured using a rapid serial process. Typical technologies used to identify

and measure metabolomics are Nuclear Magnetic Resonance(NMR) spectroscopy, and MS. NMR can detect metabolites by comparing the measured signals with databases of reference compounds. Measurements of metabolites are known for their high correlation with phenotypes and are typically used for early disease detection and disease subtyping.

Finally, genomics determines which of the millions of Single Nucleotide Polymorphisms (SNPs) scattered through the whole genome an individual carries. SNPs are single variations (mutations) in the DNA sequence that occur in about 1 of a population. Analyzing the frequencies of these variations is important because it helps to determine the relationship of specific mutations with a particular disease. Genotyping can be performed using different technologies depending on the variants under study. When looking for several different variants, genotyping arrays are the right choice. The main advantage of genomic studies over the previous omics is that genotypes are useful for pre-disease prediction because genomic aberrations are present even before diseases start phenotypic manifestations.

All these different data types provide different levels of knowledge, and they should be considered to fully understand biological processes at the cellular level. There are several computational solutions for analyzing omics data in isolated fashion [24]; however, single measures have not given enough conclusions for disease diagnosis and treatment. Some of the ultimate goals for integrating multiple-omics are the identification of relevant pathways behind a condition, treatment response prediction, and disease subtyping.

The identification of pathways that are involved in a specific phenotype is typically referred to as pathway analysis. Identifying pathways that are relevant to a condition is essential because it gives insights that can be used to further disease treatment or diagnosis. The standard input of pathway analysis techniques is the fold change of two phenotypes. Fold change is computed from gene expression of two different groups, commonly one group of control subjects and another group with patients carrying a disease. The output of these methods is a ranked list of statistically significant biological pathways. These pathways are considered to be related to the condition under study. Biological pathways are graphical

representations of common knowledge about genes and their interactions as part of biological processes. Some of these graphs have a set of genes as nodes and the biochemical and physical interactions as edges. These pathways are typically made by mining the literature and then manually curating the retrieved information [182]. The components that are mostly analyzed are genes, proteins, and metabolites. Signaling processes of the cell are captured in signaling pathways that describe the interactions between protein and DNA level of protein-coding genes [189].

Treatment response prediction refers to the identification of patients that respond to a treatment and patients that do not. Prediction of treatment outcomes in complex diseases like cancer is crucial. Tumor size reduction and side effects are commonly expected outcomes from these approaches. For example, studies have proven that the integration of gene mutation with gene expression improves outcome prediction in some myelodysplastic syndromes (MDS) [131], which are a group of cancers for which blood cells in the bone marrow do not mature and never become healthy cells. Particularly for cancers, tumor progression can be tracked by identifying differentially expressed genes across two phenotypes. The input of treatment response predictors includes clinical or biological parameters registered at baseline during treatment. The output is a likelihood of a response to a given treatment or the prognostic ability of the models to distinguish between patients that responded to the treatment and patients that did not.

Generating clinical meaningful disease subtyping is critical for prognosis and further treatment determination. Based on statistical information and the patient's profile, the objective is to identify the sub-type of the disease that the patient more likely belongs to. The input for disease subtyping is molecular and clinical data of several patients that undergo a particular condition and have different outcomes. The expected output is well-identified groups that highly correlate with the observed survival (e.g., a group of long-term survival patients and another group of short-term survival patients). It is also essential to identify possible patterns that are shared among members of each sub-type and differences with other

sub-types. This is commonly posed as a clustering problem where the main goal is to find groups of patients that are highly similar to the members of its group, very different with the other groups, and the number of groups is unknown.

All these applications show how important is integrating various biomolecular data types. There are more applications and studies on genetic data integration like drug studies [351], signaling networks reconstruction [197, 241, 77], systems biology [130], or biological networks visualization [301]. From the computer science perspective, the term data integration refers to the integration of fragmented information from different physical databases or data warehouses and different representations. Several authors have proposed platforms and languages to integrate the databases (typically using XML) [3]. In bioinformatics, the terms data integration and data fusion are used indistinctly. In computer science, data fusion is referred to as the process of integrating information acquired from various heterogeneous types into a unique compound knowledge. Here we refer to data integration and data fusion as the integration of knowledge without focusing on the representation. Additionally, data fusion is valuable for acquiring more reliable information than original measurements from a single type of source. The primary issue in data fusion (DF) is to provide fused data with increased correctness, conciseness, and completeness compared with the original data. Correctness measures whether the fused data conform to the reality of the object under study. This occurs when more than one data source can confirm the same hypothesis, which increases the confidence of the data. Conciseness refers to the reduction of ambiguity, which means that the fused data from multiple sources have decreased the set of hypotheses about the object of study. Finally, completeness measures the amount of information from the fused data, which increases the robustness because one measurement can contribute to information where other measurements are incapable. To make this process successful, we need to define resolving conflicts from the data outline. The data conflicts can occur when there is uncertainty or when there are contradictions. Uncertainty occurs when there is missing information, such as gene levels not included in the measured platform, or a particular

sample. A contradiction occurs when all the information is presented to its entirety, but the information that we can extract from one source is different from the one that can be extracted from another source.

Data fusion techniques have been applied mainly to the graphical computation context. Numerous data fusion algorithms at different levels of detail [83] have been categorized by the USA JDL [221] into three basic levels according to the amount of information that they provide. The first level corresponds to raw uncorrelated data. The second level -or feature level- provides a higher level of inference, and an additional interpretative meaning is suggested. Finally, the third level –or decision level– delivers additional interpretative meaning. It is designed to provide recommendations to users.

In the high-throughput biomolecular data context, data integration is typically performed in four different manners. One is to analyze each data type isolated first, then integrate the conclusions. Another one is to first pre-process each type of data independently. Second, normalize the data types. Third, integrate the normalized figures and finally perform an overall analysis. A different approach for integration consists of performing a statistical integration (using statistical methods). Lastly, to integrate the data types using a model based on the biological meaning of the data types and their interactions.

For instance, researchers have analyzed mRNA and microRNA paired data by analyzing each data type independently and interpret results manually [71]. Sometimes the results of these experiments can lead to contradictory and unexplained results. A second scenario consists when researchers decide to merge measurement tables, i.e., add the microRNA rows to the mRNA table (rows are the RNA molecules) and analyze and extract conclusions from the new merged table.

During the last two decades, immense progress has been made toward understanding the molecular processes that are different in cancer patients. Traditional approaches compare gene expression levels between samples of cancer patients and healthy individuals; however, recent studies have shown that monitoring only gene expression is not able to capture the

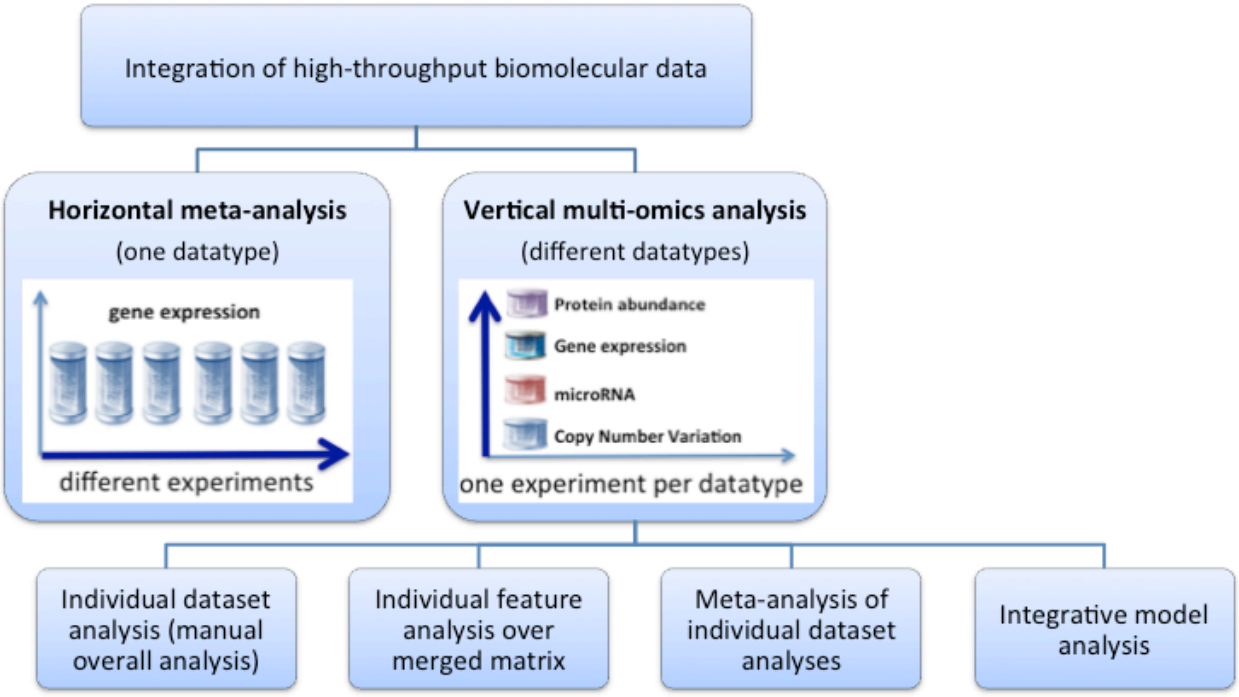


Figure 1.1: Classification of integration methods for high-throughput biomolecular data.

disease signatures in most cancers [246]. Integrating gene expression with other data types has become a new challenge in our age. Integrating gene expression with other data types is the new paradigm for studying complex diseases. Integrative approaches have shown to be successful in finding cohesive perspectives of complex cellular systems. Nevertheless, analyzing multiple data types is extremely difficult due to heterogeneity and high-dimensionality. For example, The Cancer Genome Atlas (TCGA) [325] dataset contains nine data levels (excluding clinical data and imagines) for a total of 28 data types. TCGA is an effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). Life scientists that intend to analyze these datasets by pairs would have to conduct 378 different analyses to compare every possible pair of data types. And the most significant set back is, for a basic experiment as determining if a gene is active or inactive will result in data suggesting different conclusions. To help biologists to understand this complex data flood, informaticians have been developing computational methods that integrate the information that we learned from each isolated component in a systematic approach.

Many publicly available repositories contain a vast amount of high-throughput data, such as Gene Expression Omnibus [18, 106], The Cancer Genome Atlas [325], ArrayExpress [288, 34]. To take advantage of these floods of data, many researchers are trying to integrate data from multiple datasets and multiple measurements of the same set of patients from numerous sources. There are two general directions to integrate data: i) horizontal meta-analysis and ii) vertical multi-omics analysis [332]. A horizontal meta-analysis is also known as cross-cohort data integration. Its purpose is to integrate the same type of data from independent but related studies. A vertical multi-omics analysis integrates multiple types of data measured for the same set of patients. A horizontal meta-analysis (also known as cross-cohort data integration) is used to integrate the same type of data from different sources or different laboratories. Both of these can also incorporate information from biological pathways or other knowledge databases. These studies require interdisciplinary expertise like biology, statistics, and computer science. Many publicly available high-throughput multi-omic data sets create several challenges to store, pre-process, curate, analyze, integrate, and interpret data [260, 209, 30].

This thesis contributes to the field of vertical data integration rather than data fusion because the proposed methods integrate multi-omics and clinical data prior to data analysis. In Chapter 4, we introduced disSuptyper, a method that integrates multi-omics data and biological pathways. Chapter 5 presents CLIGEN, which integrates clinical and mutation data. Lastly, we integrate clinical, mutation, and gene expression data in our TGENEX pipeline presented in Chapter 6.

1.2 Disease subtyping using multiple data types

A vast majority of the diseases develop differently, making them heterogeneous. Identifying similarities and differences among patients to ultimately identify disease sub-types reduces such heterogeneity and help us study diseases [291]. Furthermore, the precise classification of patients into sub-types can help the practice of medicine from diagnosis, treatment, and prevention. Moreover, identifying sub-types that are relevant to survival profiles

or related to biological patterns is crucial as well as identifying more homogeneous disease sub-types and their corresponding genetic signatures. Such sub-type distinction can advance diagnosis classification, which can improve clinical decision and treatment matching. Most methods for disease subtyping perform clustering using either genetic data or clinical data from patients [291, 272, 31, 87, 90, 157, 156]. These methods do not integrate clinical data with molecular measurements, and the outcome sub-types are prone to be suboptimal [314].

A contemporary method that integrates genetic data for disease subtyping is SNF (Similarity Network Fusion) [345]. Figure 1.2 shows the workflow of SNF. The input includes multiple matrices, and each represents the molecular measurement of a data type for the same set of patients. a) SNF first constructs a patient similarity matrix (PSM) for each data type (Figure 1.2b). It then constructs a network for the patients where the nodes are patients, and the edges are the similarity between them (Figure 1.2c). It then iteratively fuses these networks into one network that represents the overall similarity between patients for the multi-omics data (Figure 1.2d). In each iteration, the fused network discards the weak similarities to eliminate contradictions. After each iteration, the networks from multiple data types are more similar to each. The algorithm stops when the networks are identical (Figure 1.2f). Finally, a similarity-based clustering, such as spectral clustering, is performed on the fused network to identify sub-types of the disease.

The authors validated the discovered sub-types using Kaplan-Meier survival curves, Cox regression [69, 326], and Silhouette score. The method is compared with existing methods, such as iCluster [298] and Consensus Clustering [239]. The data analysis was done using five different cancer datasets downloaded from TCGA: glioblastoma multiform data (GBM), breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC), and colon adenocarcinoma (COAD). For all the five datasets and all the metrics used, SNF achieved the best result.

Cox log-rank test is one of the methods to determine if certain groups of patients have different survival dynamics [326, 70]. Survival Analysis, or time-to-event analysis, is used for

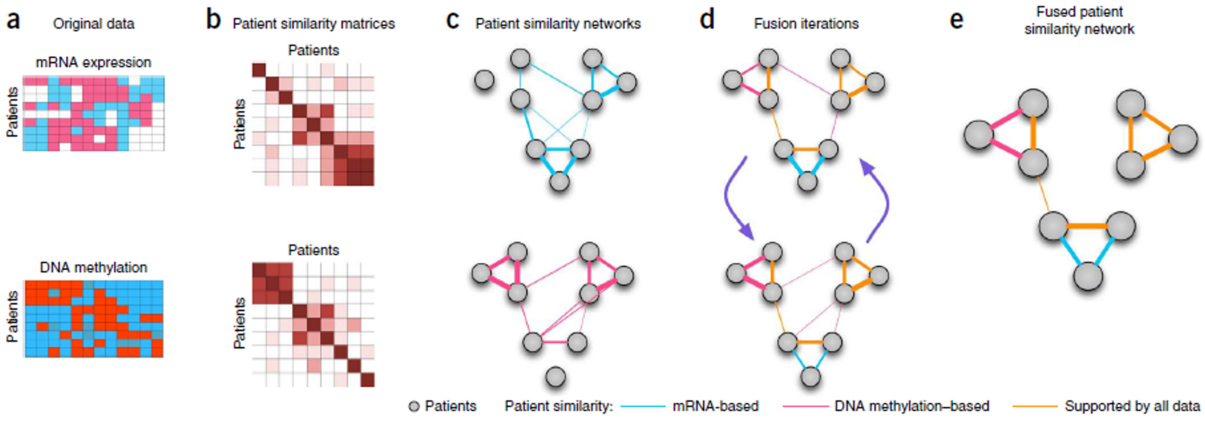


Figure 1.2: Workflow of Similarity network fusion (SNF) [345]. (a) The input consists of multiple matrices that have the same set of patients (rows) but different sets of measurements (columns). Each matrix represents the molecular measurements of a data type. (b) Similarity matrices for each data type. (c) Similarity networks built from similarity matrices. (d) Network fusion by SNF. The algorithm iteratively the networks of the data types. Each iteration makes the networks more similar. (e) The resulting fused network.

examining the expected time it takes for an event of interest, such as death, to occur [214, 26, 136, 26]. The basic setup for this analysis is that certain subjects (e.g., cancer patients) are censored over time until the event happens (e.g., death). Some subjects could get lost from the sample and cannot be censored anymore (e.g., patients drop from the study or die). Survival in this context means how long people stay in the sample. At the beginning of the study, all the subjects are in the sample; therefore, survival is 100%. Over time, events start happening, and survival starts decreasing until the study is over. Typically, the analysis will include a survival curve to visualize the behavior of survival over time. It also considers hazards, which represent the risk of failure or what is the chance that the event will happen before a specific period. In this case, the hazard is the probability of dying at a particular moment. For survival analysis, we are particularly interested in hazard ratio, which is the hazard in one group (e.g., cancer patients) divided by the hazard in another group (e.g., control subjects) [214, 26, 136, 26]. The dependent variable is the duration of measurement which is a combination of three variables the time variable (the length of time until the event happened or being in the study), the event variable (1 if the event happened

or 0 if the event has not happened yet), and the censored variable (indicating if the event was observed or not). These survival studies can have several extensions; one of these is the use of more than one group of participants in the same study. Survival analysis models can be non-parametric models, parametric or semi-parametric models. Non-parametric models are useful for descriptive purposes and to visualize the shape of the survival and hazard functions before using a parametric model. Hazard curves are nonmonotonic and survival curves are strictly non-increasing curves. There are two estimators commonly used for non-parametric models: Nelson-Aalen estimator of the cumulative hazard function and The Kaplan-Meier estimator for the survival function [136]. These models depend on the form of the survival function. The Cox proportional hazards model is a semi-parametric model [326, 70]. The cluster (i.e. sub-type) number that each patient belongs to is incorporated into this model as an independent explanatory variable. Cox's proportional hazards model, is the most common approach used to model survival data, and we use it for our survival analyses. The dependent or response variable for this model is the hazard (risk of death at a given moment) [26]. To compare two survival groups, this model assumes that the risk of death in one group is a constant ratio of the risk of death in the other group (Hazard Ratio). The Cox score test is used to test if the ratio between the groups equals 1. The Log-rank Mantel-Cox test is similar to the score test, and it tests the null hypothesis, H_0, that the survival functions of the groups have no statistically significant difference.

For disease subtyping, most methods using machine learning techniques rely on gene filtering to reduce complexity. For example, the gene expression and methylation data consist of tens of thousands or hundreds of thousand features, which make the clustering or classification problems unscalable using classical techniques. Most techniques reduce the complexity by feature selection, i.e., they focus on the set of genes that are selected beforehand by the life scientists. Similarity Network Fusion (SNF) is the only known technique that deals with molecular data on the whole genomic scale without filtering or gene pre-selection. SNF works on the space of patients instead of the space of genes. In order to do this, they construct

the similarity network between patients for each data type and then simultaneously merges the constructed networks into a single fused network that represents the overall similarity between patients across all data types.

However, SNF has several pitfalls. First, the metric fusion technique used in SNF relies on many parameters. A slight change in these parameters or the input data would significantly alter the output sub-types. Besides, the spectral clustering techniques used to partition the fused network is known to be unstable (a small change in similarity may completely alter the final clusters). Furthermore, this clustering method needs a pre-specified number of clusters.

In this thesis, we present methods for integrating multiple types of data for the purpose of disease subtyping. The first method incorporates biological pathways and integrates multi-omics data [90]. Our second method incorporates mutation data and clinical data of breast cancer patients [84], and the third method integrates gene expression, mutation data, and clinical data of several different types of cancer.

CHAPTER 2 BIOLOGICAL BACKGROUND

This chapter provides a necessary background in molecular biology, which is crucial to present the contributions of this thesis. In particular, we present the central dogma of molecular biology, introduce the function of mRNA, different types of mutation, cancer development, and breast cancer sub-typing.

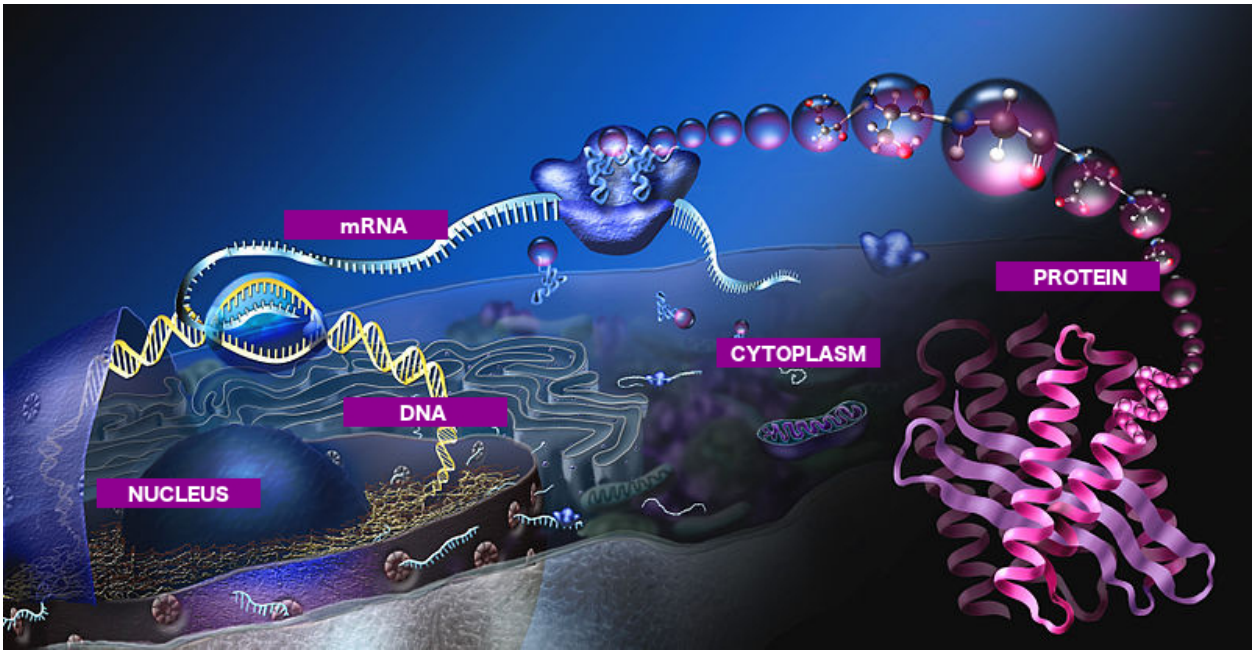
Molecular biology can be defined as the study of life at the molecular level. It is an interdisciplinary approach that combines genetics (i.e., the study of heredity) and biochemistry (i.e., the study of the chemistry of living things). Here, we are interested in the molecular biology of the genes rather than other components of the cell, i.e., the study of genes, how they translate to proteins, and its clinical significance.

Figure 2.1 illustrates a simple representation of the flow of genetic information from genes to proteins. This process is known as *the central dogma of molecular biology* which has two main steps: *translation* and *transcription*. First, a piece of information in the DNA (a gene) is *transcribed* into messenger-RNA (mRNA) in the cell nucleus. Then, the mRNA is transported to the cytoplasm to be *translated* into a polypeptide chain (protein) by the action of a ribosome and multiple transfer-RNAs.

Different technologies have been developed to measure the molecules involved in the flow of genetic information (DNA, mRNA, protein). The following subsections describe these molecules and the techniques used to measure them. The input material of transcription is deoxyribonucleic acid (DNA). The genetic information necessary for cell functioning is carried in the form of DNA, which is made up of *nucleotides*. Each DNA nucleotide contains one of these four bases: adenine (A), cytosine (C), guanine (G), or thymine (T). These bases bind nucleic acids together by complementary pairing – adenine base pairs with thymine and cytosine with guanine. The DNA structure contains two strands of complementary nucleotide chains forming a double helix [353]. Typically, DNA is represented in a linear format as a sequence of nucleotides.

The output of transcription is ribonucleic acid (RNA). DNA is transcribed to messenger-

Figure 2.1: The central dogma of molecular biology.



Credit: Nicolle Rager, National Science Foundation

RNA (mRNA), which is transported out of the nucleus. RNA is a single strand of nucleotides, where each RNA nucleotide contains one of these four nitrogen bases: adenine (A), cytosine (C), guanine (G), or uracil (U). In transcription, each thymine base is copied as a uracil base. Typically, mRNA is described in a linear format as a sequence of nucleotides.

Each triplet of mRNA nucleotides, named *codon*, is translated to an amino acid. Humans have 20 types of amino acids, and each amino acid is mapped from more than one codon. The amino acids translated from an mRNA strand bond together to form proteins, i.e., polypeptide chains. Proteins are involved in almost all functions in a cell.

There are two main categories of genes: protein-coding genes and non-protein-coding genes. Protein-coding genes are transcribed and then translated into protein. Non-protein-coding genes are transcribed but never translated; their final product is non-coding RNA (ncRNA). Gene expression is the process by which a piece of particular gene information (DNA) is transformed into a gene product, i.e., either ncRNA or protein. The essential central dogma model does not include crucial ncRNAs, such as microRNAs (miRNAs).

microRNAs (miRNAs) are small RNA molecules of approximately 22 nucleotides capable of suppressing protein production by binding to gene transcripts. More than 30% of the protein-coding genes in humans are miRNA regulated [213]. Additionally, miRNAs have been shown to play a significant role in diagnosis and prognosis for different types of diseases [211]. Several efforts have been made to identify mRNA-miRNA target interactions, i.e., which miRNAs regulate which genes. Most microRNA-target interactions are statistically predicted, and some are experimentally validated.

Given the importance of miRNAs, hundreds of thousands of miRNA-targeting-genes interactions have been experimentally validated and collected in public databases such as mirTarBase [165], miRWalk 2.0 [105], miRecords [360], and TarBase 7.0 [294]. There are also several algorithms used to predict miRNA targets [178, 213, 203] and databases with predicted interactions such as miRanda [178], TargetScan [213], PicTar [203], and TargetRank [252]. There are also find miRNA-disease interaction databases [165, 176, 218] which are growing rapidly.

2.1 Data integration in cancer studies

Cancer is a disease that involves genetic and environmental factors. Knowledge of the roles that genes play in a particular disease is rapidly helping us to understand cancer biology. These functions differ significantly; for example, some genes can contribute to determining the disease state (disease genes) while other genes can interact with particular environmental factors in causing cancer (susceptibility genes). Identifying the roles that genes play in a disease is not an easy task; it requires rigorous biological experiments followed by statistical and computational analyses to interpret the data. High-throughput technologies allow the monitoring of cellular processes at the molecular level.

One of the molecules typically measured is ribonucleic acid (RNA), particularly messenger RNA (mRNA). The mRNA is used as a proxy to determine *gene expression*, i.e., the process by which a gene synthesizes to a gene product. These measurements are taken to identify if a gene is over-expressed or under-expressed. Using these technologies, conventional data

analysis provides a list of differentially expressed (DE) genes. This analysis is done by comparing the *gene expression* from two groups and statistically identifying the genes that are significantly different between the groups, e.g., one group of healthy individuals versus one group of patients with the disease under study. Lists of DE genes are widely used. However, these lists often fail to elucidate the underlying biological mechanisms.

In the last couple of decades, several approaches have focused on the interactions between genes rather than the study of individual genes. These gene to gene interactions are captured as graphs, named *signaling pathways*, with genes as vertices and interactions as edges. Each signaling pathway describes a cellular process and contains the genes and interactions that are involved in this process. Researchers have been storing the knowledge about various pathways into many publicly available databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [182], Biocarta [28], and Nature Pathway Interaction Database (NCI) [292]. Given the availability of such a collection of pathways, researchers now could identify the pathways that are significantly impacted by a given condition. Identifying pathways instead of genes increases the explanatory power and gives us a better understanding of the underlying biological phenomenon [187, 189, 238]. Many *pathway analysis* methods have been developed to identify enriched or differentially regulated pathways [22, 25, 107, 113, 317]. These methods can be divided into three different categories: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology methods (PT) [101, 274].

The over-representation analysis (ORA) [324] identifies the pathways with differentially expressed genes that are significantly greater than expected by chance. This approach ignores all the gene interactions and assumes gene independence, resulting in an incorrect hypothesis testing thus leading to biased results. Functional class scoring (FCS) methods, such as Gene Set Enrichment Analysis (GSEA) [317] and Gene Set Analysis (GSA) [107], do not assume independence between genes [36, 95]. The hypothesis of FCS methods states that well-coordinated small changes in relevant genes can also have significant effects on pathways besides large changes in individual genes. However, these approaches still do not take into

consideration the interactions between genes as described by the pathways, resulting in information loss, which in turn leads to both false positives, as well as false negatives [101]. Topology-aware approaches, such as Impact Analysis [101, 323], analyze the pathways as graphs and take into consideration the type and direction of each gene-gene interaction.

Pathway analysis methods using gene expression (mRNA) have achieved remarkable results [22, 25, 107, 113, 187, 189, 238, 317]. However, mRNA alone is unable to capture the complete picture of cell processes, as other entities also play important roles. For instance, microRNAs (miRNAs) are newly discovered gene regulators that play a crucial role in diagnosis and prognosis for different types of cancer [211]. miRNAs are small RNA molecules capable of suppressing protein production by binding to gene transcripts. More than 30% of the protein-coding genes in humans are miRNA-regulated [213]. Given all the evidence of the miRNA's relevance, hundreds of thousands of miRNA targeting genes interactions have been experimentally validated and collected in public databases such as mirTarBase [165], miRWalk 2.0 [105], miRecords [360], and TarBase 7.0 [294]. There are also several algorithms used to predict miRNA targets [178, 203, 213] and databases with predicted interactions such as miRanda [178], TargetScan [213], PicTar [203], and TargetRank [252].

Besides, relevant work has been done to elucidate the important interplay between miRNAs and biological pathways [10, 41, 164, 165, 247, 343]. These studies focus on different directions, some methods search for pathways that are targeted by a particular miRNA [10], and others perform pathway analysis using just miRNA expression, such as mirTar [164, 165] and DIANA-miRPath [343]. Other methods incorporate both mRNA and miRNA for pathway analysis [41, 247]. The earliest tool that implements mRNA-miRNA integration is the miRNA and mRNA integrated analysis (MMIA) [247], which performs Gene Set Analysis (GSA) of the down-regulated genes that are targeted by up-regulated miRNAs. However, as mentioned before, GSA does not take advantage of the knowledge captured by the pathway topology. The state-of-the-art approach for the miRNA-mRNA pathway analysis method is microGraphite [41], which uses an empirical gene set approach. microGraphite's primary

goal is the identification of signal transduction paths that are most correlated with the condition under study [231]. Functional analysis methods that include miRNA are still needed to enhance the knowledge of disease gene regulation [74].

The major drawback of current approaches is that most of them do not take into consideration the knowledge about the interactions between the genes, as well as between genes and miRNAs. In this thesis, we present mirIntegrator, a topology-aware approach that systematically integrates miRNA and mRNA expressions to identify pathways that are significantly impacted by the studied condition. Our framework is flexible and allows users to integrate signaling pathway databases with miRNA-mRNA interaction databases to produce *miRNA-augmented pathways*. Here we show that pathway analysis performed on these *augmented pathways* offers more statistical power than performing analysis on gene-gene pathways. Our augmented pathways offer a more comprehensive view and a deeper understanding of complex diseases.

2.2 Current pathway analysis methods

High-throughput molecular biological methods perform thousands of simultaneous measurements of biological molecules to observe a particular state of cells. Recent technologies have extended the breadth of available high-throughput molecular biological data. Nowadays, most of the molecular data types are analyzed separately, which has provided essential discoveries, such as biomarker identification. However, analyzing various data types together can lead to a more consistent understanding of cell processes [124].

The term *high-throughput data* is used here as large measures of genetic data taken in a short time. These data are generated by different technologies commonly referred to as "omics technologies" which are the foundation for systems biology [268]. Omics seek to quantify, describe, and identify all of the components of cellular systems with spatial and temporal dimensions [285]. There are several data types of high-throughput measurements from which four categories are the most important: proteomics, transcriptomics, metabolomics, and genomics [362]. Proteomics is the study of proteins present in cells. Transcriptomics measures

all gene expression values. Metabolomics aims for the quantification and identification of metabolites. Genomics includes the large-scale genotyping of single nucleotide polymorphisms (SNPs). Each of these data types is unique and provide different perspectives on cellular processes, that is why is important to consider these different perspectives when analyzing live systems. There are several computational solutions for analyzing omics data in an isolated fashion [24]. However, single data type analyses have not established enough understanding to perform disease diagnosis and treatment success.

The identification of pathways that are involved in a particular phenotype is typically referred to as *pathway analysis*. Identifying pathways that are relevant to a condition is essential because it provides insights that can be used to further disease treatment or diagnosis. The standard input of pathway analysis techniques is the log-fold change of a large set of genes (around 25,000). Fold change is computed as the ratio of gene expression between two different groups, commonly one group of control subjects and another group with patients. The output of pathway analysis is a ranked list of statistically significant biological pathways. These pathways are considered to be related to the condition under study. Biological pathways are graphical representations of common knowledge about genes and their interaction with biological processes. In particular, signaling pathways are represented as graphs with a set of genes as nodes and the biochemical and physical interactions as edges. These pathways are typically made by mining the literature and then manually curating the retrieved information [182].

Disease sub-typing is an essential goal for omics integration. Generating clinical meaningful disease sub-typing is critical for prognosis and further treatment determination. Based on statistical information and the patient's profile, the objective is to identify the sub-type of disease that the patient more likely belongs to. The input for disease sub-typing is molecular and clinical data from several patients with the same condition but have different outcomes. The expected output is well-identified groups that highly correlate with the observed outcomes (e.g., a group of long-term survival patients and another group of short-term survival

patients). It is also essential to identify possible patterns that are shared among members of each sub-type and differences with other sub-types. This is commonly expressed as a clustering problem where the main goal is to search for similarities among the data points. All these applications highlight the importance of integrating various biomolecular data types. There are more applications of data integration, such as signaling networks reconstruction [77, 130, 197, 241] and biological networks visualization [301].

From the computer science perspective, the term data integration refers to the integration of fragmented information from different physical databases or data warehouses and different representations. Several authors have proposed platforms and languages to integrate databases [3]. Even though data fragmentation is a significant problem, we will not study that type of data integration here. In bioinformatics, the terms data integration and data fusion are synonymous. In computer science, data fusion is referred to as the process of integrating information acquired from various heterogeneous types into a single compound knowledge. Here, we define data integration and data fusion as the integration of knowledge without focusing on the representation. Additionally, data fusion is valuable for acquiring more reliable information than the raw measurements from a single type of source.

In the high-throughput biomolecular data context, data integration is typically performed in four different manners. One is to analyze each data type separately first and then integrate the final findings. Another manner is to pre-process each type of data independently, then perform cross-platform normalization across the data types, then combine the normalized figures and finally perform an overall analysis. The third type of integration consists of performing statistical integration. The fourth approach is to integrate the data by modeling the data types based on the biological meaning of the molecules and their interactions.

For example, researchers have integrated mRNA and microRNA paired data by analyzing each data type independently and then interpreting the results manually [71]. Sometimes the results of these experiments can lead to conflicting and unexplained outcomes. A second scenario is given when researchers having sample-paired data decide to merge the two data

tables into a single table and analyze this new merged table. This practice requires cross-normalization, and it is very dangerous because each data type has different scales, volumes, and properties.

CHAPTER 3 REVIEW OF COMPUTATIONAL METHODS FOR DISEASE SUBTYPING

Cancer development and progression are influenced by several factors, including genetic and overall patient health, which implies that one single treatment plan might not be useful for all cancer patients. Instead, health professionals should be able to design a personalized treatment plan, i.e., a plan for each individual according to their particular genetics, clinical history, and environmental factors. The stratification of patients into groups with similar biology and survival patterns it is a step forward towards personalized treatments and better prognosis. This Chapter examines many computational tools and technologies that have been applied to cancer patient stratification, also known as *disease subtyping*. Some of these tools focus on one of the many types of patient data, while others use a combination of various data types. Here we approach disease subtyping as a computational problem to bridge the gap between oncology and computer science. We also study and compare the different techniques that have been used to tackle this problem, as well as highlight opportunities for new methods. Disease subtyping is an open problem that we could address with the use of the latest machine learning techniques.

Advancements in sequencing technology and wide-spread adoption of Electronic Health Records (EHR) have enabled the large-scale collection of genetic and clinical data. Recent studies focus on analyzing each of these two types of data in isolation for identifying either phenotypes (i.e., observable characteristics that are more prevalent in some individuals with a disease than in the general population [56, 156, 157, 161, 264, 302, 350] or genotypes (i.e., genetic patterns that underlie specific diseases) [110, 163, 192, 207]. Other studies integrate different types of 'omics' data based on the hypothesis that a single data type cannot capture the whole biological system [346].

Several computational methods for analyzing genetic data have been proposed with the goal of identifying genes that play important roles in cancer. Some methods rely on feature selection to reduce the complexity of the problem [295] while other approaches adapt well-

known machine learning methods to the genetic context [110, 192, 346, 163]. Other methods use structured and unstructured Electronic Health Records (EHR) for phenotyping which is the identification of meaningful medical concepts from clinical data [264]. Phenotyping is important for disease subtyping, research subject selection, optimization of interventions, and predicting response to therapy [156, 157, 350]. Current methods for EHR-based phenotyping include rule-based, heuristic, and iterative approaches, which are not fully unsupervised [161, 56, 52, 275]. These methods need annotated data, creating which is time-consuming and requires knowledge from experts, even when phenotyping is performed for a single disease (e.g., rule-based approaches require effort to write decision rules manually). Recently, methods using tensors as a way to represent EHR data from different sources have been proposed [156, 157, 350]. These approaches extract phenotype candidates using tensor factorization methods. They represent the interactions between patients' diagnoses and medications or procedures using a tensor, and then find phenotype candidates after decomposing the tensor. Additionally, these approaches incorporate medical knowledge and constraints [350] and can distinguish the features that are common across patients from the candidate phenotypes [157]. Also, large-scale efforts have been made to automatically define and share phenotypic data, such as the Electronic Medical Records and Genomics Network (eMERGE) consortium [234], and the Observational Medical Outcomes Partnership (OMOP) [259].

Alternative methods combine genomic data and phenotypes to derive a comprehensive model of complex diseases and help to identify important genes and clinical variables [262, 33]. For instance, cancer development and progression are influenced by multiple factors, including germ-line or somatic tumor genetics and environmental or lifestyle risk factors [12]; therefore, clinical variables and genotypes should be considered together when investigating cancer sub-types. Although researchers have highlighted the importance of integrating genetic data into EHR [234, 33, 191], their main focus is on the technical aspects of including genetic reports into EHR systems. To date, limited research focused on developing

methods to analyze patients' clinical and genetic data simultaneously.

In this Chapter, we present a review of computational methods for disease subtyping with a focus on different data and methods that have been used to discover and refine disease sub-types, which can benefit science and practice of medicine [291]. Novel sub-types can drive the design of new biological studies. For example, by finding subgroups whose clinical manifestation differ, researchers can conduct targeted studies to identify molecular determinants of those differences. Such analyses can allow scientists to understand the causes of related diseases. In clinical practice, fine-grained sub-types and prognoses help to reduce uncertainty in individual patient's expected outcome. More accurate prognoses can in turn improve treatment plans. For example, the administration of therapy with substantial side effects could be well justified on an individual prognosticated to decline rapidly without this treatment. In addition, sub-types can improve the effectiveness of clinical trials by increasing the recruitment of viable patients, particularly for cancer trials, in which only a 5% of cancer patients enter clinical trials [119]. Feller et al. [119] found that 25% of cancer trials did not have an adequate number of subjects, and 18% of clinical trials ended with less than half of their goal enrolment after three or more years. The enrolment of clinical trials could be improved with sub-types that can pinpoint patients with a poor short-term prognosis, i.e., patients with a high risk of perishing soon, which might be qualified candidates for aggressive treatments and could be matched to an appropriate clinical trial. Also, the estimation of costs of care could be finetuned if we discover sub-types that can identify patients that are most likely to respond to their treatment plan.

We organized the different papers based on their main contribution, the type of input data, and the computational method applied. Figure 3.1 illustrates the complete framework of disease subtyping methods. We define disease subtyping methods to those that take information for patients with the goal of identifying distinctive groups of patients in terms of survival. Among the methods that we studied, we identified variability in the input data and the methods to be used. We classify the different input data that the studies presented in

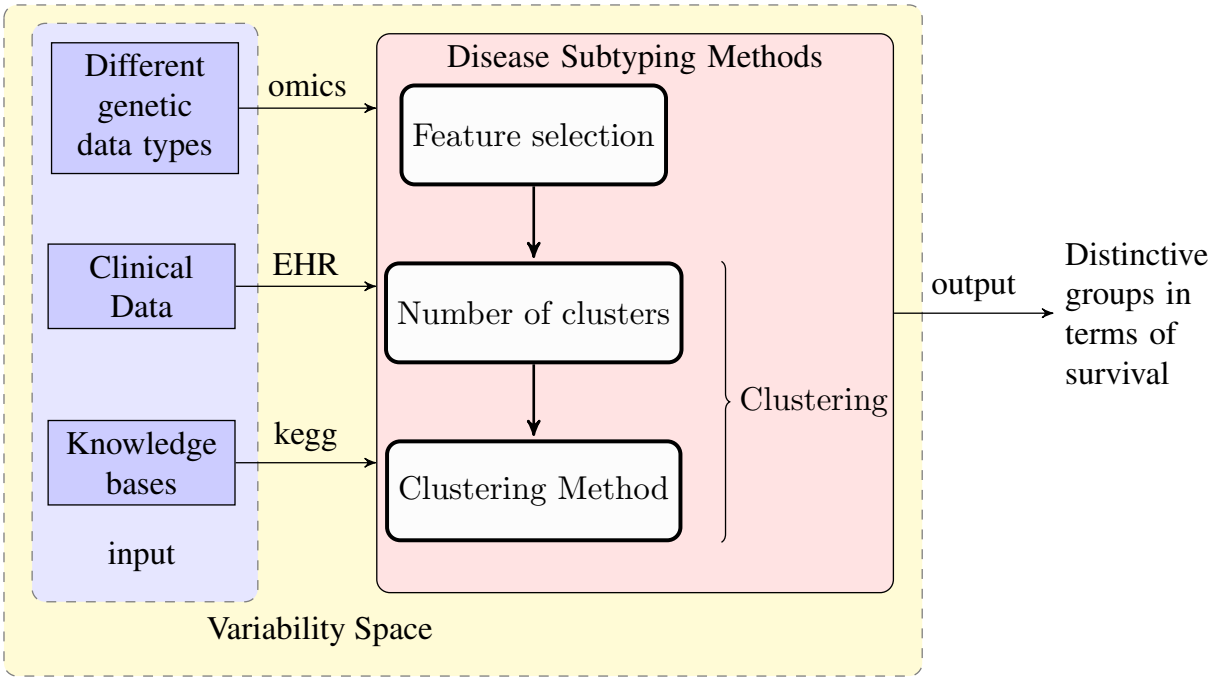


Figure 3.1: General Disease Subtyping Framework

three categories: genetic data types, clinical data, and knowledge bases. Some methods use only one of these while others analyze a combination. The second aspect of the framework that changes from study to study is the method. Some studies focus on finding distinctive features that are related to survival [13, 251, 210, 12, 87, 141, 90, 27, 127], other studies focus on identifying the number of groups of subgroups [103, 38, 37, 295, 302, 329, 318, 237, 205, 102], and others focus on the clustering method [96, 250, 195, 144, 192, 204, 111, 299].

3.1 Research Methodology

For the literature review presented in this Chapter, we collect data through research papers and a systematic literature review protocol for the formulation of research questions.

3.1 Relevant Research Questions

To compare the different computational methods for disease subtyping (DS) and their impact on cancer studies, we addressed the following questions:

RQ1. What types of data are relevant for DS?

RQ2. What type of computational methods has been applied for DS?

RQ3. Which steps on the DS pipeline have been studied?

3.1 Search Strategy

We performed a systematic search of published research available in July 2018. The target population was cancer patients. To identify search terms, we followed this protocol:

1. Derive key terms from research questions.

2. Find synonyms and acronyms for keywords.

3. For each paper, check its relevance and keywords.

3.1 Studies Selection Criteria

Following the protocol described above, we search in five recognized digital libraries (Pubmed, Google Scholar, Science Direct, IEEE Xplore and Springer Link) this query *("subtyping" OR "molecular subtyping" OR "disease subtyping" OR "molecular subtyping" OR "cancer subtyping" OR "patient stratification" OR "cancer stratification")* in July 2018. The number of papers obtained with this search is presented in Table 3.1. We manually reviewed each of these papers and discarded papers that did not present computational methods. This search was limited to English-language and peer-reviewed studies in the field of health informatics and human diseases.

Table 3.1: Query search results

Digital library	**N.o. papers**	**Date**	**URL**
IEEE Explore	158	Jul-2018	ieeexplore.ieee.org
Pubmed	105	Jul-2018	pubmed.gov
Science Direct	95	Jul-2018	sciencedirect.com
Google Scholar	90	Jul-2018	scholar.google.com.pk
Springer link	58	Jul-2018	springerlink.com

Inclusion Criteria

For each of the retrieved documents, we decide its relevance based on the degree to which it addresses any of the three research questions using its abstract section. Only journal papers, full papers from conference proceedings, thesis, and books that present computational techniques for DS, describe theoretical concepts in the context of computational methods for disease subtyping (DS), or describe the need and challenges in DS were included for a full revision.

Exclusion Criteria

Any study that did not present computational methods were excluded, for example, biological papers describing cancer pathways through lab experiments and did not present any automatic way to discover sub-types. Literature reviews and articles without original findings and studies focusing on a single gene and its influence on cancer were also excluded.

3.2 Related work

We divide the selected studies based on the data type they analyzed and the method used. Based on the input data (Section 3.2.1), we classify the DS methods in clinical, genomic, and integrative approaches. Based on the computational method (Section 3.2.2), we organize the DS methods in feature selection and unsupervised learning.

Table 3.2: Latest computational methods for disease subtyping.

Method	Data	Goal	Method characteristics
MLFS [171]	six cancer types: Breast Cancer, Hepatocellular Carcinoma (HCC), Lung Cancer, Prostate Cancer, Colon Cancer and Ovarian Cancer.	Gene/MiRNA Feature Selection	Deep Belief Nets (DBN) and unsupervised active Learning
ML [115]	Gene expression data from various types of cancer	cancer diagnosis and classification (cancer type analysis)	Unsupervised feature learning methods.
DNL [212]	alternative exons profiled from RNA-Seq data	predicting AS patterns	deep neural network
DBN and CD [219]	multi-platform genomic data (e.g., gene expression, miRNA expression, and DNA methylation) for the same set of tumor samples.	Identification of cancer sub-types cluster cancer patients from multi-platform observation data	Multimodal deep belief network (DBN) and Multimodal Deep Learning contrastive divergence (CD)
DNNs [212]	blood biochemistry reports	biomarkers of human cancer	deep neural networks (DNNs)
H2O [336]	simulated datasets and published genome-wide association dataset	detect SNP interactions for disease subtyping	deep feedforward neural network

3.2 Data types

Finding the data and features that are relevant to disease subtyping (DS) is a difficult problem. Some approaches search for sub-types using clinical variables [12, 262, 39] while other approaches use multi 'omics' data [346, 297]. Here we categorize DS methods based on the type of data they use in three categories: i) clinical data, ii) genomic data, and iii) integrative approaches.

Clinical data

Clinical variables used for subtyping include survival data [12], epidemiological data [262], clinical chemistry evaluations, and histopathologic observations [39]. These variables have shown to provide useful information for better subtyping.

Genomic

Several methods use genomic data for which has to be preprocessed and normalization. Current knowledge of cancer biology is key to prioritization and identification of candidate biomarkers. In the same direction, integrating different types of genomics data (e.g., gene expression and copy number mutation) provides crucial insights to identify the genomic alterations that characterize sub-types relevant from both biological and clinical points of view (e.g., HER2 oncogene activation through concordant DNA amplification and mRNA overexpression). Comprehensive cancer-driving mechanisms cannot be entirely captured by any of the genomic data types alone; however, integrative genomic studies can discover novel cancer sub-types and their associated mechanism.

Standard methods for multi-omics subtyping consist of first clustering each of the different genomic data types to analyze and then manually integrating the cluster assignments. These results restrict the ability to discover additional multidimensional interactions and carry a substantial loss of information and cannot identify a correlation between data types [297]. Previous review paper focused on multi-omic clustering methods [297].

One of the challenges of integrating multiple data types is that multi-omics have different scales. Another problem for validating integrative analyses is the lack of availability of independent data sets with all data types available.

Although genome-scale molecular information provides an insight into biological processes driving tumor progression, cancer subtyping based on gene expression profiles alone has been shown to have limited correlation with clinical outcomes [245, 60].

Integrative approaches

Clinical data and biological knowledge are complementary to gene expression and leverage disease subtyping. For instance, some approaches incorporate gene-expression-based subtyping with other types of data, such as clinical variables and multi 'omics' data. These types of data are more and more available nowadays. Large public repositories, including

the Cancer Genome Atlas (TCGA) [249] (now GCD) and the international cancer genome Consortium (ICGC), accumulate clinical and multi 'omics' data from thousands of patients. Despite the immense progress in computational methods for analyzing clinical or genomic data, neither of these two types of methods alone can capture all aspects of the pathogenesis of complex diseases such as cancer [284]. The emergence of EHR-linked biobanks, such as those created by the Electronic Medical Records and Genomics (eMERGE) consortium [234], enabled computational methods to discover associations between specific diseases and genes [33, 206, 104, 190] through genome-wide association studies (GWAS) [153] or between specific phenotypes and genes through phenome-wide association studies (PheWAS) [191, 81].

Since cancer development and progression are influenced by several factors, including germ-line or somatic tumor genetics, overall patient health as well as environmental or lifestyle factors [32], it is natural to assume that cancer sub-types should incorporate all these different modalities of patient data. However, there has been relatively little research on computational methods for joint analysis of clinical and genomic data for disease subtyping. To address this deficiency, CLIGEN has been proposed, a high-throughput pipeline for fully unsupervised disease subtyping based on CLInical and GENomic data [85]. This method finds verotypes (clinical disease sub-types combining phenotypes and genotypes) [31]. BioDCV [262] integrates predictive profiling from gene expression with clinical and epidemiological data by combining machine learning techniques. The modk-prototypes algorithm [39] simultaneously considers microarray gene expression data and classes of known phenotypic variables such as clinical chemistry evaluations and histopathologic observations.

3.2 Methods

Feature Selection

High-throughput technologies can measure more than ten thousand genes at the time. Subtyping patients using whole-genome data is challenging due to the curse of high-dimensionality.

Some approaches have been developed to reduce data dimensionality through feature selection techniques [297, 90].

Traditionally, data is normalized, and features pass initial variance filtering steps, which can result in highly variable clusters due to noise accumulation in estimating centroid in high dimensional feature space, for example, when using the k-means algorithm [297]. This challenge has been addressed by sparse clustering because statistical inference can be more reliable when sparsity is assumed [123, 261, 361, 348, 356].

The selection of class-discriminant features is crucial for model interpretation, accuracy, and computational complexity. For this reason, clustering methods should not be decoupled from a selection of discriminant features. [297]

For example, iCluster [297] is a feature selection based clustering method. Correct feature selection means the identification of class-discriminant features without loss of relevant information or driving factors that define biologically and clinically relevant disease sub-types. Several disease subtyping methods are based on feature selection [154, 168, 272, 216, 149].

Ranking-base methods The simplest way to perform unsupervised feature selection for subtyping is by ranking the list of genes and filtering out those with low rankings. For example, genes can be ranked using Fisher score-based methods [168, 272] or t-test based methods [216].

Filtering metrics Other methods [149] use general purpose filtering metrics like Information Gain [293], Consistency, Chi-Squared [367] and Correlation-Based Feature Selection [142].

Wrapper methods These filter-based methods are computationally efficient, but they do not account for dependency between genes or features. To address this challenge, wrapper methods [92, 296] use learning algorithms to find subsets of related features or genes. Even though these methods consider feature dependency, they have a high degree of computational complexity due to repeated training and testing of predictors. This makes them impractical

for analyzing high-dimensional data.

3.2 Unsupervised Learning

Many DS methods use unsupervised learning approaches, for which here we present methods that study how to select the number of sub-types and methods that propose the use of clustering methods for DS.

Number of sub-types

A challenge of clustering techniques is the identification of the number of clusters. Most studies get the number of clusters based on the judgment of experts instead of using an automatic method to obtain the number of clusters. Several clustering methods require the definition of the number of clusters beforehand, and their performance depends on this parameter. For example, glioblastoma multiform (GBM) has been studied to have two or three in [297], four sub-types [341], and five sub-types in [254]. From [297], the number of reproductible sub-types (K) and model sparsity (number of subtype discriminating features) are determined using resampling-based scheme. The cluster reproducibility index is explained in [300]. In [297], the concept of prediction error that typically applies to classification analysis where the true cluster labels are known now becomes relevant for clustering [103, 328, 186].

Clustering method

Selecting the correct clustering method for disease subtyping has been the main focus of many subtyping studies. Methods such as k-means [297], sparse clustering [297], PCA [297], bayesian consensus clustering [224], and spare PCA [357] have been largely applied.

3.2 Matrix decomposition based methods

Some other studies use Matrix Decomposition techniques, also known as matrix factorization, to identify sub-groups of patients [217, 364, 159, 229, 319]. In this section, we present two matrix decomposition methods that have been used for sub-typing: Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF).

Singular Value Decomposition

One approach to analyzing large volumes of data consists of reducing the dimensionality of the input data. Singular Value Decomposition (SVD) is a data dimensionality reduction technique widely used in machine learning. For SVD, we represent the input data as a matrix A, with m rows (e.g., patients) and n columns (e.g., clinical features), as a product of three matrices: U, σ, and V. Formally, $A \approx U\sigma V^T$, where the U matrix has dimensions $m \times r$ and contains the left singular vectors. σ is a diagonal matrix with dimensions $r \times r$ and has singular values. The V matrix is $n \times r$ and stores the right singular vectors (r is the rank of the matrix A).

In [319], the authors propose a multi-view SVD approach that integrates patient's clinical features with genetic markers for disease subtyping. The method partitions patients into clusters and identifies the genotypes and clinical features that define each sub-type. To validate this approach, the authors used simulated data. The authors proposed a formula for a two-view joint SVD biclustering to find consistent groups across the two matrices (clinical and genetic matrices) [319]. Their rationale is based on applying SVD to each of the matrices and obtained two left and right singular vectors, named u_1, u_2, v_1, and v_2. However, traditional SVD does not guarantee agreement between the two clusterings. The authors claim that to make the clusterings consistent, u_1 and u_2 must have their non-zero components in the same position; therefore, they extended the SVD optimization problem by introducing a common vector between the matrices.

Non-negative Matrix Factorization

Among the matrix factorization methods, Non-negative Matrix Factorization (NMF) has been widely used for genetic data because gene expressions are non-negative; therefore, NMF factors allow for a logical interpretation. For example, [217] proposes an NMF based classification of gene expression data, and [364] uses NMF to classify prostate cancer mutation profiles. Furthermore, NMF automatically clusters the data, a.k.a. has intrinsic clustering

property [94]. Formally, we represent the input data as a matrix A, by the approximation, $A \approx WH$, where the W and H can be obtained by minimizing $||A - WH||_F$, where $W \geq 0$, $H \geq 0$. If $HH^T = I$, then the minimization problem is equivalent to K-means clustering [94].

In [160], the authors cluster cancer patients based on their mutation profiles and known gene networks. They smoothed the binary mutation data by propagating the mutation occurrences through gene-gene interactions and then used NMF to cluster a sample of patients.

3.2 Tensor-based methods

Tensors are generalizations of matrices used to represent data in higher dimensions. Tensors and their factorizations were defined in 1927 [155], but used to analyze real data only recently as the computational power made the required computations possible [305]. Currently, tensor-based methods have been widely used in data science and machine learning [305]. Some studies for disease sub-typing [217, 364, 229] represent the input data as tensors and perform tensor factorization. Tensor factorization methods have shown to be beneficial for genotyping and phenotyping, as presented in [229]. Here, the authors argue and present all the advantages of using tensor factorization for precision medicine, including its facility to integrate various data modalities, reduce the data dimensionality, and identify underlying groups. In the next subsection, we present some examples of three studies that use Non-negative Tensor Factorization (NTF) for phenotyping [157, 156, 350].

Non-negative Tensor Factorization

Ho et al. proposed Limestone [156] and Marble [157] to identify observable trails in patients (phenotyping) from electronic health records (EHR). The advantage of using EHR over genetic data is that they are inexpensive and abundant; however, EHR is noisy, incomplete, and requires manual annotations. The goal is to find the clinical features that are relevant for particular phenotypes, and the authors claim that phenotyping is similar to dimensionality reduction, in the sense that the end goal is to identify relevant features. For this, the authors leverage a tensor factorization on a count tensor proposed by Chi and Kolda [58] for

both the Limestone [156] and Marble [157] methods. CP decomposition factors a tensor in the sum of rank-one tensors, and each one rank tensor is expressed as the outer product of n vectors. For counts, the authors use the generalized k divergence as the objective function to better capture Poisson data (i.e., non-negative and discrete data), with non-negative weights and stochastic constraints.

In Limestone [156], each of the rank one tensors were defined as the candidate phenotypes and the non-zero elements are the clinical variables that define such a phenotype. By post-processing rank one tensors, this method introduced computational stability and inadmissible zero problems. To solve these problems, the authors proposed Marble, which introduces a bias tensor. In Marble [157], phenotypes are defined on the signal tensor, then the baseline characteristics of the population are represented on the bias tensor, which is a special one-rank tensor. To avoid filtering out elements with a value close to zero, they add a sparsity constrain while maintaining non-negative weights and stochastic constraints.

3.2 Topological data analysis

Topological data analysis (TDA) applies three fundamental concepts in topological constructions that make extracting patterns via shape possible: 1) shape representations are coordinate-system-free, 2) shapes have properties that are invariant under small deformations, and 3) shapes can be represented using triangulations. TDA has proven to detect patterns better than other analysis methods [227].

Nicolau et al. [251] used TDA to identify sub-types among breast cancer patients. In [251], the authors use Mapper, which is a TDA method to identify shape characteristics of data sets. They also use Disease-Specific Genomic Analysis (DSGA) for transforming disease omics- data into a sum of two terms, the 'healthy' component of the data and the 'disease' component, which measures the deviation of the data from healthy samples. Then the authors proposed a method to apply Mapper in DSGA data to identify subgroups of patients.

In conclusion, this Chapter describes cancer subtyping and how to address its challenges with machine learning approaches. There had been substantial work done for discovering

molecular disease sub-types, which have helped with the identification of personalized treatments for some Cancers, such as Breast and Lung cancers. For example, the diagnosis of breast cancer is driven by the identification of mutation of a few particular genes, and depending on the mutation profile, each personalized patient's treatment targets its mutated genes. However, there are still patients that do not respond well to treatments, which is an evidence that more fine-grained sub-types are yet to be discovered.

We present a generalized view of disease subtyping (DS) studies in a framework described in Fig. 3.1. We classify DS studies based on the types of data that are analyzed and the methods. We observe that many data types have been used to stratify patients, but most of the analyses use data in isolation, there are still many opportunities for integrating different types of data from the biological domain. Analysis of clinical records remains still a complex problem that can help DS and work in this direction is in high demand. When analyzing the methods proposed so far, we observe a great variety: from purely statistical methods to Bayesian analysis, traditional machine learning and deep learning. Typically, there is a combination of clustering methods [110, 192] with data integration techniques [250, 346] and feature selection [87, 295] to identify factors that are predictive of a specific clinical outcome [12, 66, 278]. Many deep learning methods are currently being used for genetic data (see Table 3.2).

We believe that applying novel techniques to the field of oncology has a great potential to discover fine-grained sub-types that are in such great need for cancer diagnosis, treatment, and prognosis. Further research is also needed in additional oncology related questions that can greatly benefit from computational methods, such as identification of pathogenic mutation in cancer tumors [207], tumor stratification [114, 140, 159, 278], functional diagnostics [122], and tumor classification [287].

CHAPTER 4 A SYSTEMS BIOLOGY APPROACH

One main challenge in modern medicine is the discovery of molecular disease sub-types characterized by relevant clinical differences, such as survival. However, clustering high-dimensional expression data is challenging due to noise and the curse of high-dimensionality. This chapter describes a disease subtyping pipeline that can exploit the critical information available in pathway databases and clinical variables. The pipeline consists of a new feature selection procedure and existing clustering methods. Our procedure partitions a set of patients using the set of genes in each pathway as clustering features. This procedure estimates the relevance of each pathway and fuses relevant pathways to select the best features. We show that our pipeline finds sub-types of patients with more distinctive survival profiles than traditional subtyping methods by analyzing a TCGA colon cancer gene expression dataset. Here we demonstrate that our pipeline improves three different clustering methods: k-means, SNF, and hierarchical clustering.

4.1 Introduction

Identifying homogeneous sub-types in complex diseases is crucial for improving prognosis, treatment, and precision medicine [291]. Disease subtyping approaches have been developed to identify clinically relevant sub-types. High-throughput technologies can measure the expression of more than ten thousand genes at a time. Subtyping patients using the whole-genome scale measurement is challenging due to the curse of high-dimensionality. Several clustering methods have been developed [111, 193, 345, 163] to handle this type of high-dimensional data. Other approaches, such as iCluster [295], rely on feature selection to reduce the complexity of the problem.

There are many widely used feature selection methods [154, 168, 272, 216, 150]. The simplest way to perform unsupervised feature selection for subtyping is by ranking the list of genes and filtering out those with low rankings. For example, genes can be ranked using Fisher score-based methods [168, 272] or t-test-based methods [216]. Other methods, such

as [150], use general purpose filtering metrics like Information Gain [293], Consistency [223], Chi-Squared [367] and Correlation-Based Feature Selection [143]. These filter-based methods are computationally efficient, but they do not account for dependency between genes or features. To address this, wrapper methods [92, 296] use learning algorithms to find subsets of related features or genes. Even though these methods consider feature dependency, they have a high degree of computational complexity due to repeated training and testing of predictors. This makes them impractical for analyzing high-dimensional data.

Meanwhile, some approaches incorporate to gene-expression-based subtyping other types of data such as clinical variables [13, 262, 40] and multi 'omics' data [295, 54, 345]. These types of data are more and more available nowadays. Large public repositories, including the Cancer Genome Atlas (TCGA) (`cancergenome.nih.gov`), accumulate clinical and multi 'omics' data from thousands of patients. Clinical variables used for subtyping include survival data [13], epidemiological data [262], clinical chemistry evaluations and histopathologic observations [40]. These variables have shown to provide useful information for better subtyping.

Subtyping patients using gene expression data has additional challenges because genes do not function independently. They function in synchrony to carry on complex biological processes. Knowledge of these processes is usually accumulated in biological pathway databases, such as KEGG [182] and Reactome [72]. Biological pathways are graphical representations of common knowledge about genes and their interactions on biological processes. This valuable information has been used to cluster related genes using gene expression [144, 166, 276, 270] and should be used to identify disease sub-types as well. Clinical data and biological knowledge are complementary to gene expression and can leverage disease subtyping.

Here we present a disease subtyping pipeline that includes a new feature selection approach and any existing unsupervised clustering method. To the best of our knowledge, this is the first approach that integrates pathway knowledge and clinical data with gene expression for disease subtyping. Our framework is validated using gene expression and

clinical data downloaded from The Cancer Genome Atlas (TCGA) and pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG). Using the features selected with our approach and three different clustering methods (k-means, SNF, and hierarchical clustering), our pipeline can identify sub-types that have significantly different survival profiles. This pipeline was developed in R programming language. The source code is available on GitHub (`http://datad.github.io/disSuptyper`) to ease the reproducibility of the methods presented here [271, 50].

4.2 Background

This section is structured as follows. First, we briefly introduce typical comparative analysis and the importance of pathway analysis using gene expression. Second, we describe the existing knowledge-based pathway analysis methods. Third, we explain the need for multi-omics data integration to better identifying the impacted pathways for better understanding of biological mechanisms of the underlying diseases or phenotypes. Finally, we summarize the main strategies used to integrate multiple data types for pathway analysis.

High-throughput technologies for gene and protein profiling, such as DNA microarray or RNA-Seq, have transformed biomedical research by allowing for comprehensive monitoring of biological processes. A typical data analysis often yields a set of genes that are differentially expressed (DE) when comparing patients versus healthy samples. The lists of DE genes helps to identify genes that take part in the underlying phenomenon. However, there are two drawbacks. First, they often fail to reveal the underlying mechanisms [331, 189]. Second, independent experiments of the disease often yield to entirely different lists of DE genes, which makes the interpretation extremely difficult [321, 108, 109].

To address these challenges, researchers have developed a large number of knowledge bases. Biological processes, in which genes interact with each other, are accumulated in pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [182, 255], or Biocarta [28]. Pathway analysis [238, 189, 113, 201, 167] were developed to infer correlation of DE genes with the known biological processes accumulated in these databases.

The three main strategies of pathway analysis using gene/protein expression data. The input of these methods consists of two parts: i) the molecular measurements using a high-throughput technology, and ii) functional annotations (pathway database) of the corresponding genome. The first approach is the Over-Representation Analysis (ORA). From the molecular measurements of the two groups of patients to be compared, the ORA approach first calculates the list of differentially expressed genes (DE genes) between the two conditions. It then assesses whether the number of DE genes in a given pathway is likely to occur by chance. The second approach is Functional Class Scoring (FCS). It first computes a gene-level statistic for each gene. It then aggregates the gene-level statistics for all genes in a pathway to get a single pathway-level statistic. This summary statistic is then used to calculate the statistical significance of the pathway using permutation or resampling. The third approach consists of the Pathway Topology (PT) based methods. The PT approach calculates a pathway-level statistic that summarizes the expression changes of the genes in the pathway, the known interactions between genes, and the topological order of the genes in a pathway. This summary statistic is then used to calculate the statistical significance of the pathway using bootstrap or resampling. The result of each pathway analysis method is a list of pathways order by the corresponding p-values (see Figure 4.1).

The three different strategies used for pathway analysis are shown in Figure 4.1. For all methods, the input consists of gene/protein expression data and a pathway database. The gene expression data is often represented as a matrix where the columns represent the samples, and the rows represent the components of the samples. For example, a DNA microarray assay [98, 126] of 20 diabetes patients and ten healthy patients are represented as a matrix of 30 columns and about $20,000$ rows. Each column represents a patient, while each row represents the expression of a gene across all patients. The second input, the pathway database, is a list of known functional modules. A functional module can simply be a set of genes [236, 235, 17, 42, 67, 7] that are known to be involved in a biological process, or can be a complicated graph where the nodes represent genes and the edges represent interactions

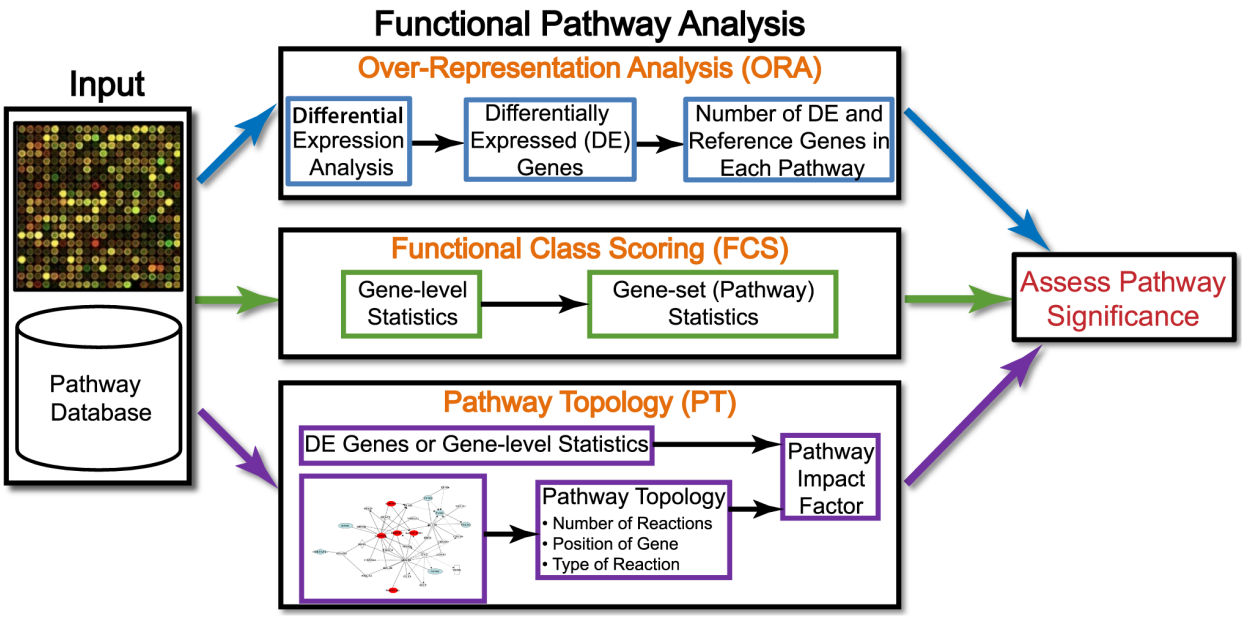


Figure 4.1: Classification of pathway analysis methods. Figure taken from [189].

between genes [184, 182, 255, 72, 232, 180].

The earliest pathway analysis methods use the Over-Representation Analysis (ORA) [230, 22, 25, 100, 97, 188] to identify the gene sets that have more differentially expressed genes than expected by chance. This approach starts by identifying genes that are differentially expressed between the two phenotypes, e.g., disease versus control. Statistical methods for identifying DE genes include t-test [139, 267], regularized t-test [147, 14], and linear models [310]. It then transforms the pathway analysis problem to the classical hypergeometric problem, in which DE genes are the red balls, and non-DE genes are the black balls. For a given gene set and the number of DE genes in the gene set, the ORA approach calculates the probability of obtaining the same number of DE genes or more, using hypergeometric or Fisher's exact test [117].

The ORA approach has obtained remarkable results and gained widespread usage. However, ORA has several limitations. First, this approach only takes into consideration the number of DE genes and completely ignores the change in expression; i.e., it ignores the actual expression measured. However, gene expression or fold change can be useful in assigning different weights to the DE genes. Second, ORA typically uses the most significant

genes and completely ignore other genes. For example, with a p-value cutoff of 0.01, i.e., 1%, only genes passing that threshold are considered significant. Genes that are marginally less significant, e.g., p-values = 0.0099, are not considered, resulting in information loss. Finally, ORA assumes that the difference in expression of a gene is independent of the other genes. However, this assumption is invalid since biological systems are complex with interaction between genes and their products. This assumption ignores the structural correlation between genes, resulting in incorrect hypothesis testing and thus leads to biased results.

The second class of methods in pathway analysis is Functional Class Scoring (FCS). Methods in this class include Gene Set Enrichment Analysis (GSEA) [317, 240], Gene Set Analysis (GSA) [107], sigPathway [327], Category [177], SAFE [20], GlobalTest [137], PCOT2 [200], SAM-GS [95], Catmap [36], FunCluster [148], and PADOG [322]. This approach hypothesizes that not only large changes in individual genes can have significant effects, but well-coordinated small changes in functionally related genes can also have significant effects on pathways. FCS methods mainly consist of three steps. First, they calculate the gene-level statistics, i.e., differential expression of individual genes between two phenotypes. Examples include correlation [266], Q-statistic [137], t-test [4], or Z-score [196]. Second, they aggregate the gene-level statistics into pathway-level statistics, one for each pathway. Existing pathway-level statistics include Kolmogorov-Smirnov (used in GSEA) [317, 240], sum, mean, or median of gene statistics for all genes in the pathway (used in Category) [177], or the max-mean statistic (used in GSA) [107].

The strategy used in FCS methods offers a significant improvement over ORA methods. However, it also has several limitations. First, although FCS methods do not assume the independence between genes, they still assume the independence between pathways. However, this is not true because a gene can function in more than one pathway. Therefore, FCS methods fail to address the crosstalk between pathways and thus lead to biased analysis and an increase in false positives. Second, they do not take into consideration the interaction between genes. For example, consider a gene that is known to interact with many other

genes in a pathway. A significant change in the expression of this gene would result in a significant perturbation in the pathway. This gene should be weighted much more than a gene that is known not to interact with any other genes.

The third class of pathway analysis methods is pathway topology-based approaches (PT) [274, 101, 323, 323, 303, 134, 174, 359, 173, 133]. Methods in this class include ScorePAGE [274], Impact Analysis [101], SPIA [323], NetGSA [303], TopoGSA [134], DE-Graph [174], MetPA [359], BPA [173], and EnrichNet [133]. These methods take advantage of the interaction between genes/proteins provided in pathway databases. Typical PT-based methods, such as Impact Analysis [101] and SPIA [323], model each pathway as a directed graph, where the nodes are genes or gene products, and the edges are the known interactions between the nodes. These methods perform two statistical tests. The first test focuses on the differential expression of genes falling on the given pathway. The p-value of this first test can be obtained from the ORA or FCS methods described above. The second test focuses on the number of perturbation factors accumulated on the given pathway. This test is concerned with the topological position, magnitude, and sign of changes in expression for genes falling the given pathway. The null distribution of the pathway perturbation is obtained by permuting the genes at different locations in the pathway graph. The two p-values obtained from the two independent tests are then combined using Fisher's method.

Although pathway analysis using gene expression has achieved excellent results, recent research has proven that integrating different types of data offers a more comprehensive view of complex cellular systems [246, 91], resulted in a wave of methods for data integration. We divide these methods into two categories: topology-aware methods and non-topology aware methods. Topology aware methods are the methods that incorporate gene topology and interactions into the analysis (i.e., methods that make use of nodes and edges of the pathways). Non-topology aware methods are methods that treat a pathway as a set of genes without considering their topology or interactions. Figure 4.2 shows the overall pipeline of integrative pathway analysis methods. The input includes a set of signaling pathways and

experimental data from multiple data types coming from the same set of patients. Integrative methods output a list of pathways ranked by statistical significance, i.e. p-value or score.

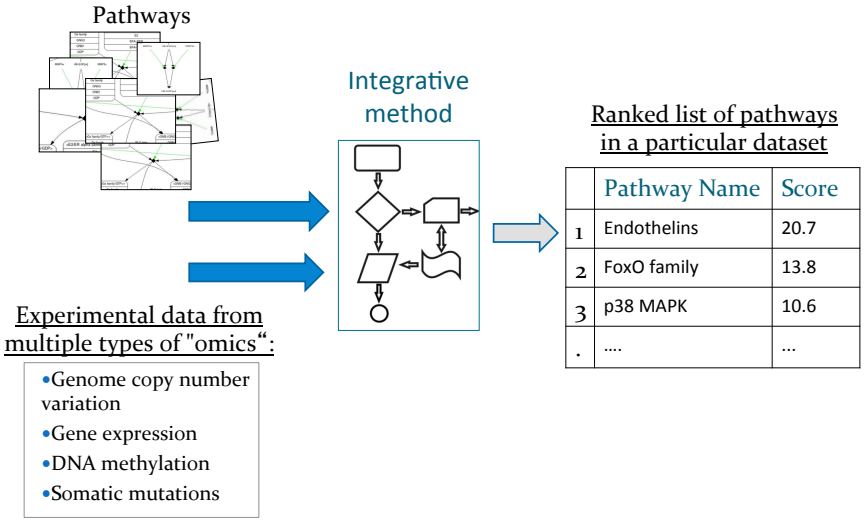


Figure 4.2: A general overview of multi-omics pathway topology techniques. The input of these techniques includes, i) different types of molecular measurements for the same set of patients, and ii) pathway knowledge from the databases. The output is a list of pathways ranked according to their statistical significance, e.g., p-values or scores.

Topology aware methods are based on the hypothesis that incorporating the structure of biological processes on the analysis will provide better results. We reviewed several methods of this nature, and we identified two main categories. Methods belonging to the first category extend the existing signaling pathways with molecules or nodes that were not included in the original pathways. Methods belonging to the second category transform pathways to probabilistic graphical models and include new relations among multiple types of data.

Methods in the first category base their algorithms in traditional statistical tests, which have been used, evaluated, and accepted by the scientific community for decades; therefore, they can be rapidly implemented in research pipelines. The main disadvantage of this approach type is that there are very few data types that can be mapped directly to gene interactions. Given that current signaling pathways databases contain information about gene interactions and ignore remaining data types, enhancing them is imperative [273, 263, 91, 89].

An example of integrating mRNA with microRNA is microGraphite [41]. The small microRNAs molecules interact with genes as gene regulators, and recent studies have shown that these molecules play essential roles in the development of cancers and many complex diseases [211, 226]. In the pipeline of microGraphite, the pathway of genes is extended to the pathway of genes and microRNA molecules. This study [41] defines a pipeline to integrate microRNA and mRNA expressions by wiring the microRNA - mRNA interactions into the formal pathway representations. After expanding the pathway, microGraphite performs pathway analysis using the existing method named CliPPER [231] (see Figure 4.3). The pipeline has been applied to ovarian cancer data, obtaining successful results.

The pipeline of microGraphite consists of five steps, as shown in Figure 4.3. In the first step, microGraphite wires microRNAs to existing pathways downloaded from pathway databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [182, 255], Nature Pathway Interaction Database (NCI) [292], Reactome [180], or Biocarta [28]. There are two types of microRNA target interactions that are wired to the pathways: in-silico predicted interactions and validated interactions. Validated interactions are obtained from miRecords [360] and mirTarbase [165]. In the second step, microGraphite performs pathway analysis to obtain a set of significant initial pathways. In the third step, it carries an analysis across the significant pathways to score the coherent paths inside the pathways. In the fourth step, microGraphite selects the paths with the highest score and then join these paths to form a connected network called meta-pathway. Finally, microGraphite performs pathway analysis among the paths to identify the most significant paths.

Methods in the second category use graphical models, such as Bayesian network or factor graphs, to represent the interaction between data types and gene expression. These models are versatile because they can describe the complex type of interactions. These methods rely on the fact that each type of genomic data contains valuable information, so integrating them in a unique figure makes the analysis more complete.

An example of these approaches is PARADIGM [339]. This method integrates and an-

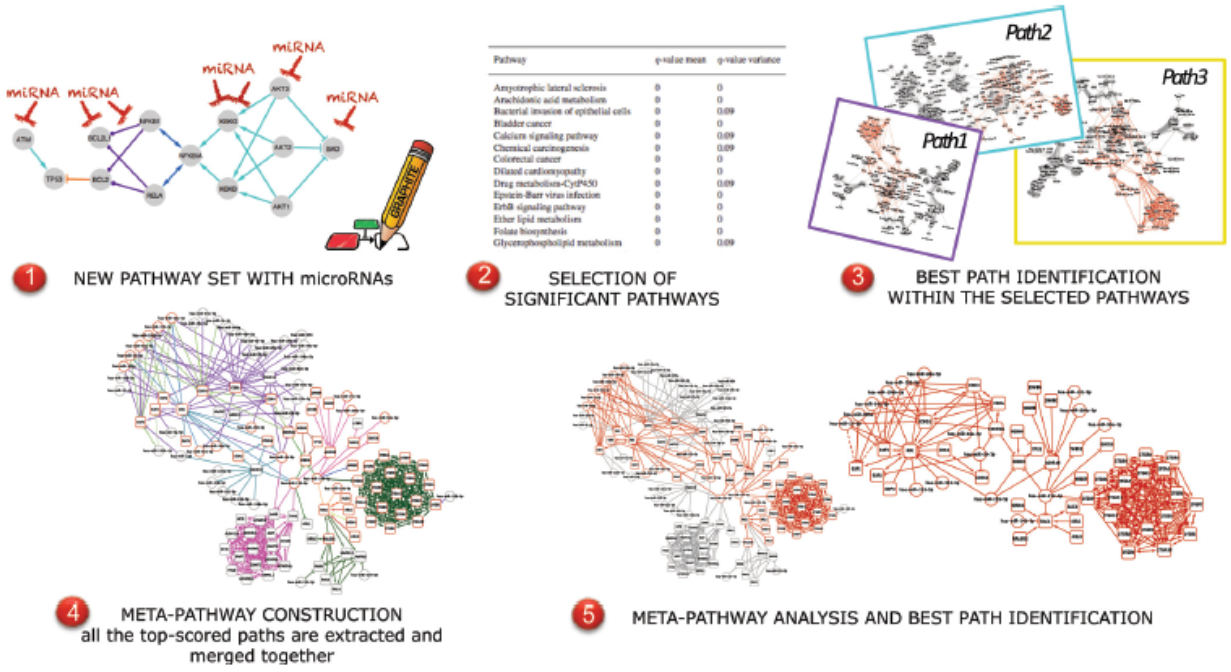


Figure 4.3: Outline of the computational approach described on [41]. (1) Signaling pathways are extended with microRNAs. (2) Pathway analysis is performed, and the significant preliminary pathways are obtained. (3) Analysis across the significant pathways to score the coherent paths. (4) Selection of the paths with the highest score to further join then in connected networks called meta-pathways. (5) Perform pathway analysis among the meta-pathways to identify the significant meta-pathway.

alyzes different types of genomic data by producing a single measurement called Inferred Pathway Activities (IPA). Obtaining a single measurement per patient is innovative because this measurement can be used as a complete signature, simplifying the disease diagnosis. Also, the IPA per patient allows us to perform pathway analysis for an individual while current approaches need a group of samples (several patients) for the comparison. In order to compute the IPA, PARADIGM connects the different types of measurements by adding causal-effect relations and the interaction between genes in a factor graph model. Then, the likelihood of having a gene activated or not in each particular cancer patient and the IPA per gene is computed by performing a Bayesian inference algorithm. The method was evaluated by performing pathway analysis in two different diseases, breast cancer and glioblastoma multiform (GBM), and comparing the results with those obtained by using SPIA. The authors

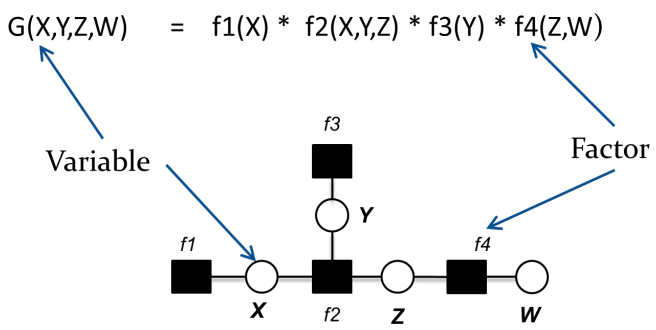


Figure 4.4: An example of a factor graph. This factor graph represents a global function G as the product of the local functions f1, f2, f3, and f4. Black squares of the graph represent local functions or factors, and circles represent variables. Each factor is a function of its neighbor variables.

concluded that PARADIGM analysis generates fewer false positives than other methods, and they were able to identify different groups of GBM with significantly different survival profiles. This method has been included as an official tool into The Cancer Genome Atlas (TCGA) [325]. For this reason, we considered PARADIGM to be the state of the art tool to integrate high-throughput data for pathway analysis.

4.3 Method

In this section, we introduce a new feature selection framework for disease subtyping. Figure 4.5 presents the overall pipeline of our framework. The input includes *i)* gene expression data, *ii)* survival data, and *iii)* biological pathways (see Figure 4.5a). The output is a set of selected genes (Figure 4.5f) for finding sub-types with significantly distinct survival patterns (Figure 4.5g).

Gene expression data can be represented as a matrix $D \in R^{M \times N}$, where the rows are different patients having the same disease, and columns are different features (i.e., genes). M is the number of patients, and N is the number of genes. For gene expression data, N can be as large as $20,000$. The survival data include the patient's vital status (dead or alive) and follow-up information (time and censored/uncensored). The biological pathways are collected from public pathway databases. In this work, our data analysis is based on KEGG

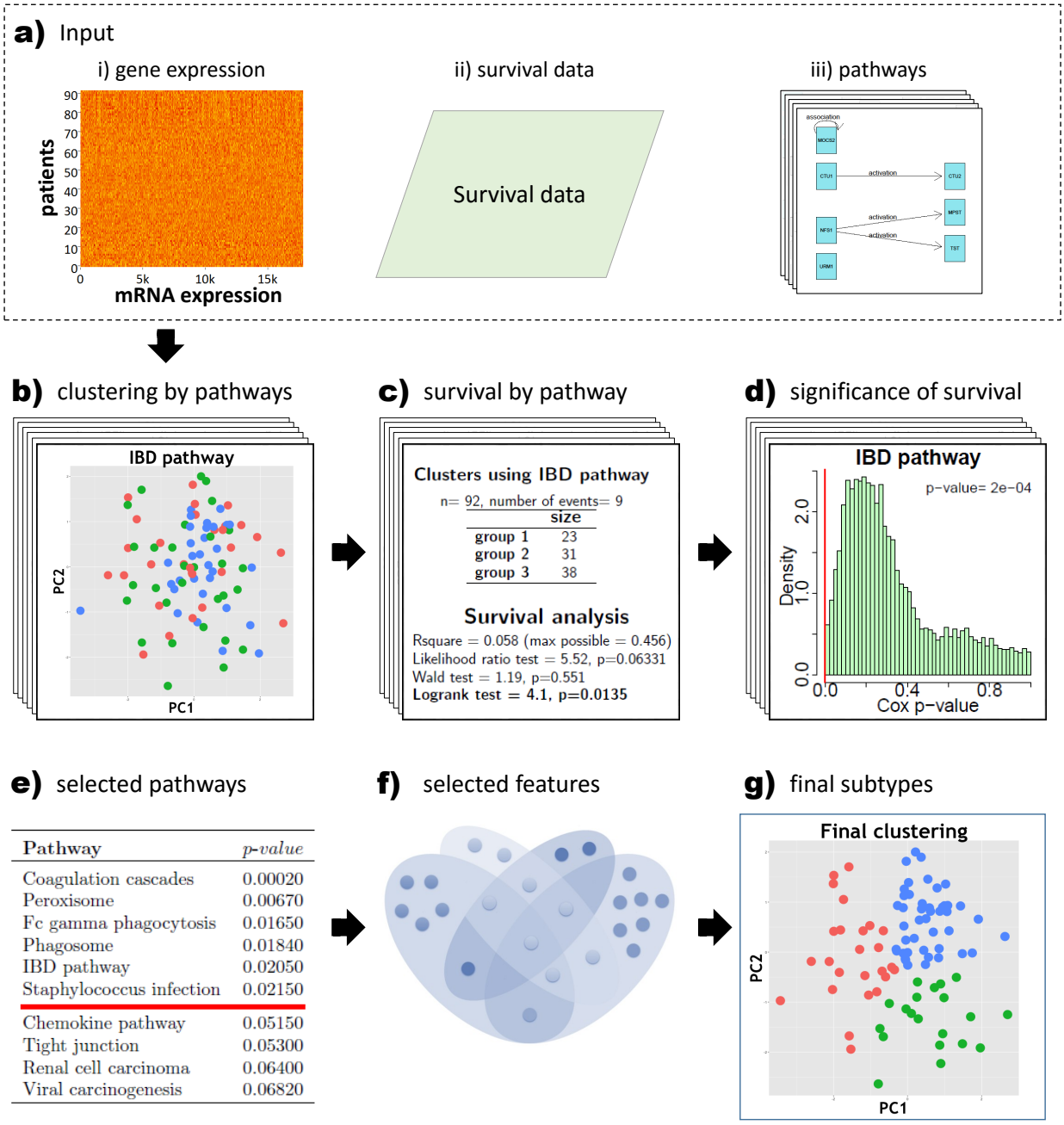


Figure 4.5: Proposed feature selection pipeline for disease subtyping using biological knowledge. (a) The input includes i) gene expression data, ii) survival data, and iii) pathways downloaded from a database. (b) First, we partition the gene expression data using the set of genes in each pathway as features. (c) Second, we perform survival analysis on each resulting partition. (d) Third, we compute the p-value that represents how likely the pathway improves the subtyping. (e) Fourth, we rank the list of pathways by corrected p-value and select pathways that have a nominal p-value less than or equal to the significance threshold of 5%. (f) Fifth, we merge the relevant pathways to construct the final set of features. (g) Finally, we sub-type the patients using the selected features. The clustering is demonstrated in the first two principal components, but we use all dimensions/genes for clustering. Note: IBD pathway stands for the Inflammatory Bowel Disease pathway.

pathways [182], but other databases can also be used.

First, we partition the rows (patients) of gene expression matrix D using the features provided by each pathway in the pathway database (Figure 4.5b). Formally, let us denote $\mathbf{P}$ as the pathway database, which has $n = |\mathbf{P}|$ signaling pathways. We have $\mathbf{P} = \{P_i\}$ where $i \in [1..n]$. For each pathway P_i, we cluster the rows using genes that belong to the pathway P_i as features resulting in a partitioning C_i.

Second, we perform survival analysis on each of the pathway-based clusterings C_i (Figure 4.5c). We calculate the Cox log-rank p-value for the sub-types defined by C_i using the input survival information. This Cox p-value represents how likely the survival curves' difference is observed by chance. So far, we have n Cox p-values, one per pathway.

Now the question is whether the features provided by the pathway P_i help to differentiate the sub-types better. We will answer this question by using a random sampling technique. Denote $|P_i|$ as the number of genes in the pathway P_i. We randomly select $|P_i|$ genes from the original set of N genes. We partition the patients using this randomly selected set of genes and then compute the Cox p-value. We repeat this random selection $10,000$ times which results in a distribution that has $10,000$ Cox p-values (Figure 4.5d). This distribution represents the distribution of Cox p-values when randomly selecting $|P_i|$ features for subtyping. In Figure 4.5d, the vertical red line shows the real Cox p-value calculated from the actual genes in P_i, whereas the green distribution shows the $10,000$ random Cox p-values. Now we compare the Cox p-value obtained from the pathway P_i with the distribution of randomly selected genes. We estimate the probability of obtaining this Cox p-value (using genes in P_i) by computing the ratio of the area to the left of this Cox p-value divided by the total area of the distribution. We denote this probability as p_i. In total, we have n values $\{p_i, i \in [1..n]\}$, one for each pathway. Each of these p-values p_i quantifies how likely it is to observe by chance a Cox log-rank statistic as extreme or more than the one observed. Therefore, this p-value of a pathway P_i represents how likely the features provided by the pathway help to improve the subtyping.

Table 4.1: List of pathways selected by our approach when using RSS k-means. We first ranked the pathways by FDR adjusted p-value ($p\text{-}value.fdr$), then selected the pathways with a nominal $p\text{-}value \leq 0.05$ as relevant pathways.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
Complement and coagulation cascades	0.00020	0.03680
AGE-RAGE signaling pathway in diabetic complications	0.00420	0.38640
Peroxisome	0.00670	0.41093
Cytokine-cytokine receptor interaction	0.01040	0.45448
Fc gamma R-mediated phagocytosis	0.01650	0.45448
Phagosome	0.01840	0.45448
Inflammatory bowel disease (IBD)	0.02050	0.45448
Staphylococcus aureus infection	0.02150	0.45448
Leukocyte transendothelial migration	0.02330	0.45448
NF-kappa B signaling pathway	0.03710	0.50048
Renin secretion	0.03850	0.50048
Malaria	0.04780	0.51326
Platelet activation	0.06980	0.54970

The third step is to choose a set of pathways that certainly help to improve the subtyping. To do this, we adjusted the p-values for multiple comparisons using False Discovery Rate (FDR), we rank the set of pathways and select those that have the corresponding nominal *p-values* less than or equal to the significance threshold of 5%. Let us name the pathways yielding significantly distinct survival curves as *relevant pathways*. For example, In Figure 4.5e, the horizontal red line shows the significance threshold of 5%. In this example, the relevant pathways are *Coagulation cascades*, *Peroxisome*, *Fc gamma phagocytosis*, *Phagosome*, *Inflammatory Bowel Disease (IBD) pathway*, and *Staphylococcus infection*.

Considering all the genes in the relevant pathways as favorable features, we merge these pathways to get a single set of genes (Figure 4.5f). We use this merged set of genes as the selected features for our final subtyping. In our example, the final selected genes are the genes in the six pathways listed above. We then use these genes to construct the final clustering, as shown in Figure 4.5g.

We note that this feature selection procedure can be used in conjunction with any clus-

tering method. In our experimental studies, we used three clustering methods that belong to different clustering models. The first method is the classical k-means. It is well-known that k-means does not always converge to a global optimal point; it depends on the initialization. To overcome this problem, we ran k-means several times and chose the partitioning that has the smallest residual sum of squares (RSS). In the rest of the manuscript, we refer to this as "RSS k-means". The second method is Similarity Network Fusion (SNF) [345], which is based on spectral clustering. The third one is the traditional hierarchical clustering using cosine similarity as the distance function. We will show that our framework helps to improve the subtyping using any of the three mentioned clustering methods.

4.4 Results

In this section, we assess the performance of our feature selection for disease subtyping framework using gene expression data (Agilent G4502A-07 platform level 3) generated by the Cancer Genome Atlas (TCGA) (`cancergenome.nih.gov`). We selected the samples that have miRNA and methylation measurements as were selected in SNF [345]. A copy of the dataset is available in the GitHub repository (`http://datad.github.io/disSuptyper`). The number of patients is $M = 92$, and the number of genes is $N = 17,814$. For all the performed clusterings, we set the number of clusters as $k = 3$ according to prior knowledge of the number of sub-types of colon cancer [345]. When running our method, we used 184 pathways from the KEGG pathway database [182].

As described in Section 4.3, our framework can be used in conjunction with any unsupervised clustering algorithm. Here we test it using three clustering methods: RSS k-means, SNF [345], hierarchical clustering [111]. For all clustering methods, we first clustered the patients using all the measured genes, then clustered the patients using only the genes selected by our technique. To contrast the difference between the three traditional clustering methods and our pipeline results, we performed survival analysis for all the cases using Kaplan-Meier analysis and Cox p-value.

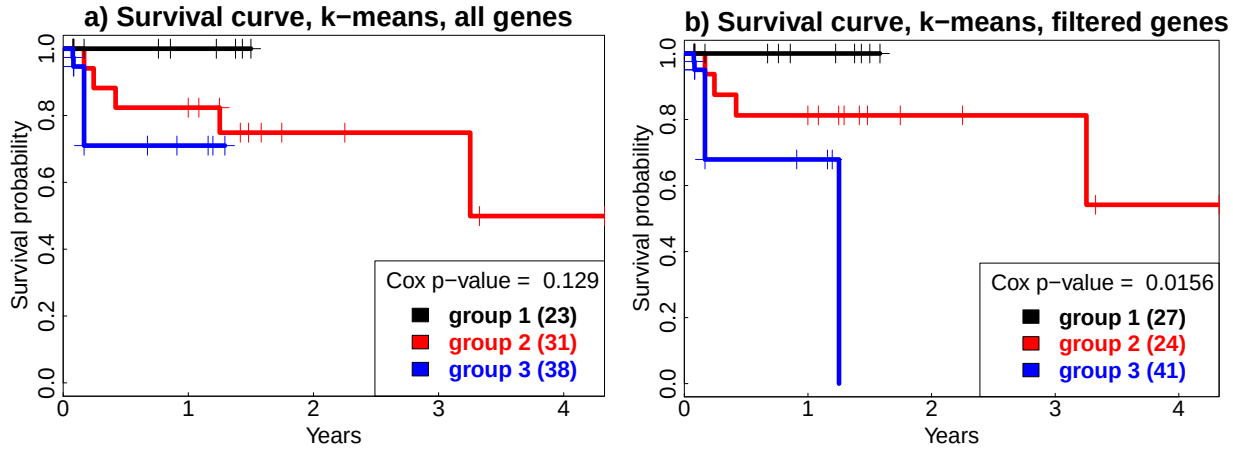


Figure 4.6: Kaplan-Meier survival analysis of the obtained sub-types using the RSS k-means algorithm. a) Survival curves using all genes. b) Survival curves using selected genes.

4.4 Subtyping using k-means

We clustered the patients from the TCGA colon adenocarcinoma dataset using our pipeline in conjunction with RSS k-means. We used the 184 signaling pathways from the KEGG database [182]. For each pathway P_i, we partitioned the patients using the genes in the pathway P_i as features to get a clustering C_i.

After this step, we got a total of 184 clusterings, one per pathway. Also, for each pathway, we constructed the empirical distribution and then estimated the *p-value* of how likely the pathway helps to improve disease subtyping. The *p-values* of relevant pathways are shown in Table 4.1. The horizontal red line represents the significance of cutoff at 5%. There are 12 relevant pathways. We then merged the relevant pathways to get a single set of genes that we used as clustering features. Finally, we performed RSS k-means clustering of patients using these 851 genes. Figure 4.6 shows the survival analysis of the resultant clusterings. Figure 4.6a shows the resultant clustering when using RSS k-means for all $17,814$ genes. The Cox p-value of this clustering is 0.129, which is not significant. Figure 4.6b shows the resultant clustering using the 851 selected genes. The resultant Cox p-value is 0.0156, which is approximately ten times lower than using all genes.

Table 4.2: List of pathways that contain relevant genes obtained with our approach when using SNF. These are the results of the third step of our pipeline, the selection of the relevant pathways. We first ranked the pathways by $p\text{-}value.fdr$, then selected the pathways with a nominal $p\text{-}value \leq 0.05$.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
HTLV-I infection	0.00400	0.37765
Endocrine and other factor-regulated calcium reabsorption	0.00680	0.37765
Complement and coagulation cascades	0.00800	0.37765
Aldosterone-regulated sodium reabsorption	0.00830	0.37765
AMPK signaling pathway	0.01410	0.51324
Phagosome	0.02150	0.54196
Fc epsilon RI signaling pathway	0.02290	0.54196
Cytosolic DNA-sensing pathway	0.02680	0.54196
Peroxisome	0.03900	0.61320
Leishmaniasis	0.04300	0.61320
Non-alcoholic fatty liver disease (NAFLD)	0.05400	0.66544

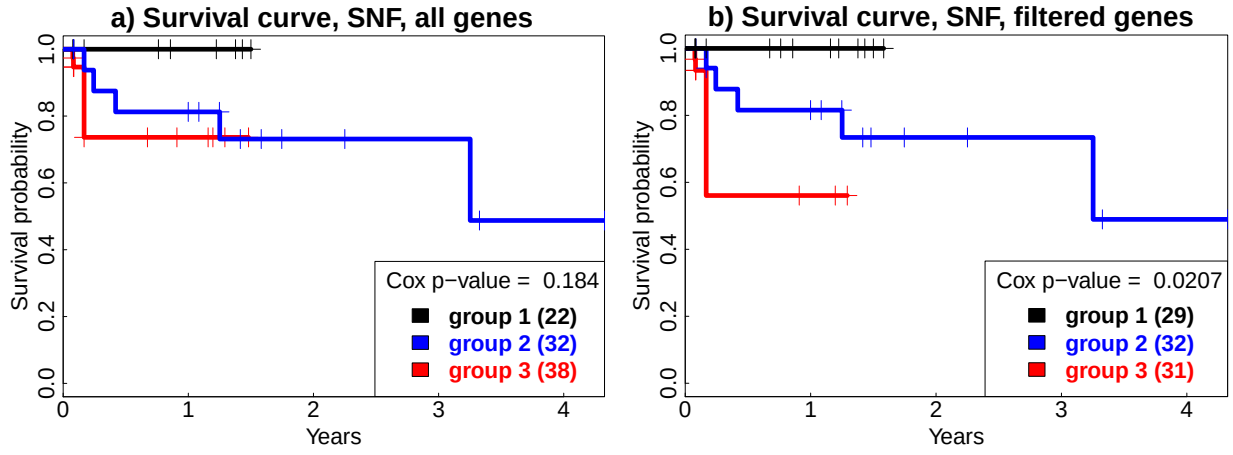
4.4 Subtyping using SNF

Similar to the assessment performed for k-means, we clustered the patients from the TCGA colon adenocarcinoma dataset using our pipeline in conjunction with SNF. To perform SNF clustering, we ran the SNFtool Bioconductor package with the parameters suggested by the authors [345]. We used the same input (KEGG pathways), settings (three clusters), and process previously described.

After this step, we obtained 184 clusterings, one per pathway. Then for each pathway, we constructed the empirical distribution and estimated the $p\text{-}value$ of how likely the pathway helps to improve disease subtyping. The estimated $p\text{-}values$ are shown in Table 4.2. The horizontal red line represents the significance threshold of 5%. There are 10 relevant pathways. We merged these relevant pathways to get a single set of genes that we used as our final set of selected features. This feature set contains 764 genes for the SNF method. Finally, we performed SNF clustering using these 764 genes.

Figure 4.7 shows the survival analysis of the resultant clusterings. Figure 4.7a shows the clustering when using SNF for all $17,814$ genes. The Cox p-value of this clustering

Figure 4.7: Kaplan-Meier survival analysis of the obtained sub-types using SNF. a) Survival curves using all genes. b) Survival curves using the selected genes.



is 0.1836, which is not significant (this result is identical to the result reported in [345]). Figure 4.7b shows the resultant clustering when using the 764 selected genes. The Cox p-value is 0.0207, which is approximately ten times lower than using all genes. Despite this meaningful improvement, none of the pathways has a corrected $p\text{-}value.fdr \leq 0.05$. This shows a lack of statistical power on our approach and an opportunity for improvement.

4.4 Subtyping using hierarchical clustering

Alike the assessment performed previously; we clustered the colon adenocarcinoma patients using our pipeline in conjunction with Hierarchical Clustering (HC) [111]. We set the dendrogram cutoff into three clusters according to prior knowledge. We used the 184 signaling pathways from KEGG [182]. The estimated *p-values* of the relevant pathways obtained with HC are shown in Table 4.3. The horizontal red line represents the significance threshold of 5%. We merged these three relevant pathways to get our final set of selected features. This feature set contains 195 genes for HC. Finally, we performed hierarchical clustering using the selected genes only.

Figure 4.8 shows the survival analysis of the resultant clusterings. Figure 4.8a shows the clustering when using HC for all 17,814 genes. The Cox p-value of this clustering is 0.799, which is not significant. Figure 4.8b shows the resultant clustering when using the

195 selected genes. The Cox p-value is 0.151, which is lower than using all genes, but it is still not significant. The sub-types obtained with hierarchical clustering do not separate the patients in clinically meaningful sub-types in any of the cases (neither using all genes nor filtered genes).

Table 4.3: List of pathways selected by our approach when using hierarchical clustering. These are the results of the third step of our pipeline, the selection of the relevant pathways. We first ranked the pathways by FDR adjusted p-value ($p\text{-}value.fdr$), then selected the pathways with a nominal $p\text{-}value \leq 0.05$ as relevant pathways.

Pathway	$p\text{-}value$	$p\text{-}value.fdr$
Cytosolic DNA-sensing pathway	0.01140	0.63874
Peroxisome	0.01200	0.63874
Fc epsilon RI signaling pathway	0.04090	0.63874
Complement and coagulation cascades	0.12390	0.80770

Given that our approach requires resampling for computing the $p\text{-}values$ p_i, this pipeline is more time consuming than traditional approaches. For the computational experiments presented here, we generated $10,000$ random samplings and clusterings per each pathway (184 pathways in total). Our pipeline took several hours to sub-type the set of patients (about 8 hours for k-means, 17 hours for SNF, and 46 hours for hierarchical clustering) while running any traditional clustering method takes only some minutes (less than 6 minutes). We ran these experiments on a typical desktop workstation with a 2.6 GHz Intel Core i5, 8GB of RAM, on a single thread, and the OS X 10.11 operative system.

4.5 Conclusions

In this chapter, we describe a framework to combine gene expression data, survival data, and biological knowledge available in pathway databases for a better disease subtyping. The performance of the new approach was demonstrated on the colon adenocarcinoma data downloaded from TCGA. The described framework was tested in conjunction with k-means, Similarity Network Fusion (SNF), and hierarchical clustering. For these clustering algorithms, our approach greatly improves the subtyping. In all cases, the Cox p-value is 3

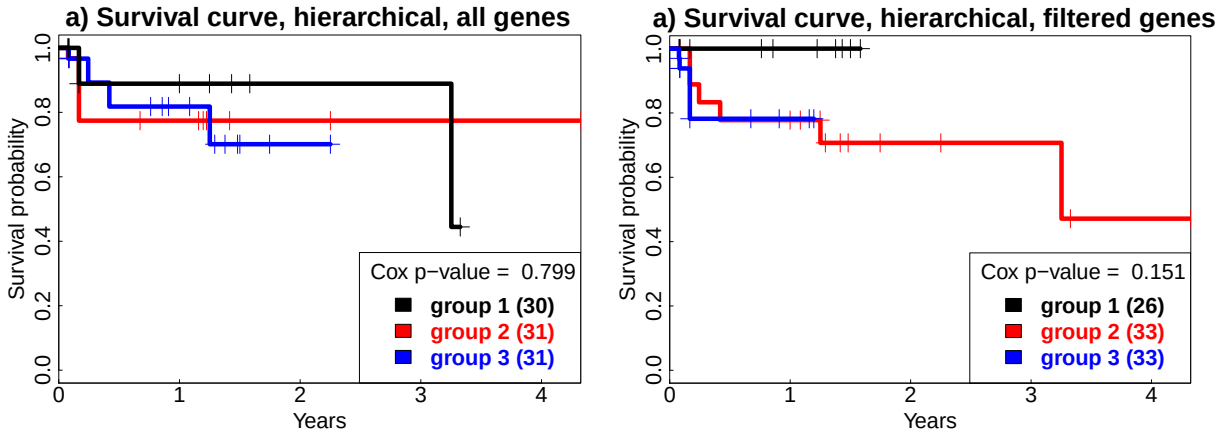


Figure 4.8: Kaplan-Meier survival analysis of the obtained sub-types using hierarchical clustering (HC). a) Survival curves using traditional HC. b) Survival curves using HC in our pipeline.

folds lower when using the selected features. Cox p-value improved from 0.129 to 0.0156 for k-means, from 0.184 to 0.0207 for SNF, and from 0.799 to 0.151 for hierarchical clustering.

Our contribution is two-fold. First, this framework introduces a way to exploit the additional information available in biological databases. Although the framework was demonstrated on KEGG pathways, it can exploit the information available in other databases, such as functional modules available in the Gene Ontology database or protein-protein interactions available in the STRING database. Second, this framework is the first one that integrates clinical data, biological pathways, and gene expression data for disease subtyping. For future work, we plan to use other clinical variables besides survival information and integrate multiple data types, such as microRNA, for a more comprehensive analysis [89]. Additionally, we plan to analyze the performance of feature selection methods from other contexts into the context of disease subtyping.

CHAPTER 5 CANCER SUBTYPING BASED ON CLINICAL AND MUTATION DATA

Successful transition into the era of precision medicine or screening, diagnostic, therapeutic, and prognostic procedures that take into account individual variability of patients [9, 64], requires comprehensive knowledge of complex relationships between molecular, biological and physiological processes in a human body. Stratification of patients into cohorts with a typical biological pattern is an essential aspect of such knowledge. Remarkable advances in the next-generation sequencing technology coupled with the widespread adoption of electronic health records (EHR) by healthcare providers in the United States have enabled the collection of unprecedented amounts of genetic and clinical patient data from which such knowledge can be discovered. Specifically, methods for high-throughput computational analysis of genetic and clinical data can help shed light on heterogeneous (molecular, biological, and physiological) markers that are highly predictive of survival as well as the outcome of therapeutic agents and treatment strategies. Prior research along this direction has focused on the methods to analyze genetic and clinical data in isolation with the goal of identifying either phenotypes (i.e. sets of biomarkers that are more prevalent in individuals with a particular disease or condition than in the general population) [56, 156, 157, 161, 264, 302, 350] or genotypes (i.e. DNA sequences that underlie specific diseases or traits) [111, 163, 193, 208, 345]. In particular, the recently proposed computational methods for discovering EHR-based phenotypes have been successfully applied to patient cohort identification [302] and determining the eligibility of patients for clinical trials [281]. On the other hand, the mapping of the human genome has enabled computational genotyping methods, which typically combine clustering [111, 193] with data integration [250, 345] and feature selection [87, 295] to identify the genes that are predictive of specific clinical outcomes [13, 66, 278]. Previous research on personalized approaches to cancer treatment has primarily focused on genetic studies, including identification of pathogenic mutations of individual genes in cancer tumors [208], tumor stratification [114, 140, 160, 278], functional diagnostics [122], and classification [287], as

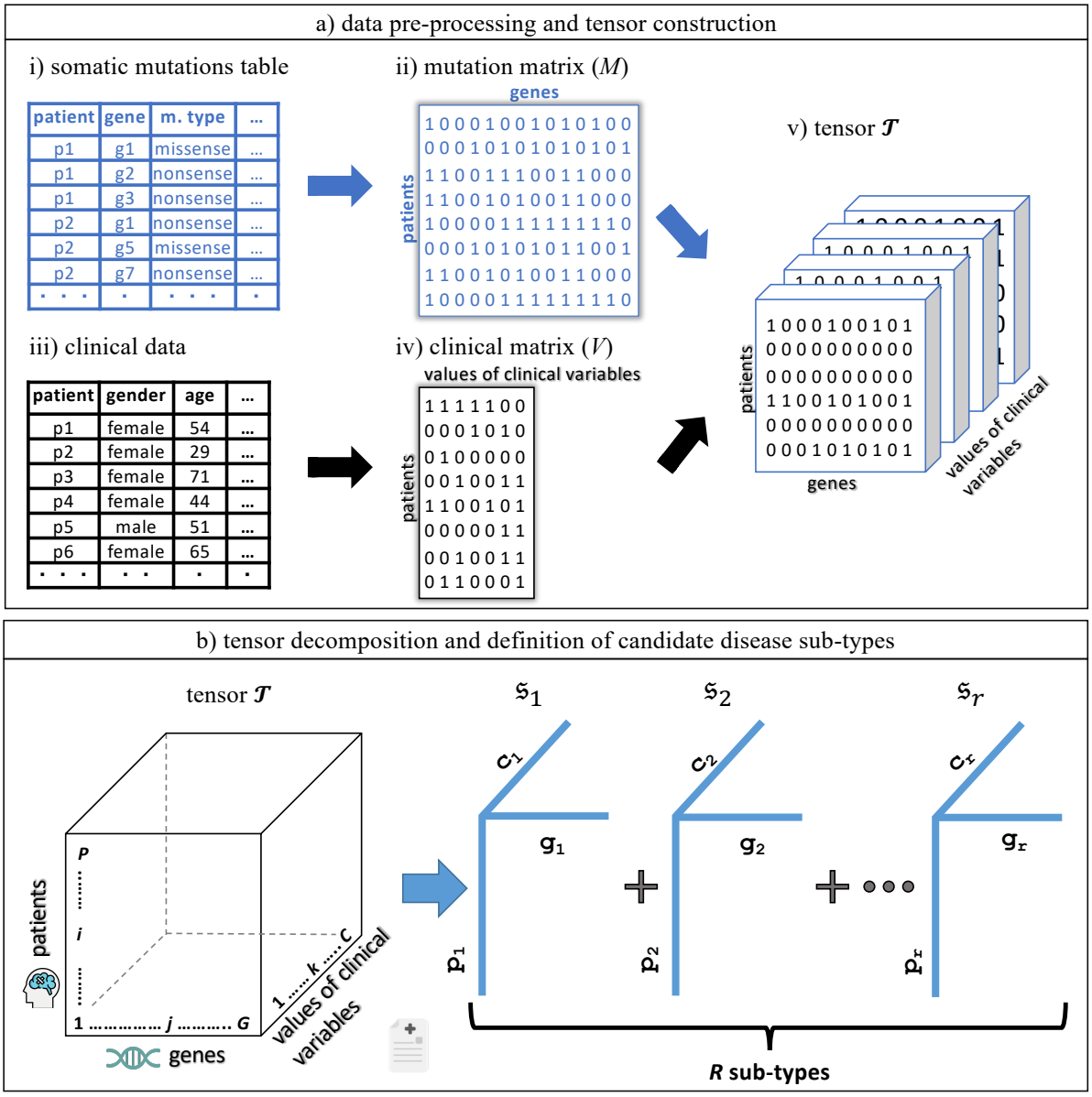


Figure 5.1: Stages of the proposed *CLIGEN* pipeline: A) data representation and construction of multi-modal three-dimensional tensor τ. B) obtaining candidate sub-types via CP decomposition of tensor τ.

well as creating centralized resources, repositories and protocols for interpreting, validating, sharing and updating the results produced by these studies [277].

Despite the immense progress in computational methods for analyzing clinical or genetic data, these methods alone cannot capture all aspects of the pathogenesis of complex diseases [284], such as cancer. The emergence of EHR-linked biobanks, such as those created by the Electronic Medical Records and Genomics (eMERGE) consortium [234], enable com-

Symbol	Definition
$M,\ V$	somatic mutation and clinical matrices
M_{ij}	cell of mutation matrix for patient i and gene j
V_{ij}	cell of clinical matrix for patient i and k-*th* value or interval of clinical markers
$\boldsymbol{\tau}$	multi-modal binary *CLIGEN* tensor
$P,\ G,\ C$	number of patients, genes as well as intervals and values of clinical markers
$i,\ j,\ k$	indices of patients, genes, and clinical markers.
t_{ijk}	value of tensor cell for the i-*th* patient, j-*th* gene, and k-*th* value or interval of clinical markers
$\circ$	outer product of two vectors
R	number of tensor components
$\mathfrak{s}_r$	r-*th* rank-one component i.e. a candidate verotype definition
$p_r,\ g_r,\ c_r$	patient, gene and clinical factor vectors

Table 5.1: Sample table title. List of notations used in this paper and their definitions

putational methods to discover associations between specific diseases and genes [33, 104, 190, 206] through genome-wide association studies (GWAS) [153] or between specific phenotypes and genes through phenome-wide association studies (PheWAS) [81, 191].

Previous tools for integrative analysis of TCGA data have been proposed [63, 369, 290, 289, 128]. The purpose of these tools is to retrieve TCGA data sets in a single environment for further performing integrative analyses. In [63], for example, the authors present an R package that includes one method to integrate DNA methylation and gene expression data and two distinct types of analysis: molecular analysis and clinical analysis. These tools are very useful (in fact, we use two of these tools cBioPortal and GDAC Firehose here in this work), but none of them include a method to analyze clinical and genetic data together.

Since cancer development and progression are influenced by several factors, including germ-line or somatic tumor genetics, overall patient health as well as environmental or lifestyle factors [32], it is natural to assume that cancer sub-types should incorporate all these different modalities of patient data. However, existing integrative approaches are specific to one particular type of clinical data, such as chemistry evaluations [40], survival [13, 90], or epidemiological data [262], and there has been relatively little research on computational

methods for joint analysis of clinical and genomic data for disease subtyping. We focus on the problem of identifying cohorts of patients, which share the same set of pathogenic gene mutations as well as the same values of clinical variables and markers. This problem is different from tumor stratification, which is based only on genetic information and is aimed at dividing the heterogeneous population of cancer tumors into biologically meaningful sub-types based on mRNA expression data [114, 278] or gene networks [160]. Although genome-scale molecular information provides an insight into biological processes driving tumor progression, cancer subtyping based on gene expression profiles alone has been shown to have limited correlation with clinical outcomes [60, 245].

Some studies combining clinical and genomic data for subtyping have been presented [215, 55, 80, 253], showing promising results for understanding Type II diabetes, traumatic brain injury, and bipolar disorder. Here we present three contributions, i) a high-throughput pipeline for fully unsupervised disease subtyping based on CLInical and GENomic data, *CLIGEN*, ii) its implementation as an open-source R package, and iii) the breast cancer sub-types discovered with this pipeline, which is presented here in two main stages. In the first stage, multi-modal patient data that includes somatic mutation profiles, as well as clinical variables and markers, is represented as a binary three-dimensional tensor. As differential measurements between a tumor and healthy tissue, somatic mutation profiles are more suitable for disease subtyping than other types of "omics" data, which are absolute measurements for each patient. Furthermore, somatic mutations capture causal genetic events underlying tumor progression, whereas mRNA or protein expression profiles are functional readouts of the current cell state and can be influenced by external factors that are unrelated to tumor biology. In the second stage, singular value decomposition (SVD) based smoothing is applied to the binary tensor to reduce the scarcity of our tensor. Then, tensor decomposition is applied to identify latent factors in each modality of the smoothed tensor. These latent factors correspond to the frequently co-occurring combinations of gene mutations and clinical markers in patients with a particular complex disease, such as cancer. We

hypothesize that the proposed pipeline enables the discovery of cancer sub-types (clinical disease sub-types combining phenotypes and genotypes) [31]. To validate this hypothesis, we applied the proposed framework to discover breast cancer sub-types based on the clinical and genetic data in the Cancer Genome Atlas (TCGA) Cancer Genome Atlas Research Network [249] and experimentally demonstrate that the discovered breast cancer sub-types can provide actionable insights, such as patient survival prognosis, to clinicians at the point of care.

5.1 Dataset Description

We used real patient data from The Cancer Genome Atlas - Genomic Data Commons Data Portal (TCGA GDC) downloaded from cBioportal [49] and firebrowse `http://firebrowse.org` on 9 April 2017. We used somatic mutation (non-silent mutation from the whole-exome sequencing level 3) profiles and clinical data of breast cancer patients. We considered only the patients for whom both somatic mutation and clinical data were available and discarded the genes that appear mutated on fewer than five patients. We manually searched for biomarkers through the 3402 variables of the clinical table downloaded from firebrowse. Then, we searched for biomarkers among the 30 variables of the clinical table from cBioportal. Next, we verified the consistency of the tables using common variables (i.e., variables on the firebrowse table and the cBioportal table). We removed two patients with inconsistent data (i.e., patients that have different values across the two databases[1]). This resulted in a dataset combining information about mutations in 499 genes and 32 dichotomized values of 11 clinical variables and markers for 482 patients.

5.1 Somatic Mutations

The downloaded somatic mutation table consists of 37 columns and 34032 registries. A registry in this table indicates a mutation in the gene reported in the column "*Hugo Symbol*"

[1] *1)* patient.bcr-patient-barcode tcga-e2-a14w has an inconsistent gender field; in one dataset, the patient is male and female in the other dataset. *2)* patient.bcr-patient-barcode tcga-b6-a0ru has an inconsistent age at initial pathology as 40 in one dataset and 49 in the other dataset.

Demographics and clinical history:
Age
Sex
Race
BMI
Prior cancer history
Family history of cancer
Diagnosis of diabetes
Menstrual status
Histopathology results:
Tumor size
Histology designations
Tumor grade
Cancer stage
Lymph node stage
Metastasis stage
Test results for molecular markers or their immunohistochemistry surrogates:
EGFR protein
Cytokeratin 5/6 (CK 5/6)
ER/PR
HER1
HER2
TP53
CA-125
Prostate-specific antigen (PSA)
KRAS
ERBB2
UGT-1A
EML4
ALK
BRCA

Table 5.2: List of biomarkers.

for the sample in the column "*Tumor Sample Barcode*". Additional details on the processing and organization of these data are available in [49]. In this work, we constructed patient mutation profiles as binary vectors, in which a bit is set, if the patient's gene corresponding to that position in the vector harbors a mutation.

5.1 Clinical Variables and Markers

The clinical variables and markers (without loss of generality, further referred to as clinical markers) in TCGA include the age of diagnosis with cancer, gender, estrogen receptor (ER)

status, progesterone receptor (PR) status, human epidermal growth factor receptor 2 (HER2) final status, American Joint Committee on Cancer (AJCC) coded tumor (T), AJCC coded lymph node invasion status (N), AJCC coded metastasis (M), histology, cancer stage, and patient ethnicity, see Table 5.2. The complete description of each of these clinical variables can be found in [249].

5.2 Proposed Pipeline

The proposed pipeline for unsupervised subtyping of complex diseases, *CLIGEN*, is illustrated in Figure 5.1. The input to the pipeline consists of genetic and clinical data of the patients with the same complex disease (e.g., cancer). The output of the pipeline is a set of definitions of disease sub-types characterized by genetic and clinical markers. The pipeline consists of three stages: *i)* data pre-processing *ii)* tensor construction and *iii)* decomposition of the constructed tensor to derive *multi-modal disease sub-types*. Each of these stages is discussed in detail below, and the notations used in these descriptions are summarized in Table 5.1. The source code for *CLIGEN* is publicly available at `github.com/datad/CLIGEN`.

5.2 Data pre-processing

The first stage of the proposed pipeline, illustrated in Figure 5.1A, involves pre-processing the input to create a *combined multi-dimensional representation* of genomic and clinical patient data for subsequent analysis. Given the input mutation table, *CLIGEN* constructs a binary mutation matrix $\mathbf{M}$ with patients as rows and genes as columns. A value of the cell $\mathbf{M}_{ij}$ of matrix $\mathbf{M}$ is set to 1, if the ith patient has a mutation in the jth gene and to 0, otherwise. Continuous clinical variables, such as the age of cancer diagnosis, are discretized into intervals. The values of clinical variables for all patients are represented as a binary clinical matrix $\mathbf{V}$ with patients as rows and values of clinical variables as columns. A value of the cell $\mathbf{V}_{ik}$ of matrix $\mathbf{V}$ is set to 1, if the ith patient has the kth value of clinical variables.

Somatic Mutation Data Representation

Somatic mutation datasets typically take the form of mutation tables, in which the rows correspond to mutations, and the columns describe the type of each mutation and its location (Figure 5.1 Ai). Based on the input mutation table, *CLIGEN* constructs a binary mutation matrix M with patients as rows and genes as columns. The value of 1 in the cell M_{ij} of matrix M indicates that patient i has at least one non-silent mutation (i.e., a mutation of any of the following types: missense mutation, nonsense mutation, non-stop mutation, in-frame insertion, in-frame deletion, or frameshift mutation) in gene j, while the value of 0 indicates that patient i has no mutations in gene j. Genes with mutations appearing in less than five patients were discarded from the mutation matrix.

Clinical Data Representation

Continuous clinical variables or markers, such as the age of diagnosis, were discretized into intervals with a total of n combined intervals of all continuous markers and levels of all discrete markers in the entire dataset. The values of clinical markers for each patient were represented as a binary clinical matrix V with patients as rows and intervals of continuous or levels of discrete clinical markers as columns. The value of 1 in the cell V_{ik} of matrix V indicates that patient i has a *k-th* level or interval of a discrete or continuous clinical marker.

Smoothing of mutation data

Due to the sparse nature of the binary mutation matrix, we smooth it with the following Singular Vector Decomposition (SVD) based approach. First, we find the singular value decomposition of the binary mutation matrix, which outputs its eigenvalues and eigenvectors. Then, matrix multiplication of the top 50% of the eigenvalues and eigenvectors was performed to obtain our *smoothed mutation matrix*.

5.2 Tensor construction

Matrices $\mathbf{M}$ and $\mathbf{V}$ are combined to create a three-dimensional binary tensor (i.e., multidimensional array of objects) $\boldsymbol{\tau} \in \mathbb{R}^{P \times G \times C}$, which captures interactions between somatic mutations and clinical variables. The first mode of tensor $\boldsymbol{\tau}$ corresponds to P patients in the population, while the other two modes correspond to G distinct genes and C distinct values of clinical variables (Figure 5.1.b). Each cell t_{ijk} of tensor $\boldsymbol{\tau}$ has a binary value (1 or 0) which is set to 1, if the ith patient has at least one mutation in the jth gene and the kth value of clinical variables or to 0 otherwise.

5.2 Tensor decomposition

Tensor decomposition [199] is a powerful mathematical technique that has been successfully applied in different domains ranging from psychology and neuroscience to computer vision [242]. In biomedical informatics, tensor decomposition has proven to be useful for understanding cellular states [366] and biological processes [256], in addition to EHR-based phenotyping [156, 157, 350]. Tensor decomposition has several advantages over matrix factorization. First, tensors explicitly account for the multiway structure of the data that is otherwise lost, when a tensor is converted into a matrix by collapsing some of its modes. Second, some tensor decomposition methods guarantee the uniqueness of the optimal solution even for very sparse tensors. The two most widely used tensor decomposition methods are the Tucker method [334] and Candecomp/Parafac (CP) which stands for Canonical Decomposition (CANDECOMP) [51] and Parallel Factor Analysis (PARAFAC) [145]. CP decomposes a tensor into a linear combination of rank-one tensor components [51].

CLIGEN utilizes CP tensor factorization [199] to identify disease sub-types as groups of latent factors in $\boldsymbol{\tau}$. CP decomposition approximates $\boldsymbol{\tau}$ with $\hat{\boldsymbol{\tau}}$, a linear combination of rank-one tensors. Formally:

$$\boldsymbol{\tau} \approx \hat{\boldsymbol{\tau}} = [\![\lambda, \mathbf{P}, \mathbf{G}, \mathbf{C}]\!] = \sum_{r=1}^{R} \lambda_r \cdot \mathfrak{s}_r = \sum_{r=1}^{R} \lambda_r \cdot \mathbf{p}_r \circ \mathbf{g}_r \circ \mathbf{c}_r \tag{5.1}$$

where R is the number of rank-one tensors $\mathfrak{s}_r$ that $\boldsymbol{\tau}$ is decomposed into, $\lambda_r \in \mathbb{R}$ is the weight of the rth rank-one tensor. Each $\mathfrak{s}_r$ is an outer product ($\circ$) of patient $\mathbf{p}_r \in \mathbb{R}^P$, gene $\mathbf{g}_r \in \mathbb{R}^G$ and clinical $\mathbf{c}_r \in \mathbb{R}^C$ latent factors. Patient, gene and clinical latent factors that correspond to each rank-one tensor can be thought of as clusters of patients with frequently co-occurring somatic gene mutations and clinical variables. Latent factors for all rank-one tensors can be grouped into the columns of the patient $\mathbf{P}$, gene $\mathbf{G}$ and clinical $\mathbf{C}$ factor matrices. CP decomposition of $\boldsymbol{\tau}$ is obtained by solving the following optimization problem:

$$\min_{\hat{\boldsymbol{\tau}}} \|\boldsymbol{\tau} - \hat{\boldsymbol{\tau}}\|_{\mathcal{F}} \tag{5.2}$$

aimed at finding the best approximation of each element t_{ijk} of the original tensor $\boldsymbol{\tau}$ from the latent factors corresponding to rank-one tensors as follows:

$$t_{ijk} \approx \sum_{r=1}^{R} \lambda_r p_{ir} g_{jr} c_{kr} \tag{5.3}$$

Uniqueness of the optimal solution to the above optimization problem is an important property of CP decomposition [199].

Molecular and clinical markers of disease sub-types are derived from the gene and clinical latent factors associated with each rank-one tensor obtained by CP decomposition of $\boldsymbol{\tau}$. Each element of a gene and clinical latent factor can be interpreted as a degree of specificity of a particular gene or a clinical variable to the corresponding disease subtype. Each element of a patient latent factor can be interpreted as a membership proportion of a particular patient in the corresponding disease subtype.

Slicing tensor $\boldsymbol{\tau}$ along each of its three modes yields the following views:

1. *Patient mode*: each slice is a matrix of co-occurrences of mutations and clinical markers for a particular patient. For instance, if a patient's health records indicate stage I breast cancer, and her mutation profile indicates a mutation in gene TP53, then a cell at the row for the "Stage I" clinical variable and the column for the TP53 gene in the matrix

corresponding to the tensor slice for this patient will have the corresponding smoothed value of gene TP53. The cells in the same column and the rows for "Stage II", "Stage III" and "Stage IV" will have the value of 0.

2. *Gene mode*: each slice is a matrix with patients as rows and clinical markers as columns, which shows how a mutation in a particular gene is correlated with clinical markers in different patients. Such matrix can be considered as a summary of phenotypic manifestations of a particular gene mutation.

3. *Clinical mode*: each slice is a matrix with patients as rows and genes as columns, which shows how gene mutations in different patients are correlated with a particular clinical marker. Such a matrix can be considered as a summary of genetic markers for a single phenotype.

5.3 Challenges and Limitations

Methods for integrative analysis of genomic and clinical data face a common challenge of dealing with large volumes and high-dimensionality of data. By utilizing sparse representations and inexpensive linear algebra operations, tensor factorization methods effectively address this challenge. Successful application of tensor decomposition in different domains led to further research into efficient optimization methods for tensor decomposition [1], which makes tensor decomposition a method of choice for high-throughput cancer subtyping.

We propose a method capable of incorporating clinical data into the pipeline of mutation based stratification by utilizing a tensor based-representations and inexpensive linear algebra operations, tensor factorization methods effectively address this challenge. Advancements in efficient tensor decomposition methods and their broad application to different domains makes tensor decomposition a method of choice for high-throughput disease subtyping.

Since tensor factorization models are parametric, selecting the optimal number of components for CP decomposition of the binary tensor (i.e. model order estimation) is an important practical aspect of the proposed pipeline. Too few components typically result in general

subtype definitions, which may combine several actual disease sub-types. Too many components typically result in specific subtype definitions, which may split the actual disease sub-types. It is important to point out that, in terms of the number of model parameters, CP decomposition, which assumes that the number of components is the same per each tensor mode, has an advantage over Tucker decomposition, which requires specifying the number of components per each mode. While it is known that the number of components that minimizes the reconstruction error of the original tensor from its components is equal to the rank of a given tensor [59, 199], finding tensor rank is an NP-complete problem [146]. Even if the rank of a tensor is known, the number of components that minimizes reconstruction error may not result in the best accuracy for a particular task, such as survival prediction. Identifying the parameters of our model, number of components and number of clusters, still remains a challenge. However, we show here that empirical experiments with the combination of the possible values of these parameters is feasible way to address this problem. Therefore, the optimal number of components is typically determined using heuristics, such as core consistency diagnostic [37], cross validation [38] (as was done in this work) or hierarchical Bayesian approach [243], if a suitable prior can be defined.

Tensor construction is another aspect of the proposed approach with possible variations. In this work, we used the presence or absence of non-silent gene mutation as a single genetic signature of patients. However, it is possible to use other types of genetic data, such as gene expression or copy number variation. It is also possible to construct a count tensor, instead of a binary tensor, by taking into account both the type of mutations and the number of mutations per gene, which we leave for future work.

5.4 Results

We performed both qualitative and quantitative evaluation of breast cancer sub-types derived from a TCGA breast cancer dataset [49] using the proposed pipeline. Here, we present the quantitative evaluation. The qualitative evaluation is presented in section 5.6.

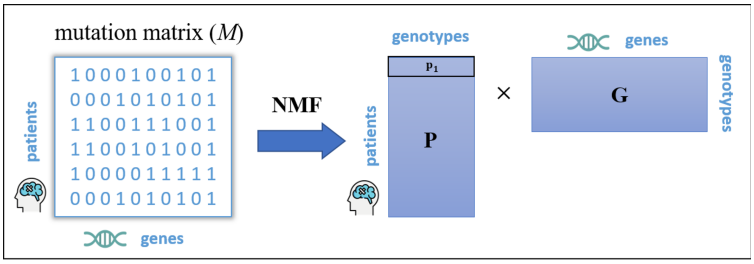


Figure 5.2: Obtaining genotypes through non-negative factorization of a binary somatic mutation matrix. The rows in matrix P correspond to a vector of genotype memberships (e.g. p_1 is a vector of genotype memberships for the first patient). Rows in matrix G correspond to genotype definitions.

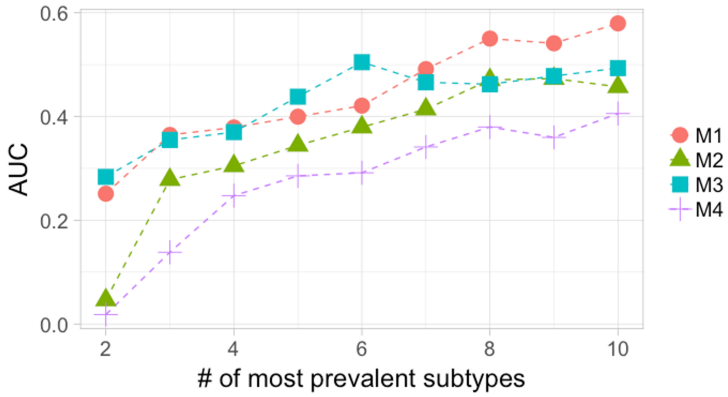


Figure 5.3: AUC of Cox models for breast cancer patient survival prediction that utilize patient membership proportions in most prevalent sub-types obtained by the proposed pipeline (M1), patient membership proportions in genotypes (M2) obtained by NMF of binary somatic mutation matrix, phenotypes (M3) obtained by NMF of binary clinical matrix as predictors, and random patient membership proportions (M4).

5.4 Quantitative Evaluation

Quantitative evaluation was conducted for the task of patient survival prognosis, which is important for personalizing cancer treatment [291]. Specifically, we compared Cox proportional hazards models that use the following predictors for survival prognosis of breast cancer patients:

- M1: membership proportions of each patient in breast cancer sub-types discovered by the proposed pipeline, which correspond to a row in the patient factor matrix $\mathbf{P}$;
- M2: membership proportions of each patient in breast cancer genotypes, which cor-

Model	Wald-rank test	Wald test
M1	8.327e-15s	0.000082
M2	0.315	0.3772
M3	0.0007	0.0124
M4	0.5060	0.5509

Table 5.3: p-values of Log-rank and Wald tests of the Cox proportional hazard models utilizing patient membership proportions in sub-types (M1), genotypes (M2), phenotypes (M3), and random membership assignment (M4) as predictors.

respond to a row in matrix **P** obtained by non-negative factorization of the somatic mutation matrix **M** (as in Hofree et al. [160] without gene network smoothing), as shown in Figure 5.2;

- M3: membership proportions of a patient in breast cancer phenotypes, which correspond to a row in the patient factor matrix obtained by non-negative factorization of the clinical matrix V;

- M4: random membership proportions of a patient in each number of breast cancer sub-types.

In the first experiment, we compared the accuracy of the Cox models using each of the above predictors for survival prognosis of breast cancer patients, while in the second experiment, we compared the goodness of fit of these models.

Accuracy of Survival Prognosis

In the first experiment, we compared the area under the ROC curve (AUC) for the models M1-M4 using randomized 10-fold cross validation. The Cox models were estimated using the data in the training splits and evaluated using the data in the testing splits.

The plot of AUC values for models M1-M3 micro-averaged over splits by varying the number of the most prevalent cancer sub-types, genotypes and phenotypes is shown in Figure 5.3, from which two major conclusions can be drawn. First, the Cox regression model that utilizes patient membership proportions in sub-types obtained by *CLIGEN* (M1) is consistently more accurate at predicting breast cancer survival than the Cox model that

uses membership proportions in genotypes obtained by NMF (M2) and phenotypes obtained by NMF (M3), which indicates the importance of taking into account both clinical and genomic data when determining cancer sub-types. In particular, the Cox model utilizing patient subtype membership proportions as predictors achieved the highest AUC of 0.5796, when ten most prevalent sub-types were used, while the Cox model utilizing patient genotype memberships as predictors achieved the highest AUC of 0.4731, when the nine most prevalent genotypes were used, and the Cox model utilizing patient phenotype memberships as predictors achieved the highest AUC of 0.5047 from the six most prevalent phenotypes.

Second, the Cox models utilizing patient membership proportions in the top-k most prevalent sub-types derived by *CLIGEN* and phenotypes and genotypes derived by NMF [129] are all more accurate at predicting breast cancer survival than the baseline Cox model utilizing random patient membership proportions (AUC = 0.4056).

Goodness of Fit

In the second experiment, we compared the goodness of fit of the models M1-M4 estimated on the entire TCGA dataset. The p-values of Log-rank and Wald tests of these models are summarized in Table 5.3.

Both tests indicate that patient membership proportions in sub-types derived by *CLIGEN* are more statistically significant predictors of breast cancer patient survival than membership proportions in breast cancer phenotypes, which in turn are more statistically significant predictors than random patient membership proportions and membership proportions in genotypes derived by NMF. This important finding illustrates the need to combine clinical and genomic data in order to more accurately predict survival of breast cancer patients.

Kaplan-Meier survival plots for the 4 most prevalent breast cancer sub-types obtained by *CLIGEN* and NMF of mutation and clinical matrices are shown in Figure 5.4. As follows from Figure 5.4, breast cancer patient cohorts that correspond to the 4 most prevalent sub-types obtained using *CLIGEN* are more distinct in terms of survival dynamics (p = 0.0493) than the patient cohorts that correspond to the 4 most prevalent molecular (p =

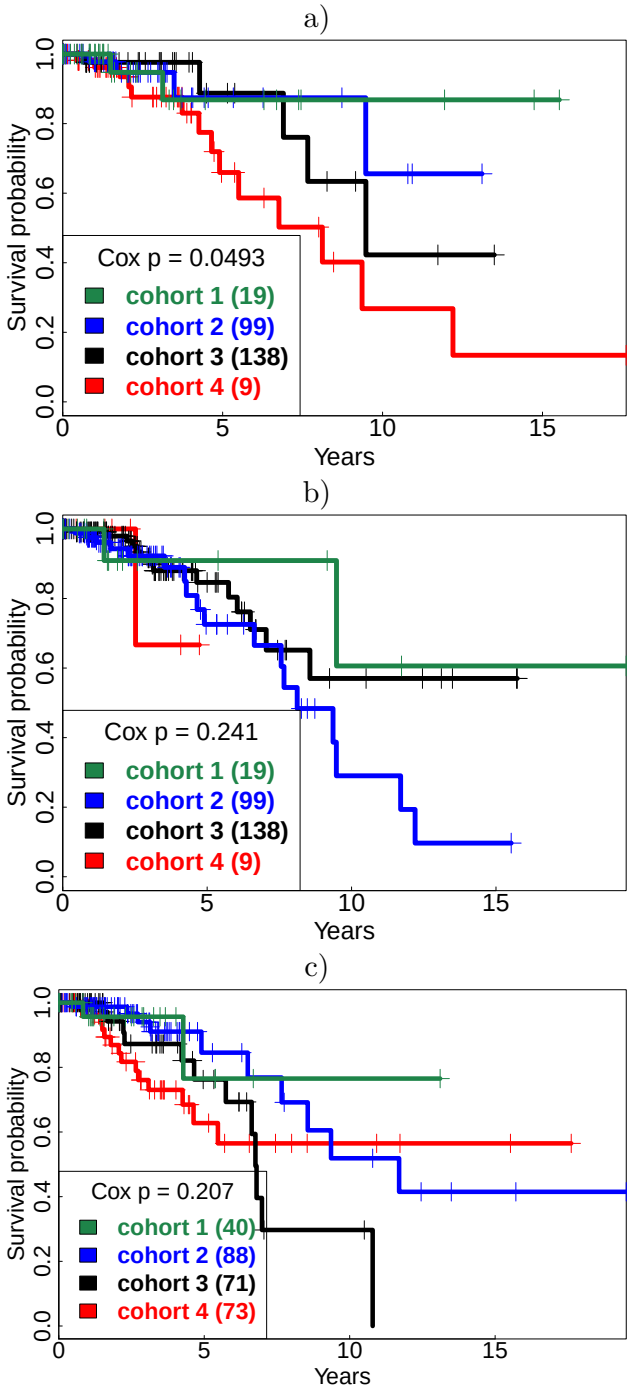


Figure 5.4: Kaplan-Meier survival plots for the four most prevalent sub-types: A) sub-types obtained using *CLIGEN*, B) genotypes obtained using NMF, C) phenotypes obtained using NMF.

0.241) and clinical ($p = 0.2073$) phenotypes. The curves in Figure 5.4a correspond to the CLIGEN components 10, 8, 2, and 1 (the patients from component 10 correspond cohort 1,

patients from component 8 to cohort 2, those from component 2 to cohort 3, and those from component 1 to cohort 4. Cohort 1, which according to the analysis in Section 5.6.1 consists mainly of Luminal B patients has worse survival than cohort 2, which mainly consists of Luminal A patients. Interestingly, GLIGEN was able to identify a small (19 patients) and a fairly large (99 patients) cohorts 1 and 2, which have significantly longer term survival than cohorts 3 and 4, but cannot be equivocally assigned to any known molecular subtype of breast cancer discussed in Section 5.5.1. This finding indicates that there may exist sub-types of breast cancer, which can be identified by the joint analysis of clinical and genomic data that require less aggressive treatment than the known molecular sub-types.

5.4 Survival Analysis

To assess if the sub-types obtained with *CLIGEN* are clinically relevant, we compared the survival curves of the obtained patient cohorts using the traditional Cox log rank test. We used two traditional clustering methods, k-means and hierarchical clustering, with different numbers of clusters, k = 2-15. To find out the best clustering, we run CP factorization on the CLIGEN tensor varying the value of R (tensor rank) from 2 to 10. In addition, we tested the effect of outliers by running these tests with all the data and comparing with the results obtained without major outliers. The outliers of the distributions of each component were detected by identifying the points that falls outside the outer fences, which was calculated by three times the interquartile range plus Q3-Q1. The lowest Cox log rank p-value (p = 2.39E-16) is obtained with Hierarchical clustering without outliers (Tables 1-4 in supplementary materials show the complete results). We compared our method with the state of the art method network-based stratification (NBS) [160] which gives us a Cox p-value = 0.00477 (see supplementary materials for complete NBS results). NBS is a method to integrate somatic tumor genomes with gene networks. This approach clusters together patients with mutations in similar network regions using Non-negative Matrix Factorization (NMF). Figure 5.4 shows the Kaplan-Meier survival plots of the clustering with lowest p-value obtained with *CLIGEN* that produces balanced number of groups and compared with the stratification obtained with

NBS. As we can see, the sub-types obtained using *CLIGEN* are significantly more distinct in terms of patient survival ($p = 0.00176$) than stratification using *NBS*.

5.5 Breast Cancer Molecular sub-types and their Implications on Diagnosis, Treatment and Mortality

In this section, we present the necessary background regarding molecular sub-types and treatment of breast cancer, followed by biological analysis of the results obtained with CLIGEN. Although CLIGEN can be utilized to analyze other complex diseases, this thesis only reports its application to the TCGA Breast Cancer (BC) data. We focused on BC because it is a well-studied cancer with known molecular sub-types. Therefore, we validate the sub-types discovered with CLIGEN against these known molecular BC sub-types, which gives us an opportunity to discover novel or refine existing sub-types. In this section, we discuss these known BC sub-types and the procedures to treat patients, which belong to each subtype.

5.5 Molecular sub-types of Breast Cancer

Based only on the genes that cancer expresses, Perou et al. have characterized the genomic diversity of breast tumors, to define five molecular (or intrinsic) sub-types of Breast Cancer (BC) using the PAM50 assay [21], i.e., Luminal A, Luminal B, Triple-negative/basal-like, HER2-enriched, and Normal-like [269, 73, 35]:

- Luminal A: ER-positive and/or PR positive, HER2 negative, and has low levels of the protein Ki-67. Luminal A cancers are low-grade, tend to grow slowly, and have the best prognosis.

- Luminal B: ER and/or PR positive, and HER2 positive or negative with high levels of Ki-67. Luminal B cancers generally grow slightly faster than luminal A cancers, and their prognosis is slightly worse.

- Triple-negative/basal-like: ER, PR, and HER2 negative. This type of cancer is more common for younger, African-American women, and women with BRCA1 gene muta-

tions than Caucasian women.

- HER2-enriched: ER negative, PR negative, and HER2 positive. HER2-enriched cancers tend to grow faster than luminal cancers and have a relatively worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.

- Normal-like: share gene expression patterns with Luminal A subtype (ER-positive and/or PR positive, HER2 negative, and has low levels of the protein Ki-67) and are characterized by a normal tissue profiling.

Ductal Carcinoma In Situ (DCIS)
Invasive Ductal Carcinoma (IDC)
(IDC type) Tubular Carcinoma of the Breast
(IDC type) Medullary Carcinoma of the Breast
(IDC type) Mucinous Carcinoma of the Breast
(IDC type) Papillary Carcinoma of the Breast
(IDC type) Cribriform Carcinoma of the Breast
(IDC type) Apocrine Carcinoma of the Breast
Invasive Lobular Carcinoma (ILC)
Inflammatory Breast Cancer
Lobular Carcinoma In Situ (LCIS)
Angiosarcomas
Male Breast Cancer
Molecular Sub-types of Breast Cancer
Paget's Disease of the Nipple
Phyllodes Tumors of the Breast
Metastatic Breast Cancer

Table 5.4: List of histological types of breast cancer.

5.5 Histological Types of Breast Cancer

There are a significant number of different histological types of BC, and a patient could be diagnosed with a combination of different types of BC. Cancer can begin in different areas of the breast; the lobules, the ducts, or the tissue in between [35] and can be divided into two main groups: the carcinomas and the sarcomas. Carcinomas are divided into two major sub-types: adenocarcinoma, which develops in an organ or gland, and squamous cell

carcinoma, which originates in the squamous epithelium. Some histological types of cancer are shown in Table 5.4 [120].

Each of these types has different implications on mortality and the required treatment; for example, Ductal Carcinoma In Situ (DCIS) is a non-invasive breast cancer (stage 0). The cancer cells are contained within the milk ducts. In Invasive Ductal Carcinoma (IDC) cancers, the cancer cells inside of a milk duct or lobule have spread to the nearby tissues.

Most invasive BCs often called 'Ductal Carcinoma NOS (not otherwise specified)' are of a generic variety and about 5 or 6 common types with an identifiable cellular appearance and behavior. Beyond this, there are many rare forms of BC and a number that is very difficult to classify because they have features from different BC types or contain a large percentage of benign tissue. Some of the names given to BC types refer to the visual characteristics of the malignant cells and cell formations. Also, patients can be diagnosed with hereditary forms of BC, such as those linked to the BRCA1 and BRCA2 genes and other gene mutations, before the appearance of any tumor.

5.5 Breast Cancer Diagnosis and Treatment

The process of diagnosis of BC involves a pathologist who is an expert in examining cells from biopsy samples under a microscope, which is crucial to diagnose the presence of cancer cells. The pathology report will include the BC stage, the tumor-lymph node-metastasis status (TNM status), and type of BC based on the cell's morphology. Current clinical practice depends on the pathology report to diagnose BC. The BC stage can fall into one of the five categories (stage I, II, III, IV, or V) and is determined by the tumor size (T). The TNM status indicates the extent of the primary BC tumor, the presence or absence of lymph node metastasis, and the presence of distant metastasis. The cell's morphology describes signs of malignant activity and cell formations. Malignant formations can include irregular nuclear borders and shapes, extra-large nuclei, cell dissociation, arrangements of cells in clusters and necrosis.

Additionally, pathologists perform immunohistochemical tests to evaluate the status of

the hormone receptors on a sample to determine the histological type of the BC tumors and predict essential aspects of BC behavior. These tests are performed on a set of biochemical markers using protein-based dyes and visually inspecting the sample. Other elements included in the histological evaluation include the genetic type of the cancer cell (which can be different from epithelial cells), calcifications, necrosis, fat, blood, and lymphatic responses.

There are three major biochemical markers for BC: estrogen receptors (ER), progesterone receptors (PR), and epidermal growth factor receptors (HER2). These are also referred to as predictive immunohistochemistry (IHC) markers for targeted treatment. ER/PR positive tumors (ER/PR+) have high levels of estrogen/progesterone receptors and can be treated with chemo/endocrine therapy, such as the use of tamoxifen. Patients with ER/PR positive breast tumors have a lower mortality risk compared to women with ER-/PR+ and ER-/PR- breast tumors. ER/PR+ cancers have a prevalence of 70% of all invasive BCs.

HER2 positive tumors (HER2+) are associated with a slightly poorer prognosis and a higher risk of local recurrence, but there are targeted therapies for this type of cancer. HER2+ cancers can be treated with targeted therapy such as Herceptin, Perjeta, Tykerb, Nerlynx, and Kadcyla [35]. The prevalence of HER2+ cancer is between 15% and 20% of all invasive BCs. Triple-negative breast cancer (TNBC) tumors lack receptors for estrogen, progesterone, and HER2 and are harder to treat because they do not respond to targeted therapies (i.e., drugs that target ER, PR, or HER2). The primary treatment for TNBC patients is chemotherapy. Studies have shown that genetic mutations are more common in women with TNBC, even if they do not have hereditary BC. The prevalence of TNBC is between 10% and 15% of all BC.

Immunotherapy

Traditional treatments of cancer include chemotherapy, radiation therapy, and surgery. An emerging type of cancer treatment that "trains" the immune system to attack cancer cells is immunotherapy. The immune system is in charge of fighting infections and other diseases by detecting foreign bodies and attaching them. It includes the lymph node system

and white blood cells [172].

Typically, the immune system is unable to detect cancer cells as malign bodies, and it does not attack them. One type of immune therapy, the immune checkpoint inhibitors therapy, disrupts cancer cell's signals to expose them to the immune system. Once the immune system is able to detect cancer cells, it can attack them and kill them [23].

Currently, immunotherapy is applied to late states of cancer and some clinical trials. In particular, checkpoint inhibitors have been FDA-approved to treat a variety of cancers, including lung cancer, melanoma, bladder cancer, kidney cancer, and lymphoma. It has not been proven effective for Breast Cancer yet.

Just recently, clinical trials to use immunotherapy in early cancer stages have been approved. Triple-negative type of cancer is a good candidate for clinical trials because there are no treatments that are effective, but CLIGEN could suggest that not all TNBC patients have a high mutation load, maybe patients that were identified in Component 1 could benefit from immunotherapy more than other TNBC patients. Without this distinction, immunotherapy would be delivered in clinical trials to TN patients that are not suitable for this treatment and possibly be marked as an ineffective treatment for breast cancer when the problem might not be the treatment, but the cohort.

5.6 Biological analysis of CLIGEN

In this section, we present the results of biological evaluation of breast cancer sub-types obtained by CP decomposition of the binary *CLIGEN* tensor constructed from the TCGA dataset into ten components, since this decomposition gives the most accurate prediction of cancer patient survival. We analyze each of the three resultant factor matrices independently and then compare the conclusions across all dimensions. First, we analyze the genetic factor matrix, then the clinical factor matrix, and conclude with the patient factor matrix. Second, we analyze the mapping of traditionally used breast cancer molecular sub-types onto the sub-types obtained with CLIGEN.

Analysis of genetic factors associated with CLIGEN components

Enrichment analysis of the genes associated with each sub-type derived by CLIGEN was performed using the Ingenuity Systems Upstream Analysis tool [202]. Breast cancer sub-types associated with the components obtained by *CLIGEN* were first analyzed to identify biological processes behind them. Component 1 (shown in Figure 5.5A) corresponds to a small cohort of patients with a high mutation load. Further investigation of this sub-type revealed a large number of mutations in the tumor suppressor genes (BRCA1, BRCA2, TP53, PTEN, RB1) that participate in DNA repair, which indicates that the high mutation load may be associated with a mutation in a DNA repair gene pathway(s). As follows from Figure 5.6, for each sample, these mutations were mutually exclusive.

a) subtype characterized by a high mutation load, which may be associated with a mutation in a DNA repair gene pathway(s).

Component 1	Top genes		
2.761% of patients	TP53	DNAH7	MACF1
	PIK3CA	MUC17	CROCCP2
	TTN	SSPO	ASXL3
0% of clinical variables	BRCA1	FAT3	SYNE1
	USH2A	OBSCN	DST
	Unknown	FLG	RYR2
	MUC16	KMT2C	LRP2

b) triple negative breast cancer verotype defined by the lack of estrogen receptor (ER), progesterone receptor (PR) and the human epidermal growth factor receptor (HER2) expression

Component 5	Top genes		
36.489% of patients	TTN	UBR4	DNAH11
Metastasis.M1	MUC16	HDAC6	UBA1
AJCC.StageStage.X	CROCCP2	SPI1	ELL3
Tumor.TX	KMT2C	SETD2	OR2B2
ER.StatusNegative	WDFY3	GON4L	UBA6
	SYNE2	ZNF821	SPEN
	VCAN	LRP2	BLM

c) subtype characterized by overexpression of progesterone receptor (PR) and estrogen receptor alpha-positive (ER+). It typically has the worst survival prognosis relative to other breast cancer subtypes.

Component 6	Top genes		
35.503% of patients	TP53	SMG8	SAMD9
AJCC.StageStage.III	KDM5C	MTO1	ITGA1
AJCC.StageStage.IIIB	NCOA6	KL	WDR52
PR.StatusPositive	HYDIN	ZFHX3	FXR1
ER.StatusPositive	PPHLN1	SYNE3	SCN7A
	BTD	ATG16L1	BID
	BRIP1	TUBA1B	ACAD11

d) subtype characterized by overexpression of the human epidermal growth factor receptor (HER2) oncogene and responsive to HER2-targeted inhibitors

Component 9	Top genes		
48.521% of patients	TTN	LRP2	EMR2
ER.StatusPositive	USH2A	ATP1A4	RP11-32B5.1
HER2.StatusPositive	MUC16	TLN2	ZFHX3
Metastasis.CodedNegative	RYR2	ERBB3	TRIP11
Nodes.DetailsPositive	FLG	COL6A3	SRCAP
TumorT3	WDFY3	DNAH11	MEP1A
	GPR98	DNAH8	PEG3

Figure 5.5: Examples of sub-types identified by the proposed pipeline.

Component 5 (shown in Figure 5.5B) appears to correspond to a sub-type of triple-negative breast cancer (TNBC), which is defined by the lack of ER, PR, and HER2 expression. Molecular aberrations driving this breast cancer sub-type remain undefined, and patients with this sub-type of breast cancer have the worst prognosis relative to the patients

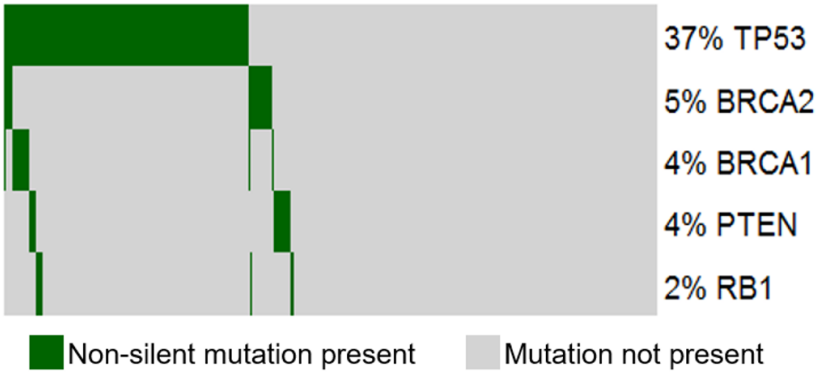


Figure 5.6: Mutual exclusivity across the characteristic genes of Component 1

with any other known breast cancer sub-type. Component 6 (shown in Figure 5.5C) corresponds to the sub-type of the progesterone receptor and estrogen receptor alpha-positive (PR+ ER+) cancers, that are responsive to anti-ER therapies. Based on the associated clinical markers, Component 9 (shown in Figure 5.5D) appears to be related to the known sub-type of breast cancer that is driven by over-expression of the epidermal growth factor receptor oncogene (HER2) and responsive to HER2-targeted inhibitors.

Further analysis using Ingenuity Systems software [202] to test Pathway Enrichment of putative cancer driver genes, potentially activated or inactivated by mutations associated with the TNBC-related fifth component identified significant enrichment of genes with a role in signaling networks that promote the function of cancer stem-like cells (CSCs), i.e., downstream of transcription factor TWIST1, and alternative mRNA splicing, i.e., downstream of serine and arginine-rich splicing factor SRSF2. CSCs have been identified in patient TNBC tumors as a fraction of self-renewing, tumor-initiating cancer cells that also give rise to drug resistance and metastatic recurrence [75, 222]. Alternative mRNA splicing has also been implicated in maintaining and generating CSCs [15].

We then compared the subtypes derived by CLIGEN with the well-established PAM50 breast cancer subtypes first introduced by Perou et al. [269] that are outlined above in Section 5.5.1. To do this, we used expression status of the biomarkers ER, PR and HER2, which are in large part available for each sample in TCGA and are acceptable surrogates

to characterize the basal, HER2 enriched, Luminal A and Luminal B breast cancer subtypes [257, 280]. Patients that do not have values for any of the three biomarkers were labeled as 'Unassigned' subtype[2]. To answer the question, "How CLIGEN subtypes compare to these established subtypes?" we performed two analyses. First, we examined patient factors associated with each CLIGEN component. Second, we examined the clinical factors associated with each CLIGEN component. Table 5.6 describes the convention that we used for BC sub-type classification and Figure 5.7 shows the distribution of patients in TCGA dataset across the PAM50 sub-types.

Analysis of patient factors associated with CLIGEN components

In this analysis, we assigned each patient to a unique CLIGEN component (the component that corresponds to the highest score in the patient's factor matrix obtained by CP decomposition of the CLIGEN binary tensor). We also assigned each patient to the PAM50 subtype based on the patient's clinical markers. Table 5.5 shows the proportions of patients assigned to each combination of CLIGEN component and PAM50 subtype.

From Table 5.5, we observe that patients assigned to Component 1 are also mostly associated with Luminal B and, to a smaller extent, with Luminal A PAM50 sub-type. Patients assigned to Component 5 include only TNBC patients, patients assigned to Component 6 are split between Luminal A and TNBC, and patients assigned to CLIGEN Components 2,3, 7-10 are associated with all PAM50 sub-types to some extent.

We also observe that patients belonging to PAM50 sub-types were not also exclusively assigned to a single CLIGEN component. Furthermore, components with higher enrichment match had less than 50% of patients assigned to them. 21% of the patients assigned to the HER2-enriched PAM sub-type were also assigned to Component 3, 29% of Luminal B patients were also assigned to Component 7, 34% of Luminal A patients to Component 2 and 26% of TNBC patients to Component 4. Majority of patients, who could not be assigned

[2]Note that Normal-like BC sub-type was not considered as a subtype here because the only difference between normal-like and luminal A is that normal-like patients have low levels of protein Ki-67, and our dataset did not have that status nor reference point to compute these levels.

to any PAM50 sub-type, are also assigned to CLIGEN components 8 and 10, which reveals that our method did not simply recapitulate the same sub-types that have been exposed by Perou et al. [269], but it elucidates a different way to sub-type BC patients. Researchers could study and learn from clustering tumors in this different way.

Comp.	HER2	Lum. B	Lum. A	Basal	Unassigned	Max
1	0	0.05 (3)	0.01 (3)	0.01 (1)	0	**0.05**
2	0.16 (3)	0.22 (12)	0.34 (105)	0.06 (5)	0.26 (5)	**0.34**
3	0.21 (4)	0.09 (5)	0.03 (9)	0.18 (15)	0.05 (1)	**0.21**
4	0	0.02 (1)	0.06 (18)	0.26 (22)	0	**0.26**
5	0	0	0	0.06 (5)	0	**0.06**
6	0	0	0.03 (9)	0.06 (5)	0	**0.06**
7	0.21 (4)	0.29 (16)	0.23 (70)	0.12 (10)	0.21 (4)	**0.29**
8	0.16 (3)	0.18 (10)	0.23 (70)	0.07 (6)	0.26 (5)	**0.26**
9	0.16 (3)	0.13 (7)	0.05 (15)	0.08 (7)	0.11 (2)	**0.16**
10	0.11 (2)	0.02 (1)	0.02 (6)	0.1 (8)	0.11 (2)	**0.11**
Sum	**1.0 (19)**	**1.0 (55)**	**1.0 (305)**	**1.0 (84)**	**1.0 (19)**	

Table 5.5: Distribution of TCGA patients assigned to different PAM50 suptypes across CLIGEN components. The column Comp. indicates the component number. Each cell shows the proportion of patients assigned to a particular CLIGEN component and PAM50 subtype, and the patient count is shown in parenthesis.

Analysis of clinical factors associated with CLIGEN components

Here we compare the assignments to the most prevalent clinical variable in the clinical factor and to the known sub-types of BC. To characterize the known sub-types, we first found the distribution of the three molecular biomarkers on the Clinical Factor **C** across the ten components, see these distributions in Table 5.7. After getting the raw values for the biomarkers, we standardized the values and summarized them in Table 5.8. Since the variables appear in more than one component, we assigned the variables to the component with the highest score. For example, the clinical variable HER2 had its highest score on the Component 2 vector compared with the score on the other Components; therefore, the equivocal status of HER2 was matched with Component 2. Equally, PR.StatusIndeterminate with Component 2, HER2.StatusNegative with Component 6, PR.StatusPositive with Component 6, HER2.StatusPositive, PAM50.sub-typeHER2.enriched, ER.StatusNegative, ER.StatusPositive,

and PR.StatusNegative with Component 9. From the Clinical Factor matrix the only clear groups are for HER2 positive, HER2 negative and PR negative as shown in Table 5.8.

Breast Cancer molecular subtype	Biomarkers
Basal-like	HER2-, ER-, PR-
HER2	HER2+, ER-, PR-
Luminal A	HER2- and ER+ and/or PR+
Luminal B	HER2+ and ER+ and/or PR+
Unassigned	Missing any HER2, ER, or PR

Table 5.6: Four molecular sub-types of breast cancer and definition of 'Unassigned' sub-type.

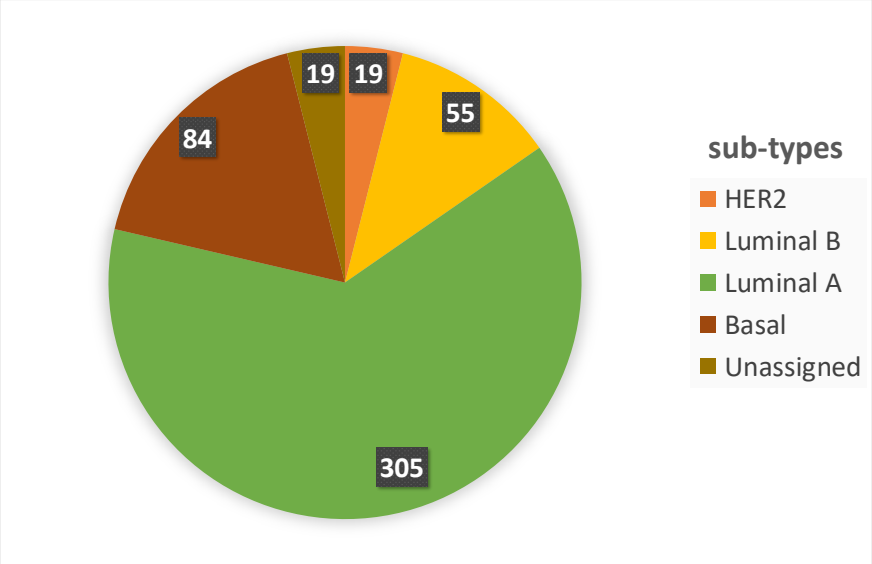


Figure 5.7: Distribution of molecular sub-types.

5.6 Degree of Overlap between CLIGEN and PAM50 Breast Cancer sub-types

Based on the analyses of patient, genetic and clinical factor matrices, we found that Components 1 and 7 correspond to Luminal B breast cancer, while more than one third of Luminal A patients are mapped to Component 2. From the genetic point of view, Components 4, 5 and 6 appear to be different sub-types of triple-negative breast cancer (TNBC), which was corroborated by our analysis of patient factor matrix obtained by CLIGEN. Component 6 corresponds to the subtype of the progesterone-receptor-positive and estrogen-receptor-positive (PR+ ER+) cancers based on the genetic factor matrix obtained

C	#	HER2	ER	PR
1	0			
2	67	Unassigned, Negative, Positive	Unassigned, Negative, Positive	Unassigned,
3	5		Unassigned	Unassigned
4	10		Unassigned	Unassigned
5	9		Unassigned	Unassigned
6	66	Unassigned, Negative, Positive	Unassigned, Negative, Positive	Unassigned, Negative, Positive
7	2			
8	2		Unassigned	
9	64	Unassigned, Negative, Positive	Unassigned, Negative, Positive	Unassigned, Negative, Positive
10	5		Indeterminate, Performed but Not Available	Performed but Not Available

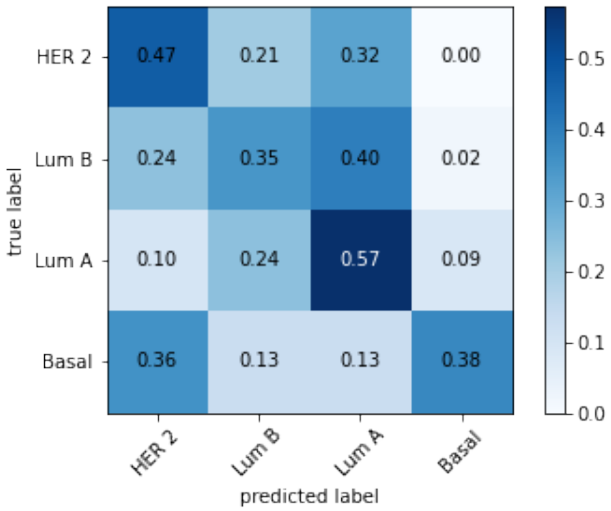
Table 5.7: Distribution of biomarkers' values among the 10 components. The **C** column indicates the component number. The # column indicates the number of clinical variables that have non-zero value in the row component.

C	HER2	ER	PR	Molecular Subtype
1	Equivocal			Unassigned
6	Negative		Positive	Luminal HER-
9	Positive	Negative, Positive	Negative	HER2 and Luminal HER2+

Table 5.8: Summary conclusive components.

by CLIGEN. From the list of patients, this component contains patients from Luminal A and Basal breast cancers. Based on the clinical factor, this component was associated with a Luminal A BC. Component 3 and 9 appear to be related to the HER2-enriched PAM50 subtype based on the genetic profile and the clinical variables. To establish the degree of overlap between GLIGEN and PAM50 breast cancer sub-types, we grouped the patients in CLIGEN components into their closest inferred PAM50 sub-types according to the mappings outlined above. Table 5.8 illustrates the confusion matrix between the PAM50 molecular sub-types of patients predicted based on the analysis of CLIGEN components and the actual PAM50 molecular sub-types of patients determined based on their clinical variables. We did not include in this analysis the patients that could not be assigned to a PAM50 subtype based

on their available clinical data (column "Unassigned" in Table 5.5). It follows from this table that overall the actual PAM50 sub-types can be inferred from the analysis of CLIGEN components with 50.76% accuracy. Luminal A PAM50 subtype was inferred with the highest accuracy of 57.38%, followed by HER2-enriched (47.37%). The greatest confusion was in distinguishing Luminal A from Luminal B as well as HER2-enriched from Basal PAM50 sub-types.



		Predicted label			
		HER2	**Lum B**	**Lum A**	**Basal**
True label	**HER2**	9 (0.47)	4 (0.21)	6 (0.32)	0 (0.0)
	Lum B	13 (0.24)	19 (0.35)	22 (0.40)	1 (0.02)
	Lum A	30 (0.1)	73 (0.24)	175 (0.57)	27 (0.09)
	Basal	30 (0.36)	11 (0.13)	11 (0.13)	32 (0.38)

Figure 5.8: Confusion matrix between the actual PAM50 patient sub-types and PAM50 sub-types predicted based on the analysis of CLIGEN components. Fractions of patients with predicted PAM50 subtype that belong to the corresponding gold standard PAM50 sub-type are shown in parentheses.

5.6 Implications for Breast Cancer Treatment

From the sub-types obtained by CLIGEN, we paid particular attention to Component 1, which had non-clinical variables associated with and a high abundance of mutually exclusive somatic mutations and mapping from CLIGEN to PAM50 breast cancer sub-types.

These analyses are important because they are in-line with ongoing research reported in the literature that a subset of patients could benefit from new therapies and specifically from immunotherapy (see Section 5.5.3), since the high mutational load is a predictor of positive response to immunotherapy [244].

It is known that the accumulation of somatic mutations is a hallmark of tumors, but the mutational burden varies dramatically among tumor types [208]. These dramatic differences in mutation burden reflect significant differences in the balance of DNA damage exposure and DNA repair fidelity among tumors. It has been shown that tumor mutational load was significantly higher in patients who achieved long-term clinical benefit and more prolonged progression-free survival from immune checkpoint blockade (ICB) therapy compared with those who had minimal benefit. This association was confirmed by two studies in patients with metastatic melanoma treated with the CTLA-4-blocking antibodies ipilimumab [311, 338] and patients with non-small cell lung cancer treated with the anti-PD-1 antibody pembrolizumab [282]. The association between mutational burden and immunotherapy response has now been observed in many cancers; however, it is becoming clear that high mutational burden alone is not sufficient to drive immunotherapy response, and there is no definitive threshold mutational burden that separates ICB responders from non-responders [244]. Our results highlight the need for further biological studies to identify a threshold for BC mutational burden and additional factors that can benefit BC immunotherapy.

5.7 Conclusion

In this chapter, we introduced *CLIGEN*, a novel machine learning based pipeline for unsupervised disease subtype discovery based on tensor decomposition of a three-dimensional tensor combining clinical and somatic mutation patient data, and applied the proposed pipeline to breast cancer subtyping. Quantitative evaluation of the discovered breast cancer sub-types indicates that representation of clinical and genetic patient data as a tensor and its subsequent decomposition is an effective computational approach to high-throughput disease

subtyping for precision medicine. In particular, our proposed pipeline was not only able to identify known breast cancer sub-types (HER2+ and ER+), but also elucidated new possible characteristics of a complex breast cancer subtype (triple negative), which provides an opportunity for further research to define new cancer sub-types. We also demonstrated that patient membership proportions in the discovered breast cancer sub-types are more effective predictors of breast cancer survival than patient membership proportions in computationally identified genotypes and phenotypes.

CHAPTER 6 CANCER SUBTYPING BASED ON CLINICAL, SOMATIC MUTATION AND GENE EXPRESSION DATA

It has long been understood that stratification of patients into fine-grained cohorts corresponding to disease sub-types is a foundation of accurate diagnosis and personalized cancer treatment. Nevertheless, current disease sub-typing methods have not addressed the integration of genetic and clinical data for obtaining comprehensive sub-types yet. In this chapter, we propose TGENEX, a new computational pipeline for high-throughput data-driven stratification of cancer patients into cohorts corresponding to multi-modal cancer sub-types based on clinical and genomic data (mutation and gene expression data). We applied TGENEX to discover sub-types of seven different cancers using publicly available datasets from the Cancer Genome Atlas (TCGA). Quantitative evaluation of the sub-types discovered by TGENEX indicates that they are clinically meaningful and can provide insights for cancer patient survival prognosis to clinicians at the point of care. We conclude that enriching gene expression and somatic mutations with clinical data can elucidate novel cancer sub-types.

6.1 Introduction

It is well recognized that the personalized medicine approach is the key to effective cancer treatment [65]. Methods for patient stratification into cohorts corresponding to cancer sub-types and markers defining these sub-types are a central tenet of precision medicine approach to cancer treatment [291]. However, many cancers are not fully understood, and stratifying patients into cohorts that can predict patient survival is still an open problem.

Remarkable advances in next-generation sequencing technology coupled with the widespread adoption of electronic health records (EHR) by healthcare providers in the United States have enabled the collection of unprecedented amounts of genomic and clinical patient data. However, despite the immense progress in computational methods for analyzing clinical or genomic data, these methods in isolation cannot capture all aspects of the pathogenesis of cancers [284]. Furthermore, there has been relatively little research on computational

methods for joint analysis of clinical and genomic data for cancer sub-typing. Topological analysis of the patient-patient similarity network has been applied to stratify patients with type II diabetes [215] and traumatic brain injury [253] based on their pathoanatomical and molecular data. As opposed to the proposed pipeline, which is implemented end-to-end and has minimum hardware requirements, topological analysis requires defining patient similarity lenses and requires the use of a cloud-based supercomputer [253].

Cancer treatment decisions are often based on characteristics of a tumor (e.g., size, pathologic stage, histologic grade, hormonal receptor status, and lymphovascular invasion) or a patient (age, race, menopausal and performance status) [62]. These and other characteristics are utilized in cancer clinical decision support software, such as PREDICT Plus [355]. However, this software does not take into account the molecular characteristics of tumors. As a result, many cancer patients are over-treated by being exposed to the risk of toxic effects from adjuvant chemotherapy without deriving significant benefits from it [179]. Advances in high-throughput sequencing and micro-array technology enabled the utilization of genomic data for cancer patient sub-typing. Early research on cancer patient stratification primarily focused on gene expression profiles and have distinguished at least four molecularly distinct sub-types of breast cancer [313]. In transcriptomics, agglomerative hierarchical clustering (HC) is a frequently used approach for clustering genes or samples that show similar expression patterns [112, 5, 269]. HC provides for a structural view of the data that makes it appealing in exploratory data analysis. However, classical HC imposes a tree structure on the data that might not reflect the underlying structure and is highly sensitive to the metric used to assess similarity among elements. Divisive clustering methods, such as k-means [93, 333], global k-means [11], fuzzy modification of k-means [79, 125], have been applied for the same application. These methods provide clear cluster boundaries and tighter clusters, but they lack the visual appeal of HC. Another group of methods is neural network clustering, such as self-organizing maps (SOM) [198, 320, 138], Self-Organising Tree Algorithm (SOTA) [151], and Dynamically Growing Self-Organising Tree (DGSOT) [228]. Neural networks can be

modeled as a collection of nodes with weighted interconnections, which can be adaptively learned. The common drawbacks of both k-means based methods and neural networks-based methods is the need to specify the number of clusters beforehand.

Although methods analyzing single types of patient data have advanced the understanding of cancer [112, 5, 269], they are unable to capture the complex interactions among biomolecules, and the outcome sub-types are prone to be suboptimal [314]. A vast majority of the diseases develop differently, making them heterogeneous. Contemporary methods integrate genetic data for disease sub-typing and have shown that the discovered sub-types result in better survival models than the methods using gene expression or mutation data alone [347, 298, 239]. Cox log-rank test is one of the methods to decide if certain groups have different survival behavior or not. The improvement in prognosis due to the integration of different types of genetic data can be because two patients rarely have identical genotypes even though cancer is a genetic disease. Similarly, patients differ in their clinicopathological parameters, and integrating their clinical variables can benefit sub-typing and prognosis models. In particular, gene expression profiling has uncovered many differences between cancer and healthy cells and has enabled the definition of relevant disease sub-types based on 'biomarkers' that correlate with the clinical outcome without indicating causality [131]. On the other hand, genomic information aims to identify driver genes (also know as driver genes) [315, 344]. However, the vehicle by which driver mutations cause cancer is gene transcription acting through a complex cellular signaling circuitry that links the genomic variants to cancer. Many of the consequences of genetic alterations will affect gene expression in different ways, such as aberrant transcription, cell signaling, gene dosage, and epigenetic regulation. For this reason, studies using only gene expression or somatic mutation data have fundamental limitations due to the unknown genetic background of the samples.

In the previous chapter, we presented *CLIGEN* to analyze somatic mutation data and clinical data for disease sub-typing [86]. In order to overcome the variability of diagnos-

tic and prognostic predictors derived from genomic data alone, we propose to integrate it with demographic and clinical cancer patient data in a computational pipeline (named *TGENEX*) for fully unsupervised disease sub-typing based on clinical data, gene mutation and gene expression data. We chose gene expression and mutation profiles for the genetic vector because it has been demonstrated that driver mutations are correlated with general gene expression, and combining them improves outcome prediction for some cancers such as myelodysplastic syndromes (MDS) [131]. Our pipeline takes demographic information of a given cancer patient population as well as somatic point mutation profiles, gene expression, and clinical properties of their tumors as input and identifies a set of patient cohorts that share the same set of pathogenic gene mutations and cancer characteristics, with each identified cohort corresponding to a cancer multi-modal sub-type. Here, we hypothesize that the sub-types discovered by *TGENEX* capture co-occurrences, which allow us to gather insights surrounding molecular, demographic and clinical features of new cancer sub-types and refine the known ones as well as shed light on molecular aberrations in tumors that are correlated with gene expression and clinical outcomes. *TGENEX* consists of two main stages. In the first stage, multi-modal patient data that includes somatic mutation profiles, gene expression, and clinical data (clinical properties of tumors and cancer patient demographics) is represented as a three-mode tensor. In the second stage, non-negative tensor decomposition is applied to identify latent factors in each modality of the constructed tensor.

6.2 Dataset

For experiments in this chapter, we used patient data from The Cancer Genome Atlas - Genomic Data Commons Data Portal (TCGA GDC) [49] downloaded from RTCGA snapshot from 2016-01-28 [289]. More concretely, we use somatic mutation (non-silent mutation from the whole exome sequencing level 3 profiles, mRNA expression (level 3 Agilent g4502), and clinical data of seven cancer types: Breast invasive carcinoma (BC), Colon adenocarcinoma (COAD), Colorectal adenocarcinoma (COADREAD), Kidney renal clear cell carcinoma (KIRC), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarci-

noma (OV), and Pan-kidney cohort (KICH+KIRC+KIRP) (KIPAN). For all the datasets, we considered only the patients for whom somatic mutation, gene expression, and clinical data are available and discarded the patients with less than 10 somatic mutations. For example, for breast ductal carcinoma patients (BC), out of the 825 patients on TCGA, we considered only 493 patients because only these have somatic mutation, gene expression and clinical data available and discarded 4 patients with few somatic mutations. The resulting BC dataset consists of genetic profiles over 11,996 genes and 87 values and value ranges of 17 discrete and dichotomized continuous clinical variables of 489 patients. For colon adenocarcinoma (COAD), 12,256 genes were filtered out of a total of 17,558 present genes on 153 patients. For the kidney and renal clear cell carcinoma dataset (KIRC), 12,238 genes were filtered out of a total of 17,522 present genes in a cohort of 72 patients. For the lung squamous cell carcinoma dataset (LUSC), 12,226 genes were filtered out of a total of 17,510 present genes in a cohort of 154 patients. For the ovarian serous cystadenocarcinoma dataset (OV), 12,155 genes were filtered out of a total of 17,467 present genes in a cohort of 541 patients. In addition to these individual cancer datasets, we analyze the following pan-cancer dataset. For colorectal adenocarcinoma (COADREAD), 12,221 genes were filtered out of a total of 17,521 present genes in 222 patients. For Pan-kidney cohort (KIPAN), which includes kidney and renal papillary cell carcinoma, kidney and renal clear cell carcinoma, and kidney Chromophobe datasets, 11,976 genes were filtered out of a total of 17,260 present genes with 88 patients.

Genetic data

For BC, the somatic mutation table consists of 37 columns and 34032 registries. A row in this table indicates a mutation in the gene reported in the column "*Hugo Symbol*" for the sample in the column "*Tumor Sample Barcode*". Additional details on the processing and organization of these data are available in Koboldt et al. [49]. We constructed patient mutation profiles as binary vectors, in which a bit is set if the patient's gene corresponding to that position in the vector harbors a mutation. All the other somatic mutation tables

Dataset	**Name**	**Num. patie.**	**Num. genes**	**Num. fil. g.**
BC	Breast invasive carcinoma	489	17327	11996
COAD	Colon adenocarcinoma	153	17558	12256
COADREAD	Colorectal adenocarcinoma	222	17521	12221
KIRC	Kidney renal clear cell carcinoma	72	17522	12238
KIPAN	Pan-kidney cohort (KICH+KIRC+KIRP)	88	17260	11976
LUSC	Lung squamous cell carcinoma	154	17510	12226
OV	Ovarian serous cystadenocarcinoma	541	17467	12155

Table 6.1: Datasets mRNA from RTCGA. Num. patie. stands for the number of patients from the mRNA dataset. Num. genes stands for the number of genes in the mRNA dataset from a total of 17815. Num. fil. g. is the size of the final list of genes.

have 37 columns but a variable number of registers. See Table 6.1 the number of filtered genes that were obtained for each cancer type. Additionally, we obtained preprocessed level 3 mRNA expression Agilent g4502 with LOWESS (Locally Weighted Scatterplot Smoothing) normalization at gene level from the RTCGA snapshot [365, 61, 289]. We downloaded only samples that were taken from the primary tumor for both mutation and gene expression, neither metastatic nor healthy samples were considered.

Clinical and Demographic Variables

Similarly, we downloaded clinical data from several cancer types using RTCGA and selected clinical variables that are considered relevant for sub-typing based on the literature. [29, 57, 68, 82, 118, 121, 132, 152, 170, 248, 286, 306, 307, 309, 335, 337, 342, 352, 354]. The downloaded clinical tables have patients as rows and clinical variables as columns. Table 6.2 shows the dimensions of the raw clinical datasets downloaded from TCGA. To identify the variables that we could use for our analysis, we selected the clinical variables common across all the cancer types, which reduced the number of variables to 1,574. Then, we discarded eight columns with administrative data, one column with additional studies, 231 columns with details about drugs, 47 columns with details about follow-ups, four columns about new tumor events, 37 columns about radiation sessions, 61 columns related to the biospecimen, 544 variables related to sample1 portions, 321 variables related to sample2

portions, 268 variables related to sample3 portions, three identification variables, and the variable informed consent. Next, we discarded 28 variables that were missing for more than 40% of patients. After filtering out all these categories of clinical variables, we ended up with a list of 20 clinical variables. Among these, we have four variables for survival analysis: days-to-death, vital-status, days-to-last-followup, and days-to-last-known-alive; therefore, we removed these variables from our input. Finally, we use 17 discrete and continuous clinical and demographic variables relevant to cancer (including one variable to identify the sample), which we used as an input of our model after dichotomizing[1] them into 87 variables (without loss of generality, further referred to as clinical variables).

Here we present the final list of clinical variables with some literature that evidences their relevance to cancer. 1) bcr-patient-barcode: Patient's barcode from The Biospecimen Core Resource, 2) age-at-initial-pathologic-diagnosis (evidence: [82]), 3) sex (evidence: [68]), 4) race (evidence: [352, 121]), 5) ethnicity (evidence: [352, 342]), 6) cqcf.country: Patient's country from the Case Quality Control Form (CQCF). (evidence: [337, 132, 57]), 7) cqcf.prior-dx: Patient's prior diagnosis of cancer from the CQCF. (evidence: [248, 335]), 8) cqcf.tumor-type: Patient's tumor-type from the CQCF. (evidence: [306, 170]), 9) cqcf.normal-tissue-anatomic-site: Anatomic site of patient's normal tissue from the CQCF (evidence: [29, 354]), 10) tumor-tissue-site (evidence: [152]), 11) patient-canonical-status (evidence: [152]), 12) person-neoplasm-cancer-status (evidence: [286]), 13) site-of-disease (evidence: [309, 307]), 14) normal-tissue-proximity (evidence: [118]), 15) drugs measure of response with values 'clinical progressive disease', 'complete response', 'partial response', and 'stable disease'[2] (evidence: [152]), 16) radiation-therapy with values 'yes' or 'not' (evidence: [152]), and 17) sample-type (evidence: [152]).

Dataset	Name	Nr clinical vars
BC	Breast invasive carcinoma	2129
COAD	Colon adenocarcinoma	3166
COADREAD	Colorectal adenocarcinoma	3509
KIRC	Kidney renal clear cell carcinoma	2809
KIPAN	Pan-kidney cohort (KICH+KIRC+KIRP)	2875
LUSC	Lung squamous cell carcinoma	2698
OV	Ovarian serous cystadenocarcinoma	1440

Table 6.2: Number of clinical variables for each Cancer from RTCGA.

6.3 Methods

An overview of *TGENEX*, our proposed pipeline for unsupervised sub-typing of complex diseases[3], is shown in Figure 6.1. The input to the pipeline consists of genetic and clinical data of cancer patients. The output of the pipeline is a set of definitions of cancer sub-types characterized by genetic and clinical variables. The pipeline consists of the two stages: i) data pre-processing and tensor construction and ii) non-negative decomposition of the constructed tensor to obtain cancer sub-types. Detailed descriptions of these stages are provided below.

6.3 Data representation

The first stage of the proposed pipeline illustrated in Figure 6.1.a involves pre-processing the input genomic and clinical patient data to create a combined representation for subsequent analysis. Somatic mutation datasets typically take the form of mutation tables, in which the rows correspond to mutations, and the columns describe the type of each mutation and its location (see Figure 6.1.a.i). Based on the input mutation table, *TGENEX* constructs a genetic matrix M with patients as rows and genes as columns. Normalized mRNA expression data is represented in the matrix R with patients as rows and genes as

[1]converting continuous and categorical data to binary values (two groups) which is a common approach in clinical research [78]

[2]A record is marked as 'complete response' when all target tumors have disappeared during treatment, 'partial response' when largest tumors have decreased for at least 30%, 'stable disease' when target tumors have no decreased its size, and 'progressive disease' when targeted lesions have increased for at least 20%

[3]source code for TGENEX is publicly available at `https://github.com/teanalab/TGENEX`

Symbol	Definition
M, R, V	somatic mutation, gene expression and clinical matrices
M_{ij}	cell of mutation matrix for patient i and gene j
R_{ij}	cell of mRNA matrix for patient i and gene j
V_{ij}	cell of clinical matrix for patient i and k-th value or interval of clinical variables
$\boldsymbol{\tau}$	multi-modal *TGENEX* tensor
G_{ij}	cell of genetic matrix (mutation and gene expression) for patient i and gene j
$\|P\|, \|G\|, \|C\|$	number of patients, genes as well as intervals and values of clinical variables
i, j, k	indices of patients, genes, and clinical variables.
t_{ijk}	value of tensor cell for the i-th patient, j-th gene, and k-th value or interval of clinical variables
$\otimes$	outer product of two vectors
R	number of tensor components
$\mathfrak{s}_r$	r-th rank-one component i.e. a candidate Sub-type definition
p_r, g_r, c_r	patient, gene and clinical factor vectors

Table 6.3: List of notations used in this chapter and their definitions

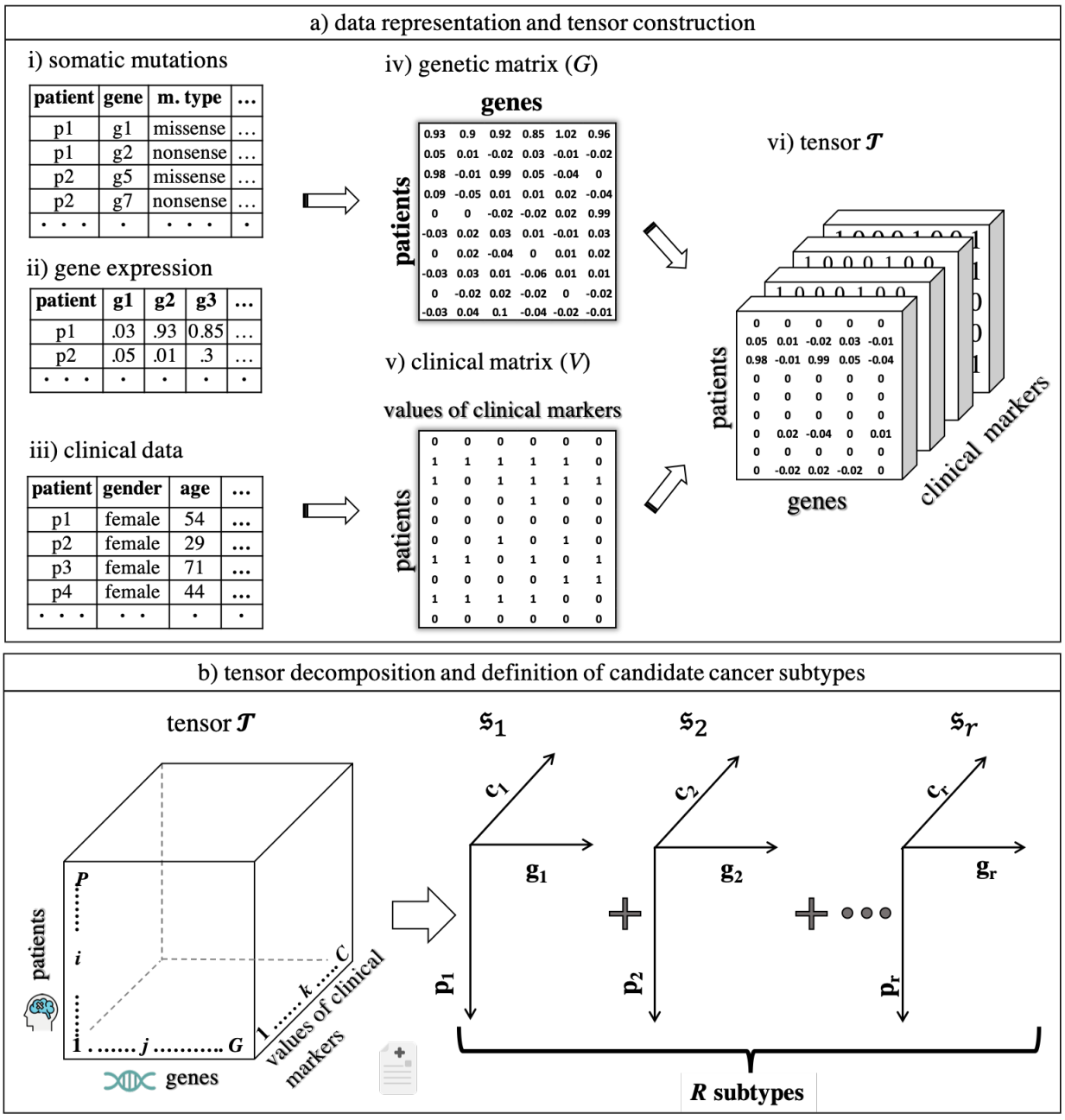


Figure 6.1: Two stages of the proposed *TGENEX* pipeline: a) data representation and construction of three-modal tensor τ. b) obtaining candidate sub-types via CP decomposition of tensor τ.

columns (see Figure 6.1.a.ii). To construct the genetic matrix G (see Figure 6.1.a.iv), we assign the gene expression value to a cell if there is a mutation for that particular gene and patient, i.e. $G = M \circ R$. With this representation, we can capture both gene expres-

sion abundance and the presence of mutation per each patient. In contrast with CLIGEN, the genetic matrix here is not binary but has gene expression values of the mutated genes. Formally, the element-wise multiplication ($\circ$ or pointwise) between the mRNA expression matrix R and the mutation matrix M requires the matrices to have the same dimensions (same number of samples and the same number of genes), and it multiplies mRNA and mutation values per each patient-gene pair. From the biological perspective, this operation captures the co-occurrence of mutation and gene expression.

Continuous clinical variables, such as the age of diagnosis, were discretized into intervals with a total of n combined intervals of all continuous variables and levels of all discrete variables in the entire dataset. The values of clinical variables for each patient were represented as a binary clinical matrix V with patients as rows and intervals of continuous or levels of discrete clinical variables as columns. A value of one in the cell V_{ik} of matrix V indicates that patient i has a *k-th* level or interval of a discrete or continuous clinical variable.

Tensor Construction

Here we present how the clinical V and genetic G matrices are combined to create a three-dimensional tensor (i.e. multidimensional array) $\boldsymbol{\tau} \in \mathbb{R}^{P \times G \times C}$ which captures interactions between clinical variables and genetic data, see Table 6.3. The first mode of tensor $\boldsymbol{\tau}$ corresponds to patients P , while the other two modes correspond to the values of clinical variables V and genes G (Figure 6.1.b). A value of the cell t_{ijk} of the tensor $\boldsymbol{\tau}$ has the gene expression value for gene j, if the $i-th$ patient has at least one mutation in gene j and $k-th$ value of clinical variables, otherwise it is set to zero.

Slicing tensor $\boldsymbol{\tau}$ along each of its three modes yields the following views:

1. *Patient mode*: each slice is a matrix of co-occurrences of gene patterns and clinical variables for a particular patient.

2. *Gene mode*: each slice is a matrix with patients as rows and clinical variables as columns, which shows how a gene signatures are correlated with clinical variables in

different patients.

3. *Clinical mode*: each slice is a matrix with patients as rows and genes as columns, which shows how gene mutations in different patients are correlated with a particular clinical variable.

Obtaining Candidate Sub-type Definitions Through Tensor Factorization

Tensor decomposition [199] is a powerful mathematical technique that has been successfully applied in different domains ranging from psychology and neuroscience to computer vision. Tensor decomposition has several advantages over matrix factorization. First, tensors explicitly account for the multiway structure of the data that is otherwise lost, when a tensor is converted into a matrix by collapsing some of its modes. Second, some tensor decomposition methods guarantee the uniqueness of the optimal solution even for very sparse tensors. The two most widely used tensor decomposition methods are the Tucker method and CANDECOMP/PARAFAC (CP), which stands for Canonical Decomposition (CANDECOMP) and Parallel Factor Analysis (PARAFAC) or CP tensor factorization [199]. CP decomposition can be considered as a particular case of the Tucker decomposition when the size of each modality of the core array is the same, and the only non-zero elements in the core are the elements along the main diagonal [199]. An essential property of CP decomposition is that the restriction imposed on the Tucker core leads to the uniqueness of the optimal solution [199].

Here, we use CP factorization to identify disease sub-types as groups of latent factors in $\boldsymbol{\tau}$. CP decomposition approximates $\boldsymbol{\tau}$ with $\hat{\boldsymbol{\tau}}$, a linear combination of rank-one tensors. Formally:

$$\boldsymbol{\tau} \approx \hat{\boldsymbol{\tau}} = [\![\lambda;\ \mathbf{P},\ \mathbf{G},\ \mathbf{C}]\!] = \sum_{r=1}^{R} \lambda_r \cdot \mathcal{S}_r = \sum_{r=1}^{R} \lambda_r \cdot \mathbf{p}_r \otimes \mathbf{g}_r \otimes \mathbf{c}_r \tag{6.1}$$

where R is the number of component rank-one tensors s_r that $\boldsymbol{\tau}$ is decomposed into, $\lambda_r \in \mathbb{R}$ is the weight of the *r-th* rank one tensor. Each sr is an outer product ($\otimes$) of partient

$p_r \in \mathbb{R}^P$, gene $g_r \in \mathbb{R}^G$ and clinical variable $c_r \in \mathbb{R}^C$ latent factors. Patient, gene and clinical vectors for all tensor components correspond to the columns of the patient, gene and clinical factor matrices **P, G, C**.

CP decomposition is obtained by solving the following optimization problem:

$$\min_{\hat{\tau}} \|\tau - \hat{\tau}\|_{\mathcal{F}} \tag{6.2}$$

aimed at finding the best approximation $\hat{\tau}$ of each element t_{ijk} of the original tensor τ from the factor vectors corresponding to the components tensors as follows:

$$t_{ijk} \approx \sum_{r=1}^{R} \lambda_r p_{ir} g_{jr} c_{kr} \tag{6.3}$$

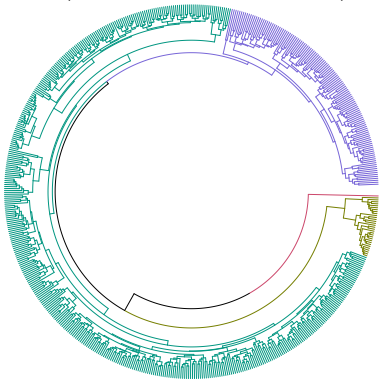
Sub-type definitions are constructed from the patient, gene, and clinical factor vectors corresponding to each of the component tensors obtained by CP decomposition of the *TGENEX* tensor. Each element in the vectors corresponding to gene and clinical tensor modes can be interpreted as the importance of a gene or clinical variable to a sub-type. After finding the sub-types, we perform survival analysis, which is essential for personalized cancer treatment [291].

6.4 Results and Discussion

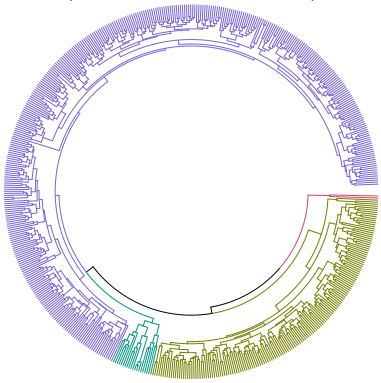
We assess the performance of our method for disease sub-typing by analyzing the seven cancer datasets described in Section 6.2. After obtaining the factor matrices with our method, we performed hierarchical clustering of the patient factor matrix. To decide the number of clusters (k) that generate the more distinctive sub-types, we performed survival analysis by estimating Cox log-rank p-value using the survival information downloaded from RTCGA. This Cox p-value represents how likely the survival curves' difference is observed by chance. We performed 8 experiments per each disease with the number of clusters varying from 3 to 10, i.e. $k = [3 - 10]$. To validate our results, we clustered the raw mRNA data (described in section 6.2) using hierarchical clustering varying the number of clusters from

3 to 10 also. Table 6.4 shows a dendrogram per each cancer to visualize each hierarchical clustering.

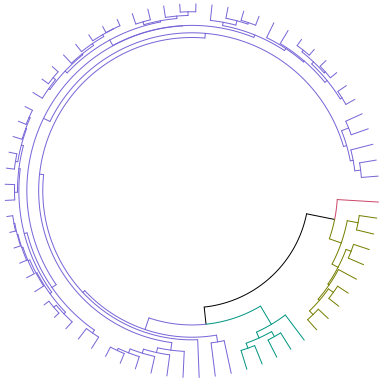
BC (nr. p = 526, k = 4)



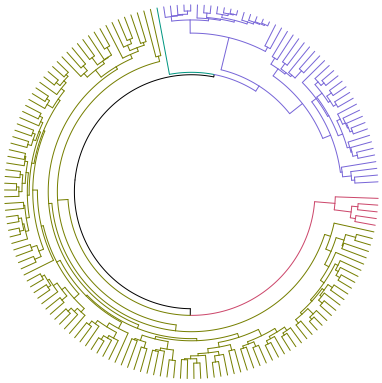
OV (nr. p = 541, k=4)



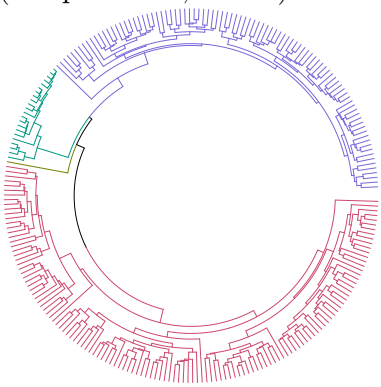
KIRC (nr. p = 27, k=3)



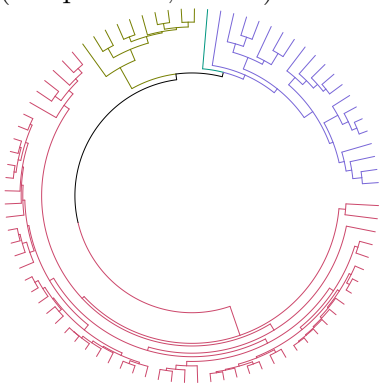
COAD (nr. p = 153, k=3)



COADREAD
(nr. p = 222 , k = 3)



KIPAN
(nr. p = 88 , k = 3)



LUSC
(nr. p = 154 , k = 3)

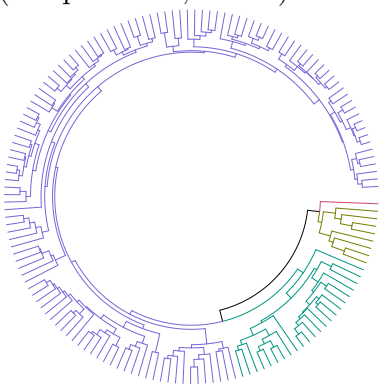


Table 6.4: Dendrograms of hierarchical clustering of patients on the gene expression matrix of the seven cancers. In parenthesis, the number of patients (nr. p) per each cancer and the number of clusters that performed the best overall clustering methods (k).

Tables 6.5, 6.6, 6.7, 6.8, 6.9, 6.10, 6.11, and 6.12 show the Kaplan-Meier survival plots

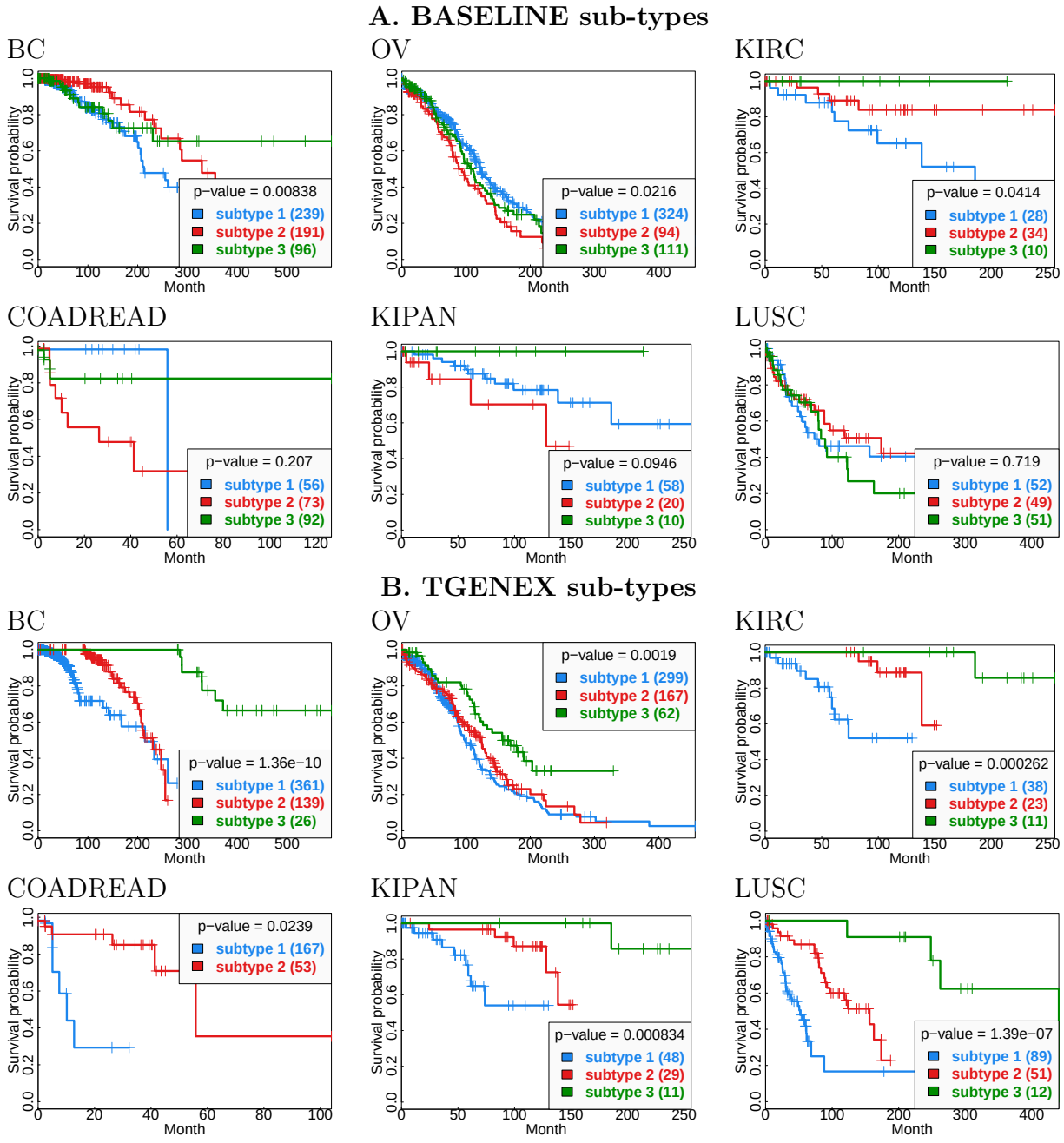


Table 6.5: Kaplan Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=3. (A) Baseline sub-types obtained with hierarchical clustering of mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=3.

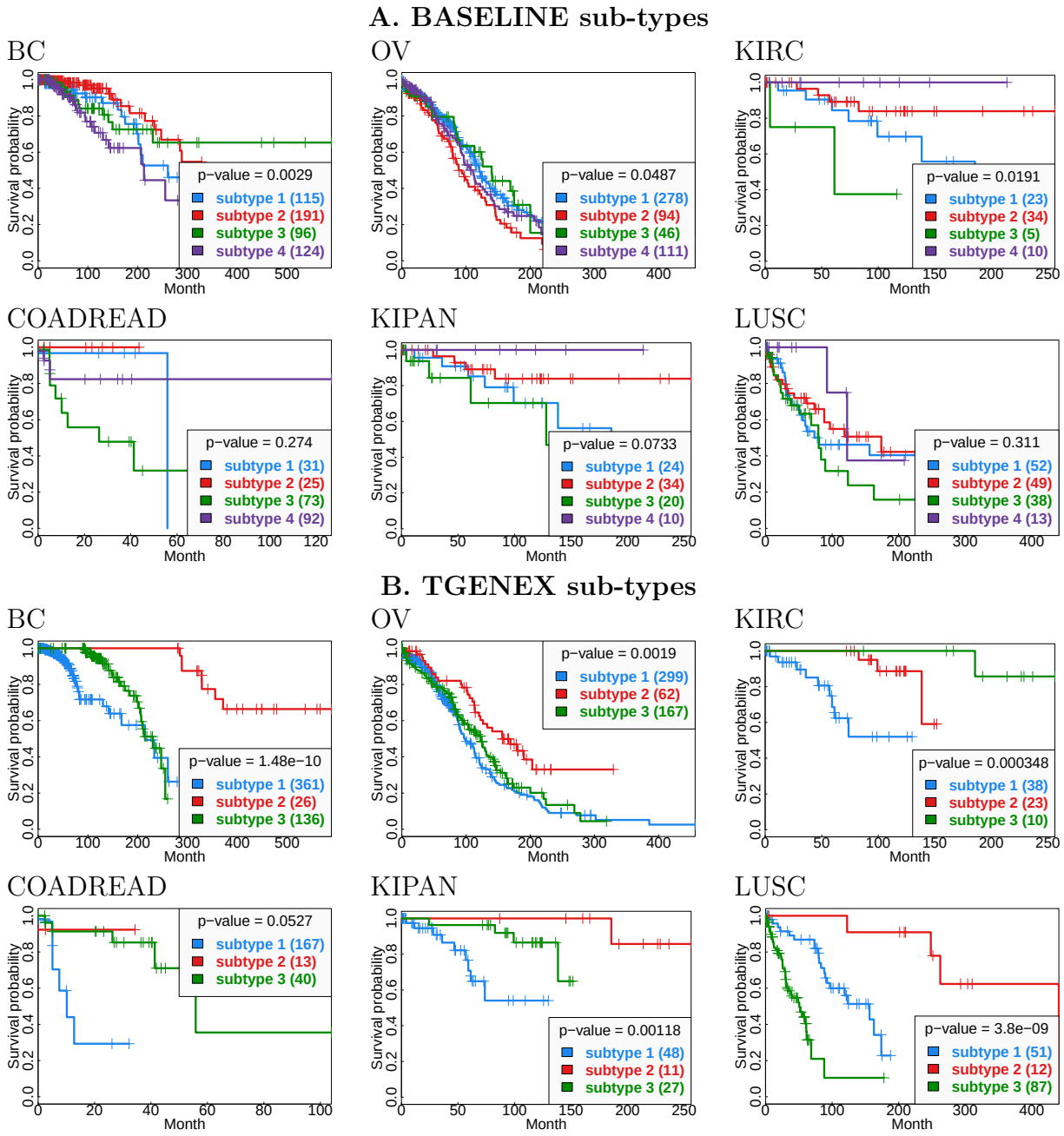


Table 6.6: Kaplan Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=4. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=4.

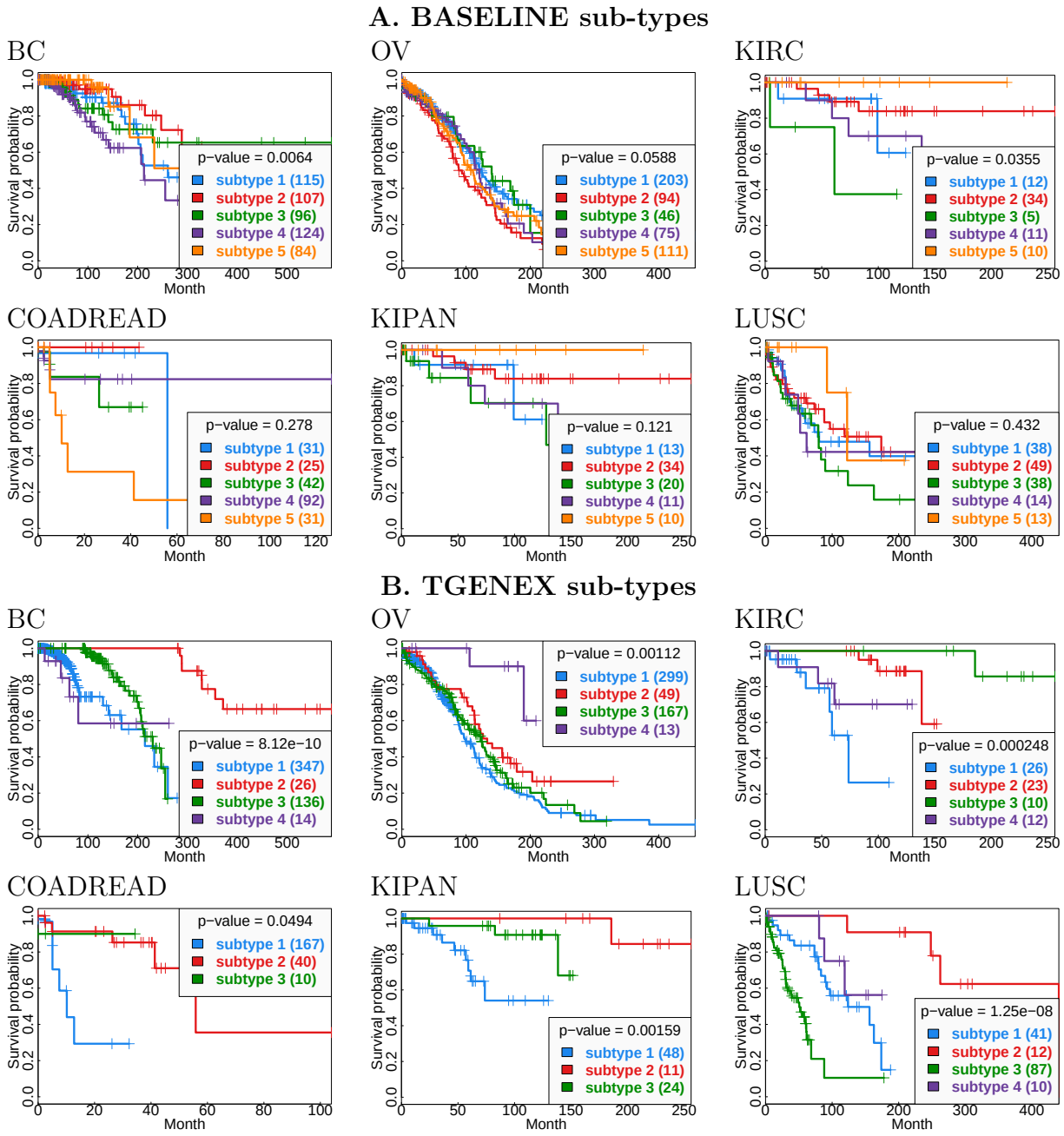


Table 6.7: Kaplan Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=5. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=5.

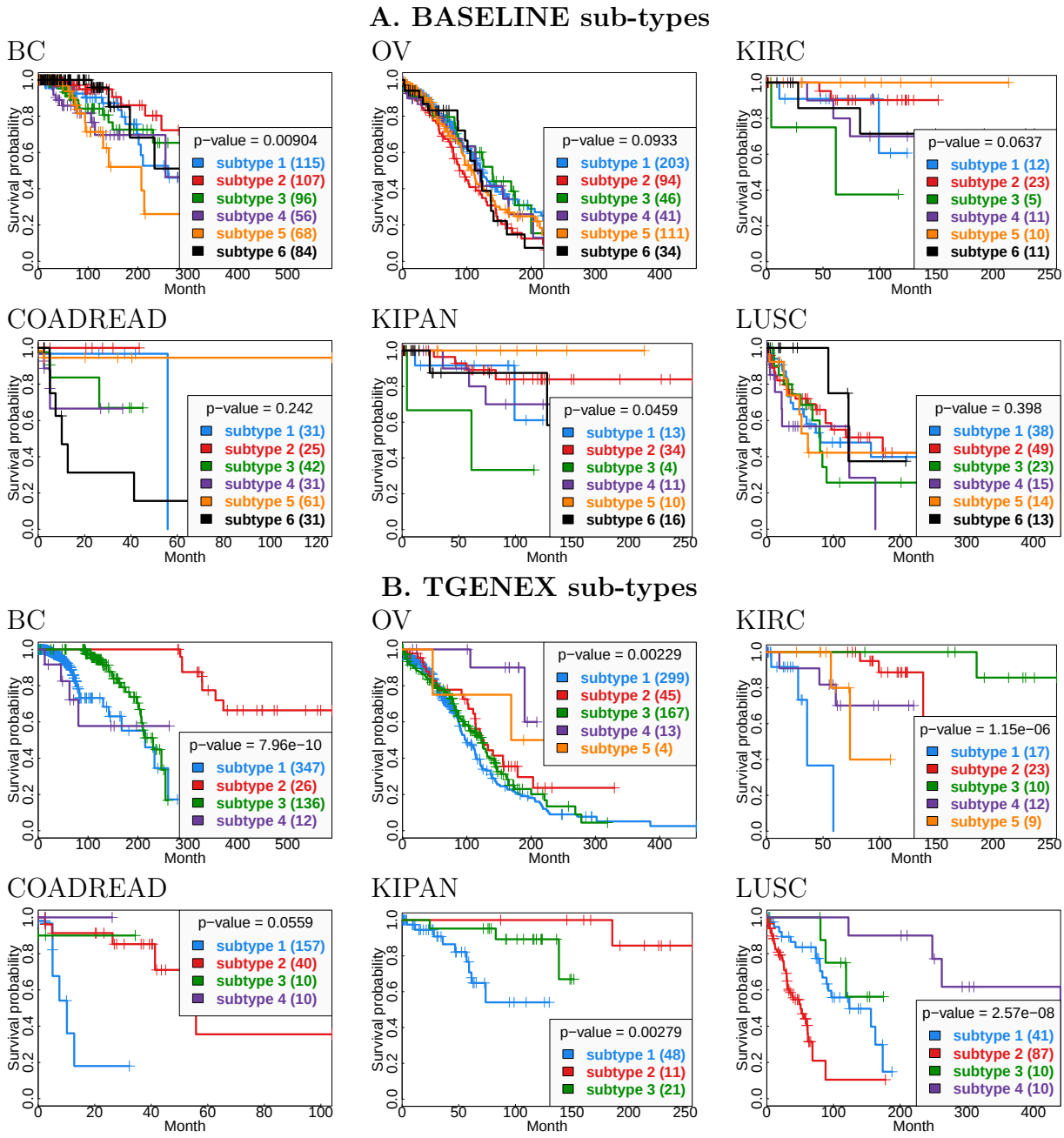


Table 6.8: Kaplan Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=6. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=6.

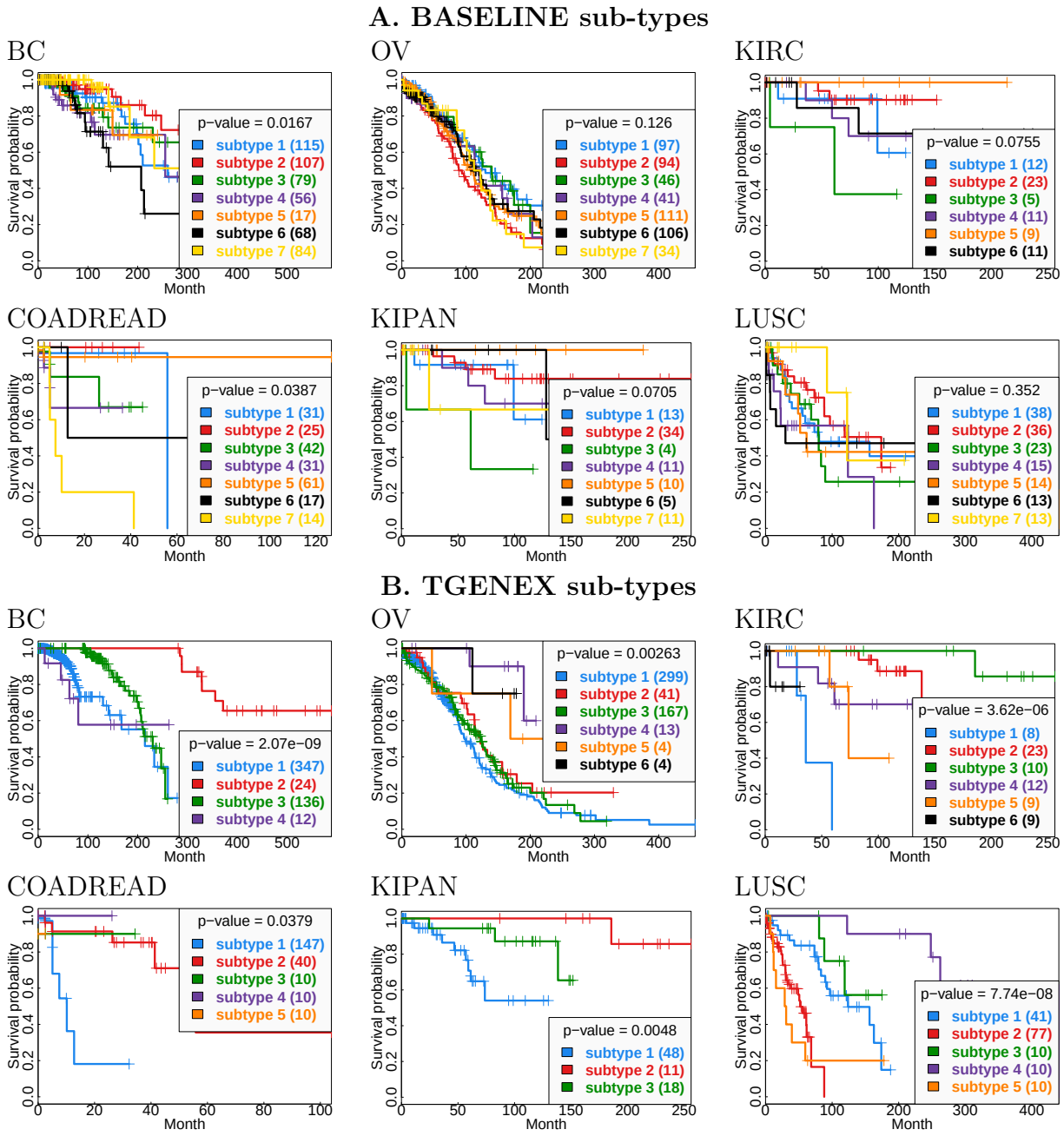


Table 6.9: Kaplan-Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=7. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=7.

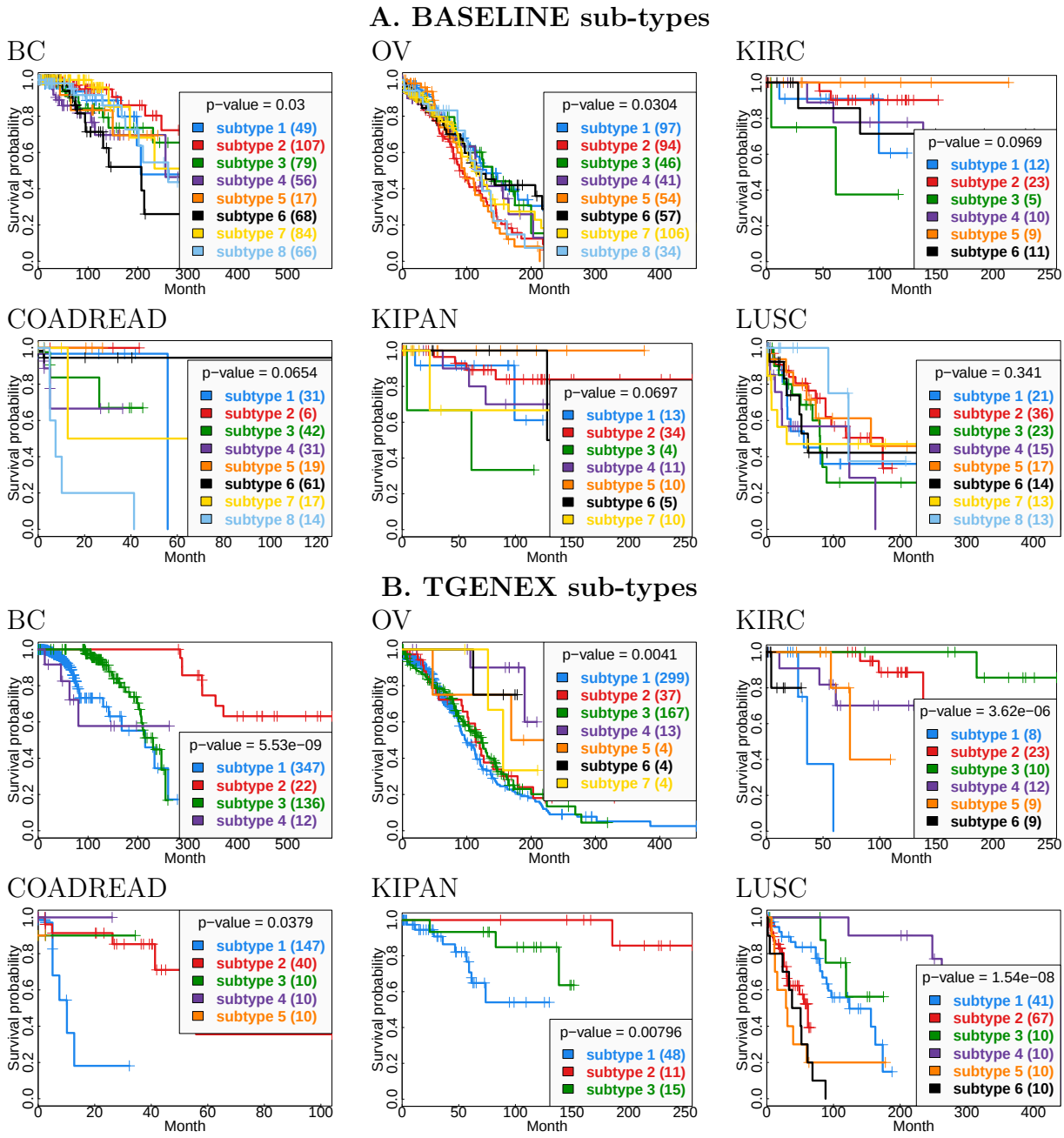


Table 6.10: Kaplan-Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=8. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=8.

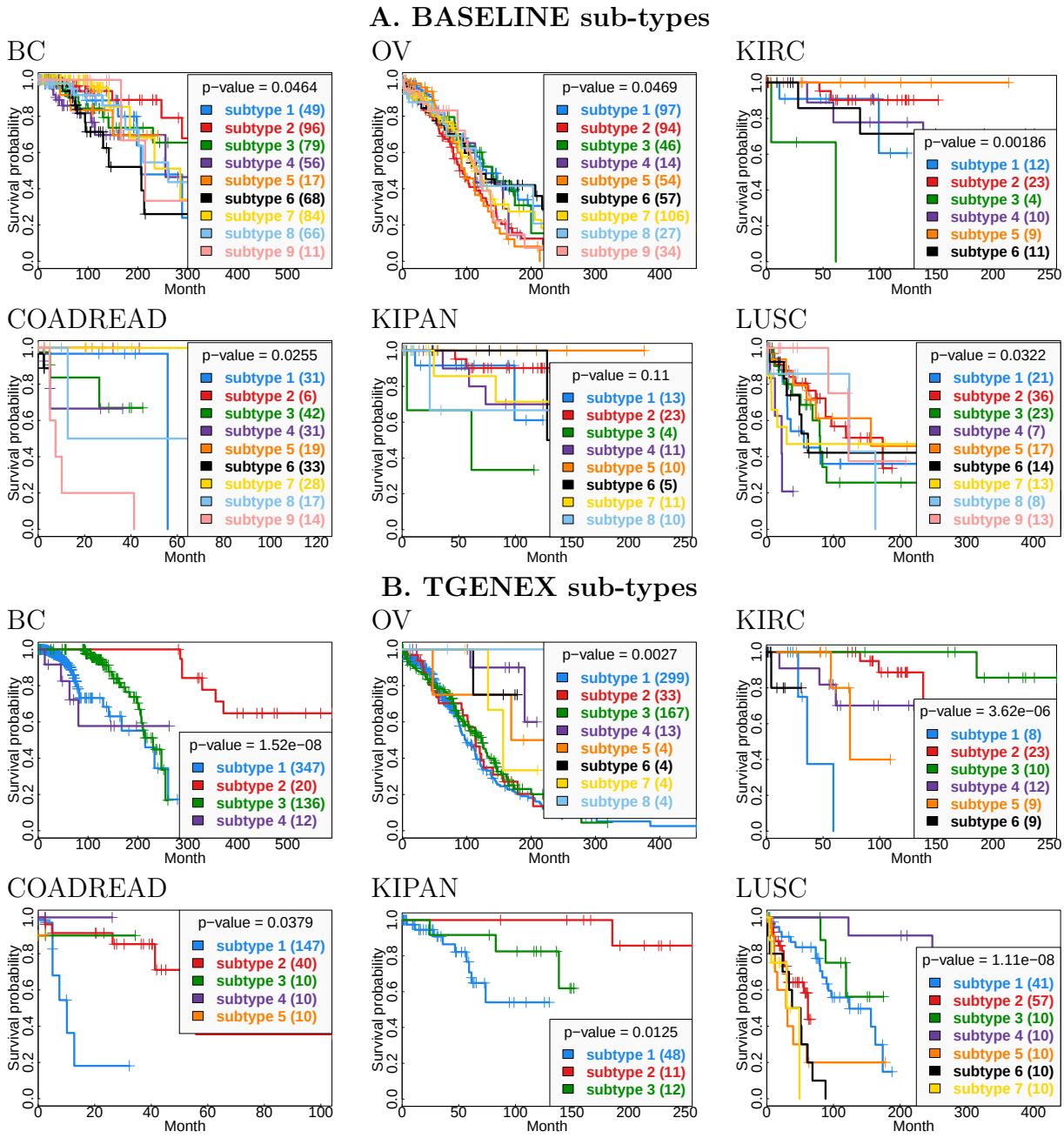


Table 6.11: Kaplan-Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=9. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=9.

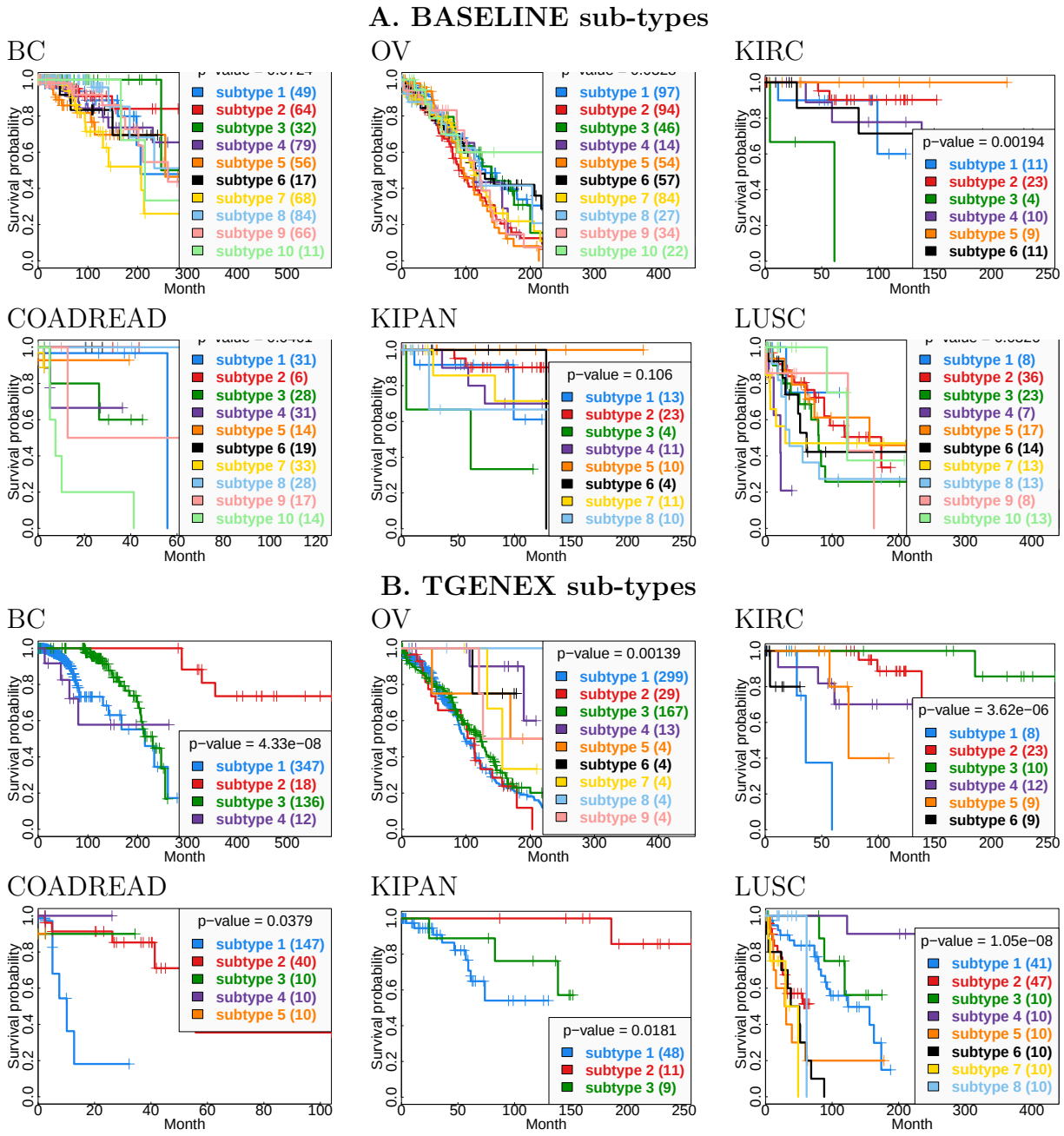


Table 6.12: Kaplan-Meier curves and Cox Log rank test p-value of the sub-types per each disease with k=10. (A) Baseline sub-types obtained with hierarchical clustering on mRNA data. (B) TGENEX sub-types were generated with the proposed method and k=10.

of the sub-types obtained using hierarchical clustering on the raw data and compares them with the sub-types obtained using *TGENEX* for the different numbers of clusters k. The two best Cox p-values obtained with mRNA expression for each cancer where obtained with the following values of k. For BC the best p-value was obtained with $k = 4$ ($p-value = 0.0029$) followed by $k = 5$ ($p - value = 0.0064$), for OV with $k = 3$ ($p - value = 0.0216$) followed by $k = 8$ ($p - value = 0.0304$), for KIRC with $k = 9$ ($p - value = 0.00186$) followed by $k = 4$ ($p - value = 0.0191$), for COADREAD with $k = 9$ ($p - value = 0.0255$) followed by $k = 8$ ($p - value = 0.0654$), for KIPAN with $k = 6$ ($p - value = 0.0459$) followed by $k = 8$ ($p - value = 0.0697$), for LUSC with $k = 9$ ($p - value = 0.00186$) followed by $k = 4$ ($p - value = 0.311$). In all these cases *TGENEX* performed better than the baseline.

The best two Cox p-values obtained with *TGENEX* for each cancer where obtained with the following values of k. For BC the best p-value was obtained with $k = 3$ ($p - value = 1.36e^{-10}$) followed by $k = 4$ ($p-value = 1.48e^{-10}$) , for OV with $k = 7$ ($p-value = 2.07e^{-9}$) followed by $k = 5$ ($p-value = 0.00112$), for KIRC with $k = 6$ ($p-value = 1.15e^{-6}$) followed by $k = 7$ and $k = 8$ ($p-value = 3.62e^{-6}$), for COADREAD with $k = 3$ ($p-value = 0.0239$) followed by $k = 7$ ($p - value = 0.0379$), for KIPAN with $k = 3$ ($p - value = 0.000834$) followed by $k = 4$ ($p-value = 0.00118$) , for LUSC with $k = 4$ ($p-value = 3.8e^{-9}$) followed by $k = 10$ ($p-value = 1.05e^{-8}$). In all these cases *TGENEX* performed better than mRNA expression alone. Our method shows that enriching gene expression data with mutation and clinical data improves the Cox p-values for these six different cancer datasets.

6.5 Biological analysis of TGENEX

In this section, we explore the biological contributions that our method can potentially elucidate for squamous cell carcinoma (LUSC). We decided to study this cancer because the molecular drivers of LUSC remain unclear and LUSC patients have limited therapeutic options [220], underscoring the potential to gain new understanding for this cancer sub-type. We focus on the sub-types obtained with TGENEX for LUSC with $k = 3$, which indicated that these survival groups are significantly different (p-value = 1.39e−7). Figure 6.2 display

the three survival curves, which allow us to validate that the three curves are distinctive visually. We can notice that sub-type 1 (curve in blue) corresponds to a short term survival group with 89 patients that decline in survival much faster than the other two groups; and sub-type 2 (curve in red) with 51 patients, has a slower pace of deaths than sub-type 1. Also, we observe that sub-type 3 (curve in green) has a better survival pattern than the other two groups, but it has only 12 patients, which makes it too small to allow us to extrapolate any conclusions to a greather population of LUSC patients; therefore, we did not include sub-type 3 in this study.

We performed the following four steps to explore the biological differences between the two major sub-types of LUSC (short-term versus long-term survival) obtained the experiments performed with TGENEX (sub-types 1 and 2, respectively). First, we identify a list of differentially expressed genes (DEG) and differentially mutated genes (DMG). Second, we compare the list of DE genes and DM genes with The COSMIC Cancer Gene Census [312]. Third, we performed pathway analysis with both lists (DEG and DMG) and discussed the list of pathways obtained. Forth, we compare the lists of significant pathways with the knowledge in The COSMIC Cancer Gene Census.

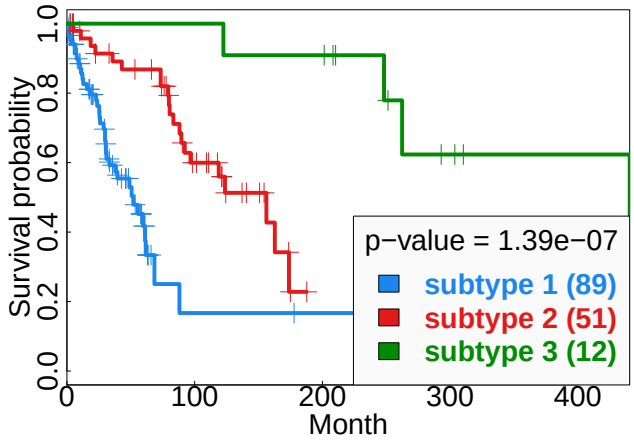


Figure 6.2: Kaplan-Meier plot of the sub-types of squamous cell carcinoma (LUSC) obtained using TGENEX with $k = 3$. Patients in subtype 1 are characterized by the worst survival.Patients belonging to subtype 2 has a slower decay in survival (long-term survival). Subtype 3 has the best survival dynamics, but consists of only 12 patients.

6.5 Identifying Differentially Expressed Genes and Differentially Mutated Genes

Here we present the process that we follow to find the list of genes that charaterize each sub-type in terms of gene expression and gene mutation. For gene mutation, we first find the relative frequency of gene mutation per each sub-type. Then, we find the intersection of the genes that are frequent in each group and remove the intersected genes from each sub-type. At the end, the list of diferentially mutated genes (DMG) for each sub-type is composed by the genes that are frequently mutated in such sub-type and are not frequently mutated in the other sub-type. Table 6.13 shows the top 10 differentially mutated genes for sub-types 1 and 2.

Differentially Mutated Genes

Here we detail the process to identify the differentially mutated genes (DMG) for each group. For this process, we use the original somatic mutation dataset from TCGA represented as a gene-level binary matrix for the patients that are part of our survival sub-types (see the details on this matrix construction in section 6.3.1). To identify the DMG between sub-type 1 and 2, we statistically tested if the proportion of gene mutations between the two groups are significantly different or not, obtained a p-value, and ranked the genes by their p-value.

The first step is to find per each gene the proportion of patients having a mutation in such a gene. We computed the sample proportion of mutated genes in sub-types 1 and 2 and ranked the genes decreasingly by proportion. Figures 6.4, and 6.5 show the top 50 mutated genes for sub-type 1 and 2, respectively. To visualize the co-occurrence of multiple alterations and to compare mutation patterns, we plot oncoprints of sub-type 1 in Figure 6.6 and sub-type 2 in Figure 6.7. To contrast, we plot the top 50 genes with the highest proportions for all the LUSC patients in Figure 6.3 and the oncoprints of all LUSC patients in Figure 6.8. As we can visually observe, the patterns for sub-type 1 (short-term survival) and the overall population look more similar than the pattern for sub-type 2.

After obtaining all proportions of mutated genes for each sub-type[4], we find the list of genes that have significantly different proportions between the two sub-types. For each gene g we performed a 2-sample Chi-squared test of proportions where the null hypothesis (H_0) is that the proportion of patients in sub-types 1 and 2 with a somatic mutation in gene g are the same [16]. We use a 10% threshold to define the area of rejection, i.e. if the p-value of Chi-squared test is < 0.10 we reject the null hypothesis and conclude that the proportion of mutation in the two sub-types is significantly different for gene g; therefore, we consider g a differentially mutated gene. In total we got 508 significantly differentially mutated genes.

gene	p-value	group
ARID4A	0.005344333	2
PCDHA13	0.007358202	2
MGC26647	0.007358202	2
EP400	0.007358202	2
ZCCHC12	0.007358202	2
AKAP4	0.007358202	2
ARHGEF9	0.007358202	2
TXLNB	0.007358202	2
C14orf145	0.007358202	2
DLGAP2	0.01459629	2
USP6	0.01459629	2

gene	p-value	group
NF1	0.009693621	1
CDH12	0.014423365	1
RYR2	0.021268981	1
RYR3	0.022152951	1
NLGN1	0.025714996	1
ARID1A	0.025714996	1
FREM1	0.025714996	1
TFAP2D	0.025714996	1
OR5D18	0.025714996	1
BOD1L	0.025714996	1
LRRC4C	0.025714996	1

Table 6.13: Top 10 differentially mutated genes of sub-type 2 on the left (long-term survival) and sub-type 1 on the right (short-term survival). The column 'gene' corresponds to each Gene Symbol, column 'p-value' contains the test of proportions p-value, and column 'group' indicates the sub-type that has the higher mutation load of such gene.

Differentially Expressed Genes

For gene expression, we downloaded normalized results of gene expression quantification (IlluminaHiSeq RNASeqV2) with RSEM of the TCGA LUSC dataset from the NCI Genomic Data Commons (GDC) and filtered genes with a quantile cut of 0.25. Differential Expression Analysis (DEA) of sub-type 1 (short-term survival) over sub-type 2 (long-term survival) was then performed with TCGAbiolinks package v3.10 [63] using the traditional generalized

[4]For each gene g, the proportion of mutation in a sub-type s is the count of patients in s with a mutation in g divided by the total number of patients in s.

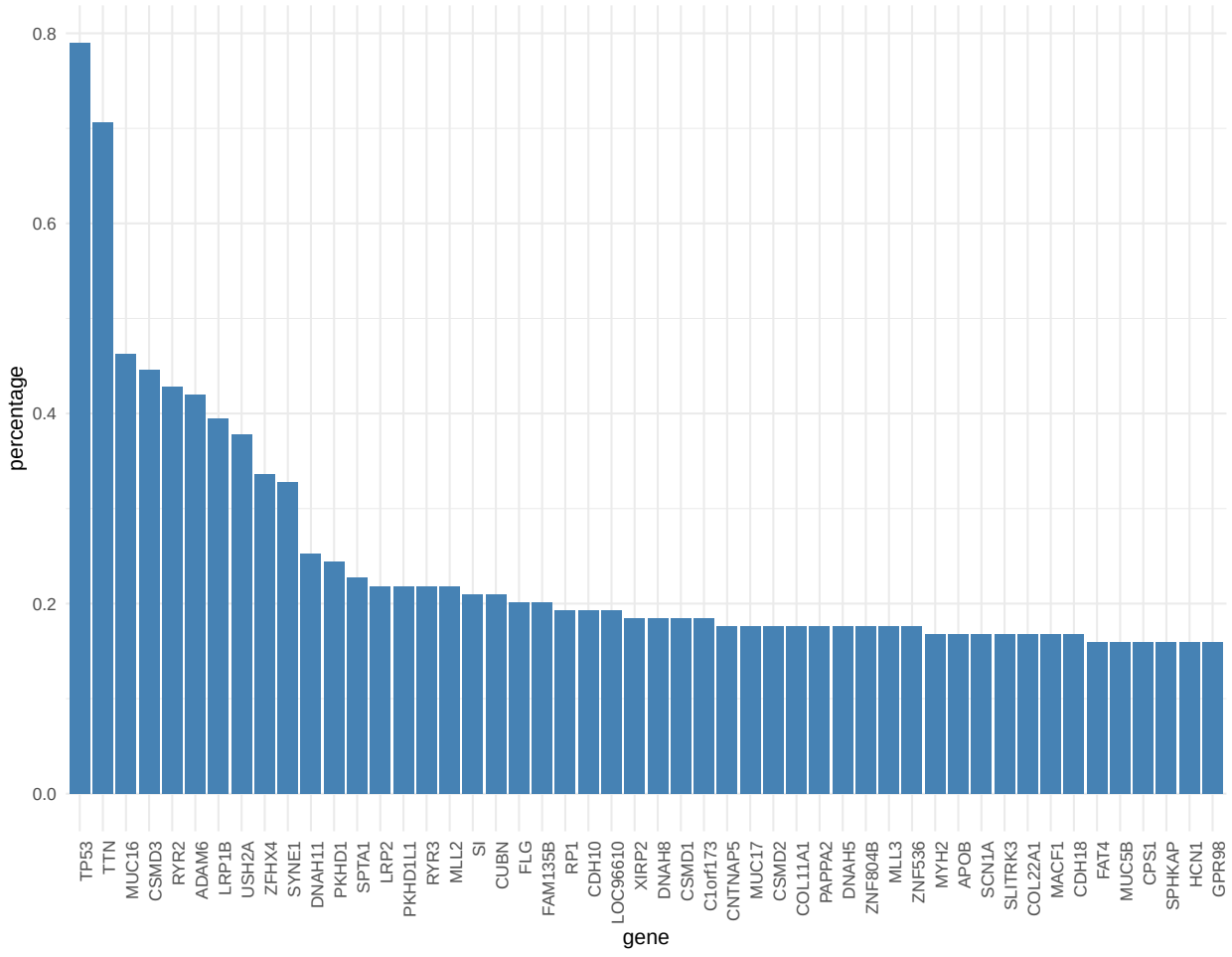


Figure 6.3: Distribution of mutations in all LUSC samples sorted descendingly by sample proportion, top 50.

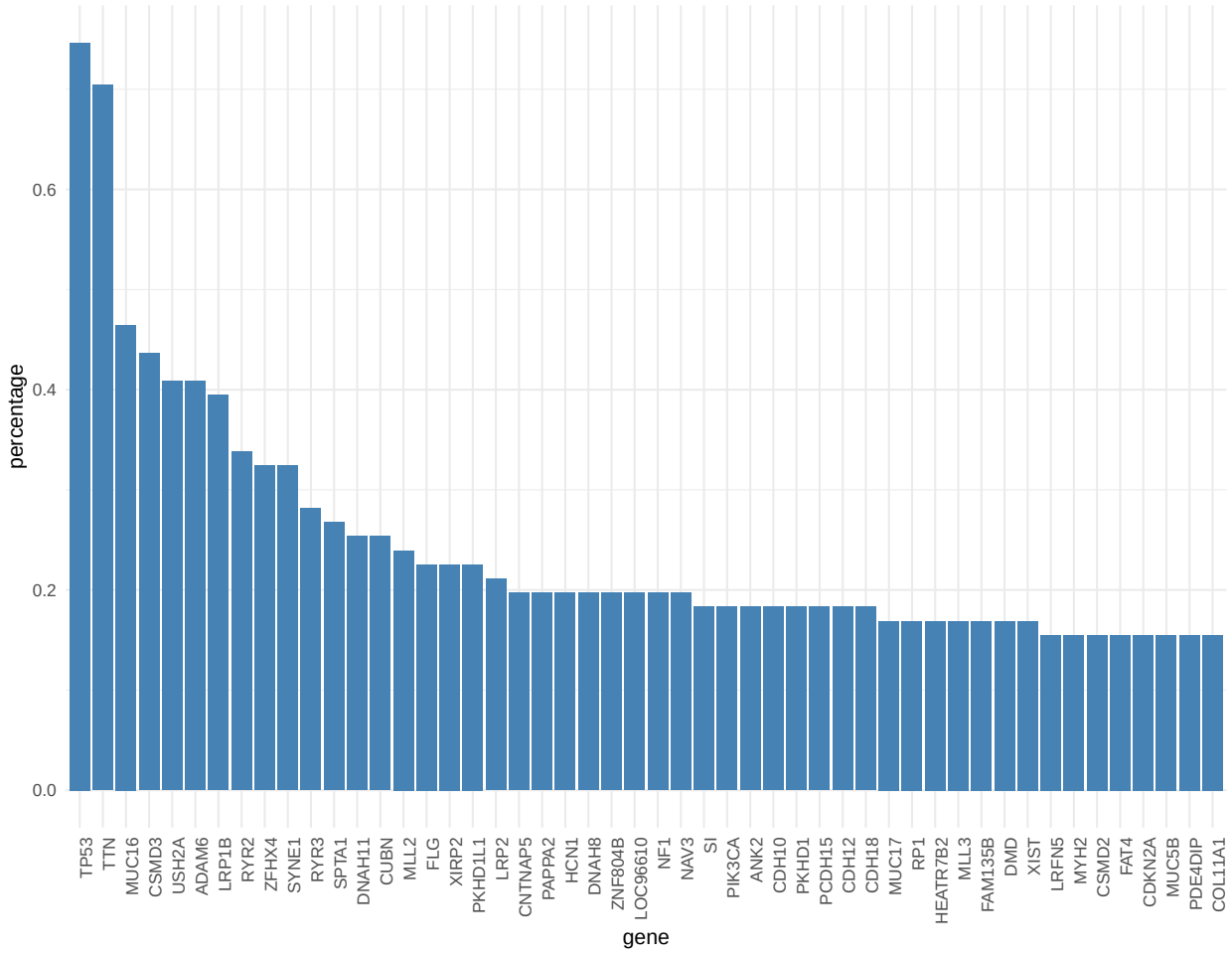


Figure 6.4: Distribution of mutations in sub-type 1 (short-term survival) sorted descendingly by frequency, top 50.

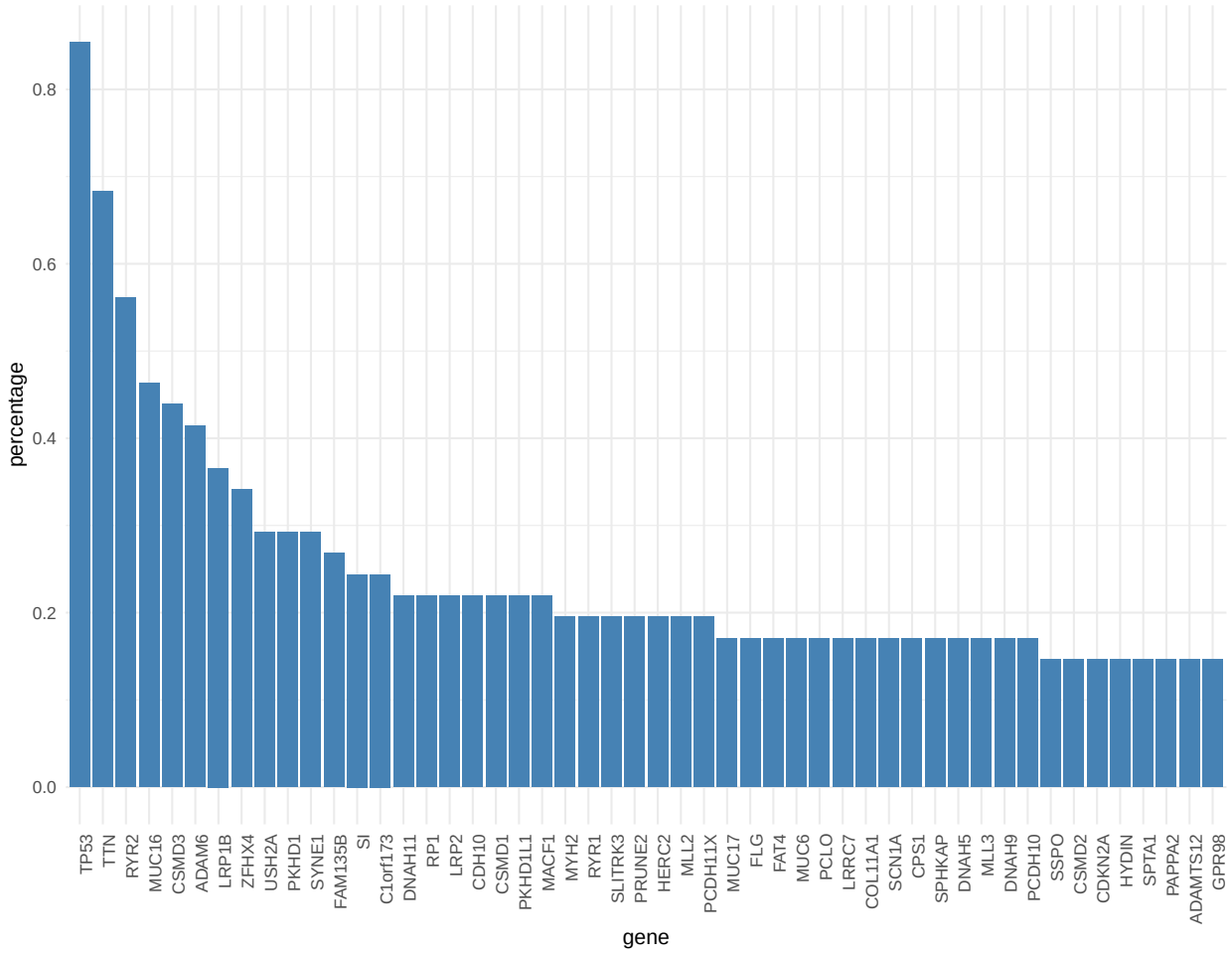


Figure 6.5: Distribution of mutations in sub-type 2 (long-term survival) sorted descendingly by frequency, top 50.

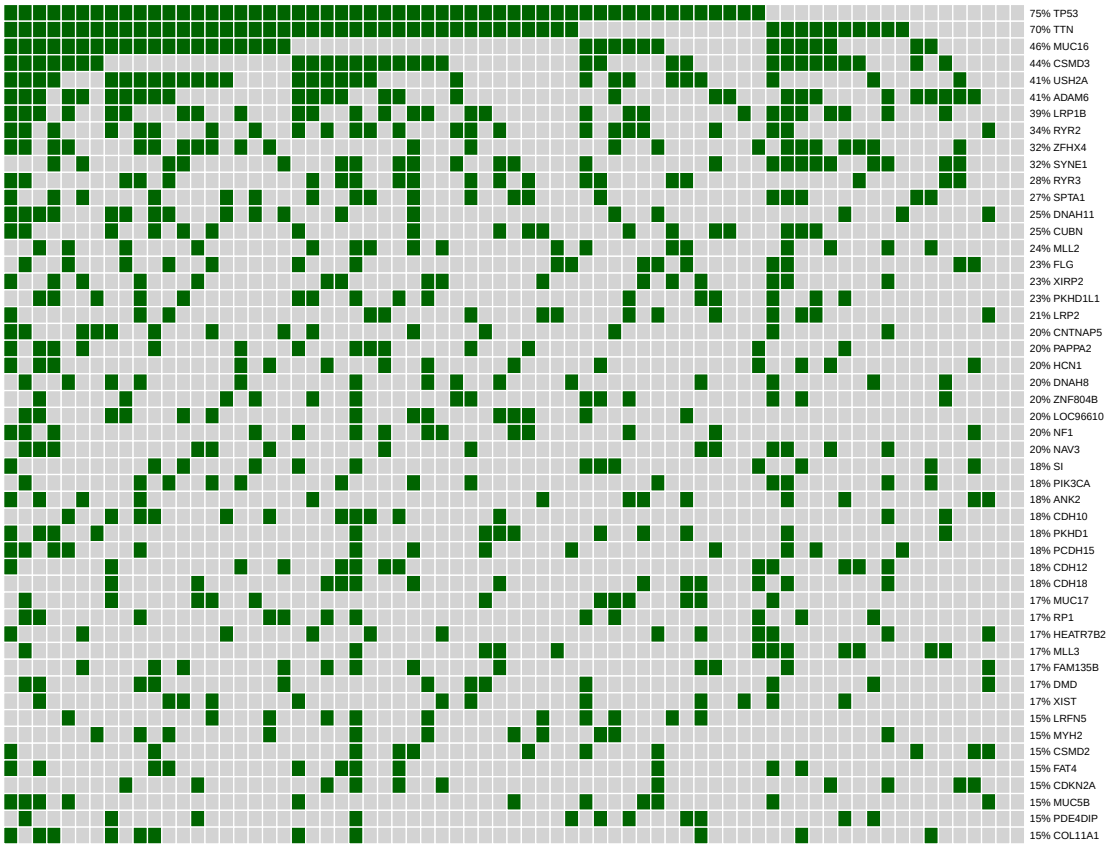


Figure 6.6: Oncoprint of mutations in sub-type 1 (short-term survival) samples.

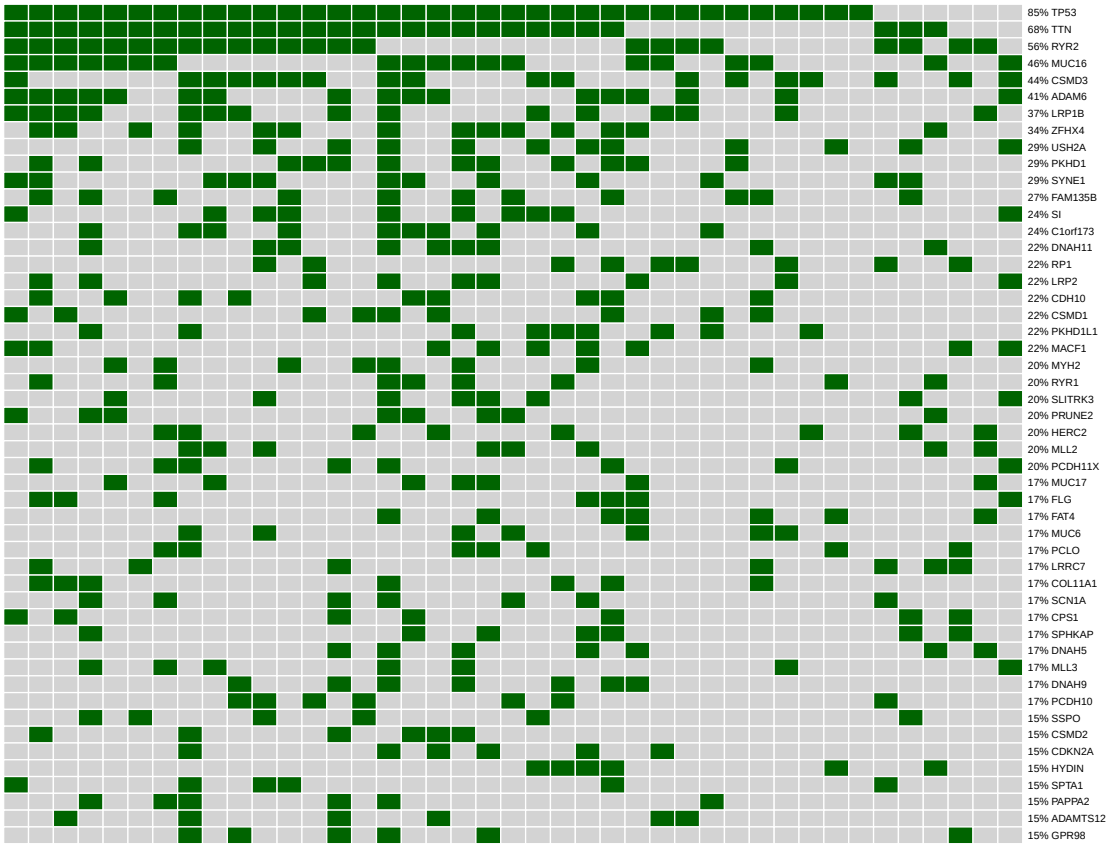


Figure 6.7: Oncoprint of mutations in sub-type 2 (long-term survival) samples.

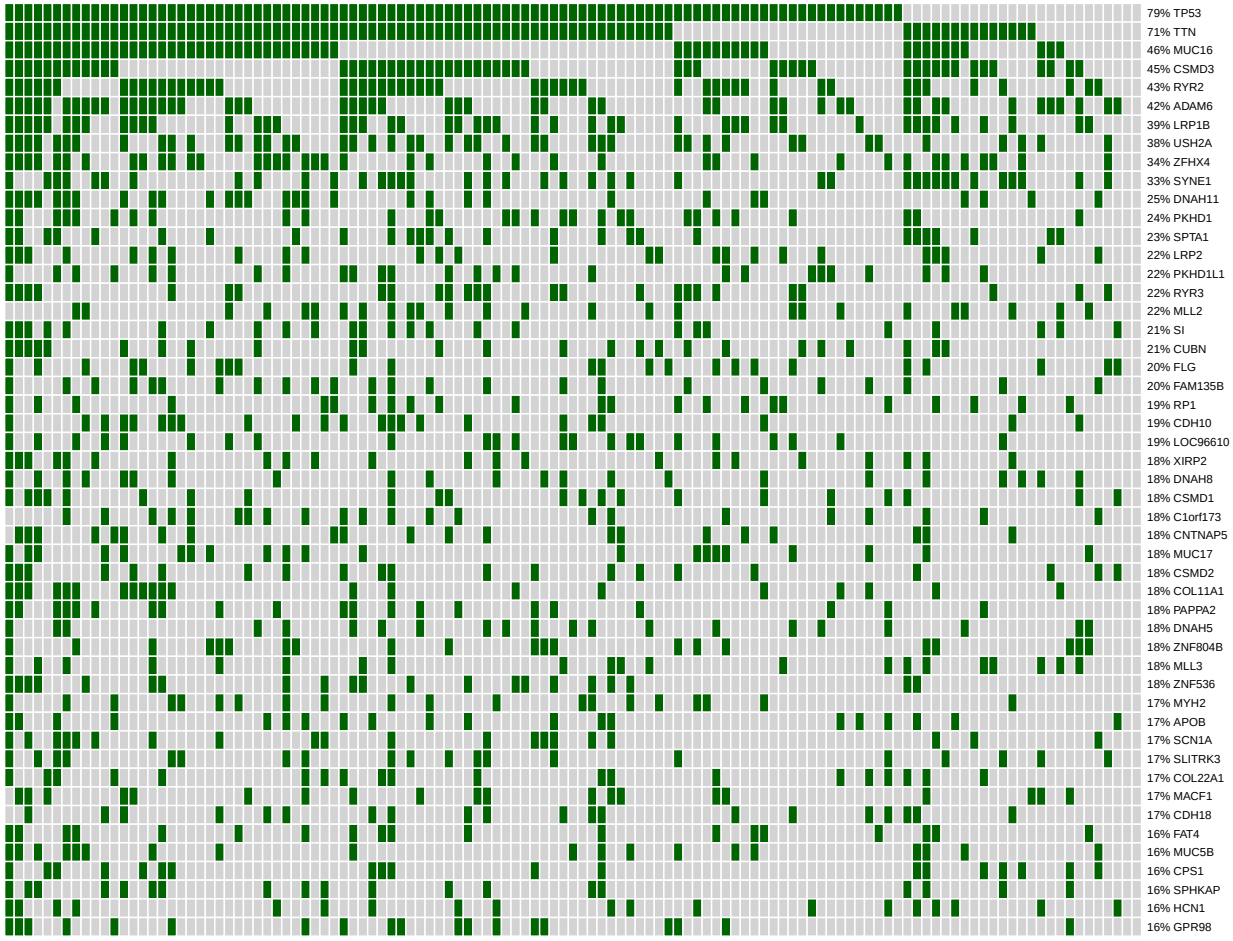


Figure 6.8: Oncoprint of mutations in all LUSC samples.

linear model (GLM) likelihood ratio test from edgeR [233, 63]. Genes were considered to be differentially expressed with adjusted p-value < 0.05 and fold-change $>= 2$. Table 6.14 shows the top 10 differentially expressed genes (DEG).

gene	logFC	p-value	FDR	FC
PSG8	-4.825603218	1.79E-10	7.05E-08	0.035265389
MOV10L1	-3.542744636	3.87E-08	1.04E-05	0.085807964
CYP11B2	-3.311367553	3.68E-05	0.004047234	0.100734687
MAGEB16	-2.96994675	3.09E-20	6.90E-17	0.127631226
TP53TG5	-2.126415445	1.30E-05	0.001678954	0.2290262
RNF14	-2.095209471	3.87E-06	0.000552361	0.234034079
GPR141	-1.379412809	1.54E-06	0.000259177	0.384375208
NVL	-1.358472558	9.58E-07	0.000183643	0.389994975
FKBP4	-1.285591138	4.41E-05	0.004775453	0.410202691
PTH2R	-1.034534905	3.03E-06	0.000483605	0.488173232

Table 6.14: Top 10 Differentially expressed genes of sub-type 1 (short-term survival) over sub-type 2 (long-term survival). The column 'gene' corresponds to each Gene Symbol, column 'logFC' contains the $\log_2 FC$, column 'p-value' contains the likelihood ratio test p-value, column 'FDR' shows the false discovery rate adjusted p-value, and column 'FC' contains the fold change measured as the ratio of sub-type 1 over sub-type 2.

Pathway name	Pathway Id	p-value
Cell adhesion molecules (CAMs)*	04514	0.011
Ribosome*	03010	0.038
Platelet activation	04611	0.046
Leukocyte transendothelial migration	04670	0.055
D-Glutamine and D-glutamate metabolism*	00471	0.057

Table 6.15: Top pathways and their associated p-values using DEG.
∗the p-value corresponding to the pathway was computed using only over-representation analysis.

6.5 Pathway analysis

Once we obtained the lists of DEG and DMG, we performed pathway analysis using gene expression and mutation. We performed pathway analysis using iPathwayGuide [2] (Advaita Corporation, 2020[5]) with KEGG pathways to identify the pathways that are modulated

[5] `https://apps.advaitabio.com/ipg/home`

in the two sub-types. We set the thresholds for log2-fold change of sub-type 1 (short-term survival) over sub-type 2 (long-term survival) of 1.0 and FDR $< 5\%$. iPathwayGuide implements Impact Analysis (IA) which combines two types of evidence: first, the classical over-representation of DEGs in a particular pathway and second, the perturbation of each pathway which is computed by propagating each gene expression change across the topology of a particular pathway. We used Impact Analysis (IA) because studies have shown that IA outperforms the classical over-representation analysis [99, 323]. The pathway graphs contain genes as nodes and their signaling interactions as edges and are obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [183, 6, 181, 185].

GO Term	**p-value**
chromatin organization	0.002
epithelial cell differentiation	0.002
mitochondrial membrane organization	0.003
negative regulation of fat cell differentiation	0.003
epithelial cell development	0.005

Table 6.16: Top identified biological processes obtained with IPathwayGuide analysis of DEG. Only the top scoring biological process.

Pathway name	**Pathway Id**	**p-value**
Circadian entrainment	04713	8.654e-5
ECM-receptor interaction	04512	5.231e-4
Oxytocin signaling pathway	04921	6.981e-4
Amoebiasis	05146	0.002
Long-term depression	04730	0.002

Table 6.17: Top pathways and their associated p-values using DMG

GO Term	**p-value**
ryanodine-sensitive calcium-release channel activity	0.002
extracellular matrix structural constituent	0.002
structural molecule activity	0.004
intracellular ligand-gated ion channel activity	0.010
calcium-induced calcium release activity	0.026

Table 6.18: Top identified biological processes obtained with IPathwayGuide analysis of DMG. Only the top scoring biological process.

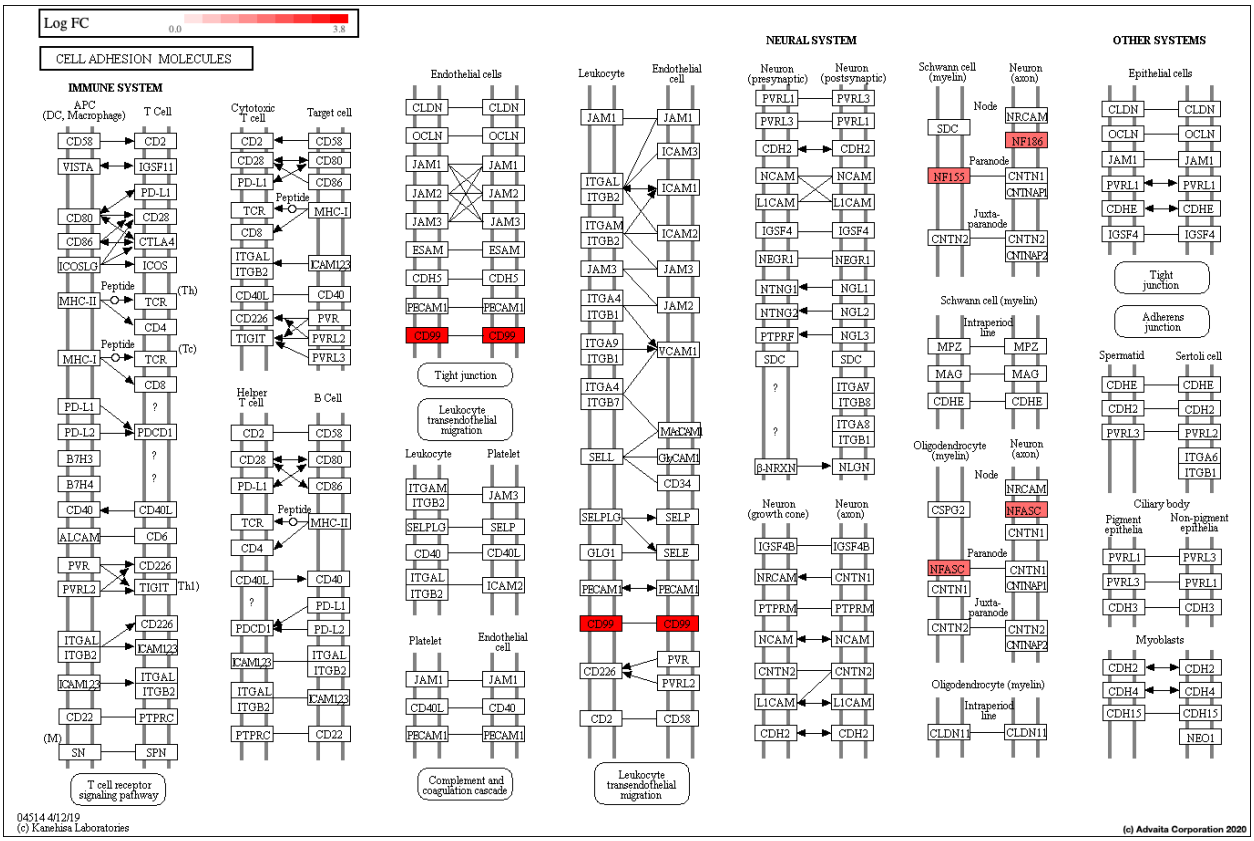


Figure 6.9: Most significant pathway when using DEG of sub-type 1 (short-term survival) over sub-type 2 (long-term survival), Cell adhesion molecules (CAMs) (KEGG: 04514): The pathway diagram is overlayed with the computed perturbation of each gene. The perturbation accounts both for the gene's measured fold change and for the accumulated perturbation propagated from any upstream genes (accumulation). The highest positive perturbation on sub-type 1 (short-term survival) with respect to sub-type 2 (long-term survival) is shown in dark red. The legend describes the values on the gradient. Note: For legibility, one gene may be represented in multiple places in the diagram and one box may represent multiple genes in the same gene family. A gene is highlighted in all locations it occurs in the diagram.

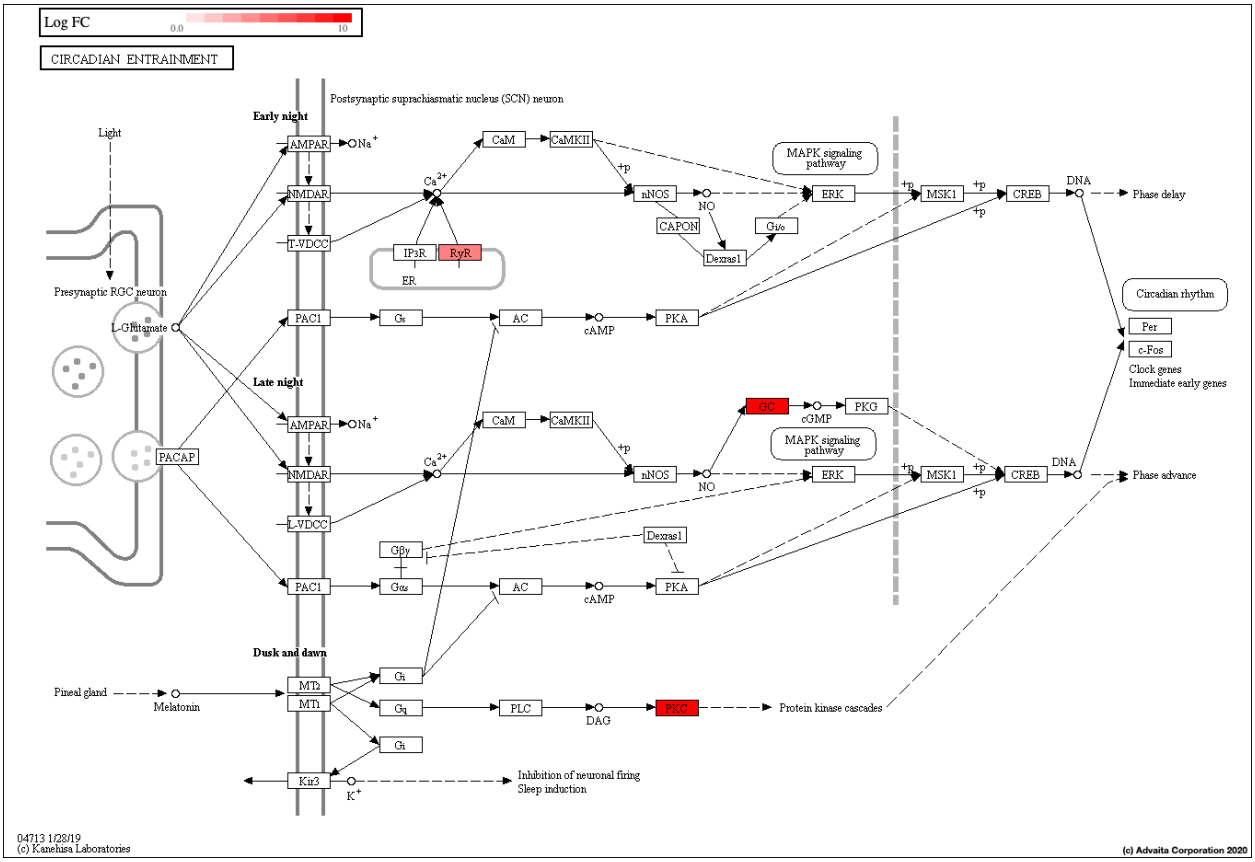


Figure 6.10: Most significant pathway using DMG of sub-type 1 (short-term survival) over sub-type 2 (long-term survival), Circadian entrainment (KEGG: 04713): The pathway diagram is overlayed with the computed perturbation of each gene. The perturbation accounts both for the gene's measured fold change and for the accumulated perturbation propagated from any upstream genes (accumulation). The highest positive perturbation on sub-type 1 (short-term survival) with respect to sub-type 2 (long-term survival) is shown in dark red.

Table 6.15 shows the top pathways and their associated p-values using differentially expressed genes (DEG). Figure 6.9 show the Cell adhesion molecules (CAMs) pathway which is the most significant pathway when using DEG. We also performed Gene Ontology (GO) analysis [8, 135] using iPathwayGuide and display the top GO terms in Table 6.16.

To run IA with mutation profiles, we used the list of differentially mutated genes with its p-value and proportion ratio of proportion of sub-type 1 over proportion of sub-type 2. Table 6.17 shows the top pathways and their associated p-values using differentially mutated genes (DMG) and Table 6.18 shows the top GO terms. Figure 6.9 show the Cell adhesion molecules (CAMs) pathway which is the most significant pathway when using DMG.

6.5 Comparison with The Cancer Genome Census

To identify if the genes that we found are novel we use the COSMIC Cancer Gene Census [312] as the baseline. "The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) is an expert-curated description of the genes driving human cancer that is used as a standard in cancer genetics across basic research, medical reporting and pharmaceutical development" [312]. For this, we downloaded the data release v90 (5th September 2019) of the catalog and found the list of genes shown in Table 6.19. Interestingly, gene NOTCH1 has been associated with squamous-cell carcinoma of the lung [175, 308, 258, 19] which evidences the potential of the rest of the genes that we found to be promising candidates to differentiate between short-term and long-term survival patients.

6.5 Identifying Differential Clinical Variables

In addition to identifying DMG and DEG, we analyzed the clinical variables and identified the clinical variables that are significantly different between sub-type 1 and 2 (short-term and long-term survival, respectively), which we detailed as follows. First, we compute per each clinical variable the proportion of patients having a positive value of such a clinical variable for each sub-type[6]. Second, we find the list of clinical variables that have significantly

[6]For each clinical variable c, the proportion of presence in a sub-type s is the count of patients in s with a positive value in c divided by the total number of patients in s.

Gene	**Tumor Types(Somatic)**	**Role in Cancer**
NOTCH1	T-ALL, breast, bladder, skin SCC, lung SCC, head and neck SCC	oncogene, TSG, fusion
NF1	neurofibroma, glioma	TSG, fusion
USP6	aneurysmal bone cyst	oncogene, fusion
CTNNB1	colorectal, ovarian, hepatoblastoma, pleomorphic salivary gland adenoma, other tumour types	oncogene, fusion
BCL9L	colorectal cancer, endometrial carcinoma, gastric cancer	oncogene, TSG
SRGAP3	pilocytic astrocytoma	fusion
ARID1A	clear cell ovarian carcinoma, RCC, breast	TSG, fusion
BCL11A	B-CLL	oncogene, fusion
PCM1	papillary thyroid, CML, MPN	fusion
TBL1XR1	splenic marginal zone lymphoma, primary central nervous system lymphoma, colorectal carcinoma, gallbladder carcinoma	oncogene, TSG, fusion
NFKB2	B-NHL	oncogene, TSG, fusion
CLTCL1	ALCL	TSG, fusion
MSN	ALCL	fusion
PRKCB	adult T-cell lymphoma-leukaemia	0
CARS	ALCL	TSG, fusion
TET1	AML	oncogene, TSG, fusion
ATRX	pancreatic neuroendocrine tumours, paediatric GBM	TSG
CNTNAP2	glioma, melanoma	TSG

Table 6.19: List of DEG and DMG that were found in the COSMIC Cancer Gene Census.

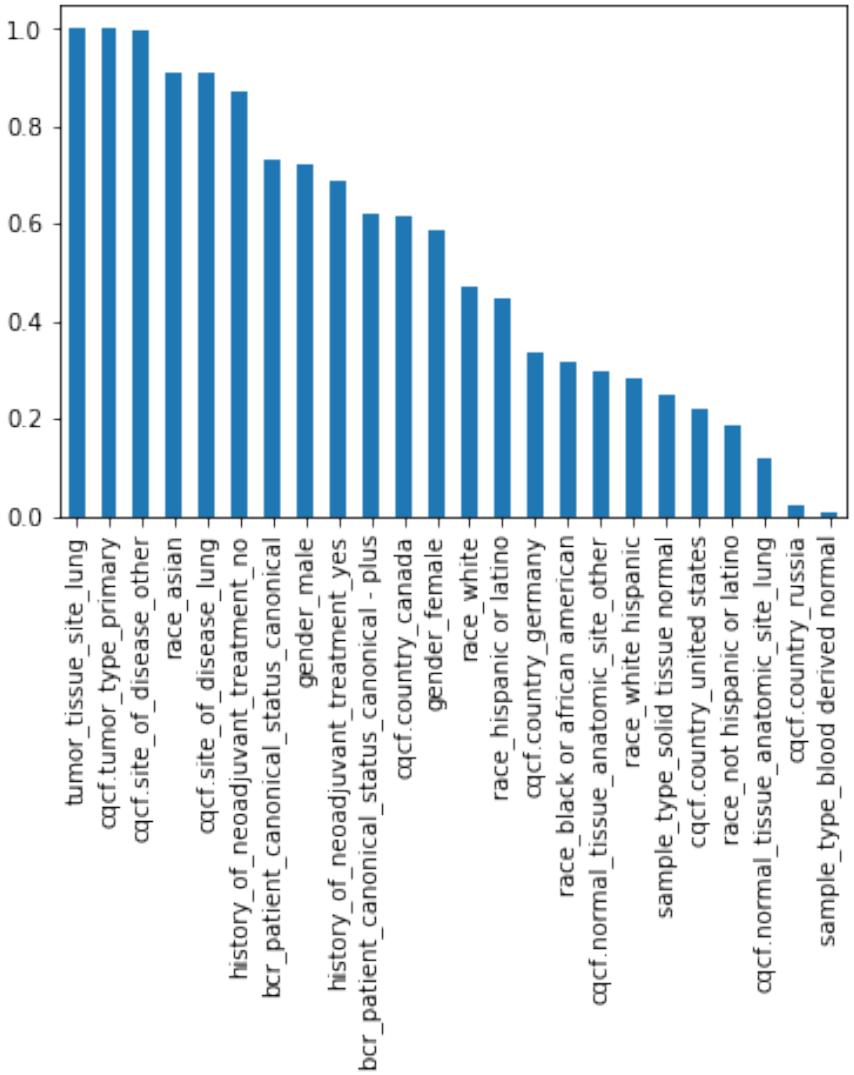


Figure 6.11: Differential Clinical Variables. Plot of the Chi-squared p-value of the top clinical variables.

p-value	Clinical Variable
0.007710	sample_type_blood derived normal
0.023149	country_russia
0.116514	normal_tissue_anatomic_site_lung
0.186494	race_not hispanic or latino
0.222184	country_united states
0.249933	sample_type_solid tissue normal
0.284374	race_white hispanic
0.295797	normal_tissue_anatomic_site_other
0.314466	race_black or african american
0.337901	country_germany
0.449056	race_hispanic or latino
0.472795	race_white
0.585688	gender_female
0.617191	country_canada
0.618592	bcr_patient_canonical_status_canonical - plus
0.690018	history_of_neoadjuvant_treatment_yes
0.721205	gender_male
0.731868	bcr_patient_canonical_status_canonical
0.872053	history_of_neoadjuvant_treatment_no
0.908088	site_of_disease_lung
0.911296	race_asian
0.996429	site_of_disease_other
1.000000	tumor_type_primary
1.000000	tumor_tissue_site_lung

Table 6.20: P-values Clinical Variables. Chi-squared p-value of top 25 clinical variables.

different proportions between the two sub-types. For each clinical variable c we performed a standard 2-sample Chi-squared test of proportions where the null hypothesis (H_0) is that the proportion of patients in sub-types 1 and 2 with the presence of clinical variable c are the same [16]. We use a 5% threshold to define the area of rejection, i.e., if the p-value of Chi-squared test is < 0.05 we reject the null hypothesis and conclude that the proportion of the clinical variable c in the two sub-types is significantly different; therefore, we consider c a differential clinical variable. Figure 6.11 and Table 6.20 show the p-values of the top clinical variables. We can see that the sample type and country of origin are clinical variables that significantly differentiate the two sub-types. Sub-type 2 (long-term survival) has only two samples blood-derived out of 51 total samples (49 solid-tissue derived samples), and none of the patients are from Russia, while 20 out of 89 samples in sub-type 1 (short-term survival) are blood samples and all Russian patients (9) belong to sub-type 1.

Also, we evaluated the features associated with sub-type 1 (short-term survival) to realize their potential relevance and identify possible new insights for LUSC. We began by looking at the clinical attributes associated with sub-type 1, the highest ranked and significant clinical variable was sample_type_blood derived normal. This was somewhat unexpected, however, NOTCH1 and NF1 are two examples of well-established cancer genes overrepresented in sub-type 1 for which mutations are somatic and constitutional, and therefore they do arise in patient blood samples. The data that we analyzed included sample sets from a number of countries of origin. Identifying Russia significantly associated with sub-type 1 lead us to realize that Russia ranks among countries with the highest lung cancer mortality rates [330], perhaps the molecular elements associated with LUSC sub-type 1 are contributing to that.

Among the mutated genes associated with sub-type 1, NF1 is the most statistically significant. This tumor suppressor gene is inactivated by mutation. NF1 inactivation promotes mutant KRAS-driven lung adenocarcinoma [349]. Our results may be an indication of the importance of KRAS signaling absent KRAS mutation in LUSC. Further down on the full list of mutated genes associated with sub-type 1 is TET1. According to a recent Pubmed search,

there are no publications describing TET1 mutations in lung cancer. TET1 mutations are known to be activating and oncogenic in other cancer types [194, 169, 225]. Moreover, for other cancer types TET1 has a role in driving a subset of aggressive cancer cells in tumors known as cancer stem cell like-cells (CSCs) [358, 162, 368, 304, 316, 116]. CSCs are the recognized source of primary malignant tumor initiation and they give rise to therapy resistance and metastases [279, 53, 265]. Further research is warranted to learn if TET1 can also drive CSCs in LUSC, if it does this may help to explain how TET1 is associated with the short-term survival LUSC sub-type 1.

6.6 Conclusion

In this chapter, we introduced *TGENEX*, a framework to combine gene expression data, mutation data, and clinical data for unsupervised cancer subtyping based on non-negative tensor decomposition. The performance of the new approach was demonstrated on seven different cancer datasets downloaded from TCGA. *TGENEX* was applied in conjunction with k-means, spectral, and hierarchical clustering. For these three clustering algorithms, our approach dramatically improves the subtyping. Our contribution is two-fold. First, this framework introduces a new way to combine clinical and genetic data. Although the framework was demonstrated on cancer datasets, it can be applied to analyze data from other complex diseases. Second, this framework is the first one that integrates clinical data, mutation, and gene expression data for disease subtyping and make these sub-types available for further biological exploration and experimentation. As a proof of concept, our method shows that enriching gene expression data with mutation and clinical data allows to obtain cancer sub-types with more distinct survival dynamics. For future work, we plan to study in-depth the discovered sub-types and integrate other data types, such as microRNA, for a more comprehensive analysis [88].

REFERENCES

[1] Acar, E., Dunlavy, D. M., and Kolda, T. G. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics 25*, 2 (Feb 2011), 67–86.

[2] Ahsan, S., and Drăghici, S. Identifying significantly impacted pathways and putative mechanisms with ipathwayguide. *Current protocols in bioinformatics 57*, 1 (2017), 7–15.

[3] Akula, S. P., Miriyala, R. N., Thota, H., Rao, A. A., and Gedela, S. Techniques for integrating omics data. *Bioinformation 3*, 6 (Jan. 2009), 284–286.

[4] Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics 21*, 13 (2005), 2988–2993.

[5] Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson Jr, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature 403*, 6769 (2000), 503–511.

[6] Arakawa, K., Kono, N., Yamada, Y., Mori, H., and Tomita, M. KEGG-based pathway visualization tool for complex omics data. *In Silico Biology 5*, 4 (Jan. 2005), 419–423.

[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.,

Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene Ontology: tool for the unification of biology. *Nature Genetics 25* (2000), 25–29.

[8] Ashburner, M., et al. Creating the Gene Ontology Resource: Design and Implementation. *Genome Research 11* (2001), 1425–1433.

[9] Ashley, E. A. The precision medicine initiative. *JAMA 313*, 21 (Jun 2015), 2119.

[10] Backes, C., Meese, E., Lenhof, H.-P., and Keller, A. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Research 38*, 13 (2010), 4476–4486.

[11] Bagirov, A. M., and Mardaneh, K. Modified global k-means algorithm for clustering in gene expression data sets. In *Proceedings of the 2006 Workshop on Intelligent Systems for Bioinformatics* (2006), vol. 73, Australian Computer Society, Inc., pp. 23–28.

[12] Bair, E., and Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. In *PLOS Biol* [13], p. e108.

[13] Bair, E., and Tibshirani, R. Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLOS Biol 2*, 4 (Apr. 2004), e108.

[14] Baldi, P., and Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics 17*, 6 (2001), 509–519.

[15] Bao, B., Mitrea, C., Wijesinghe, P., Marchetti, L., Girsch, E., Farr, R. L., Boerner, J. L., Mohammad, R., Dyson, G., Terlecky, S. R., and Bollig-Fischer, A. Treating triple negative breast cancer cells with erlotinib plus

a select antioxidant overcomes drug resistance by targeting cancer cell heterogeneity. *Scientific Reports 7* (Mar. 2017), srep44125.

[16] Baron, M. *Probability and statistics for computer scientists*. CRC Press, 2019.

[17] Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. The GOA database in 2009—an integrated Gene Ontology Annotation resource. *Nucleic acids research 37*, suppl 1 (2009), D396–D403.

[18] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research 41*, D1 (2013), D991–D995.

[19] Barros-Filho, M. C., Guisier, F., Rock, L. D., Becker-Santos, D. D., Sage, A. P., Marshall, E. A., and Lam, W. L. Tumour suppressor genes with oncogenic roles in lung cancer. In *Tumor Suppressor Genes*. IntechOpen, 2019.

[20] Barry, W. T., Nobel, A. B., and Wright, F. A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics 21*, 9 (May 2005), 1943–1949.

[21] Bastien, R. R., Rodríguez-Lescure, Á., Ebbert, M. T., Prat, A., Munárriz, B., Rowe, L., Miller, P., Ruiz-Borrego, M., Anderson, D., Lyons, B., et al. Pam50 breast cancer subtyping by rt-qpcr and concordance with standard clinical molecular markers. *BMC medical genomics 5*, 1 (2012), 44.

[22] Beissbarth, T., and Speed, T. P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics 20* (June 2004), 1464–1465.

[23] Benada, J., and Macurek, L. Targeting the checkpoint to kill cancer cells. *Biomolecules 5*, 3 (2015), 1912–1937.

[24] Berger, B., Peng, J., and Singh, M. Computational solutions for omics data. *Nature Reviews Genetics 14*, 5 (May 2013), 333+. 333.

[25] Berriz, G. F., King, O. D., Bryant, B., Sander, C., and Roth, F. P. Characterizing gene sets with FuncAssociate. *Bioinformatics 19*, 18 (2003), 2502–2504.

[26] Bewick, V., Cheek, L., and Ball, J. Statistics review 12: Survival analysis. *Critical Care 8*, 5 (2004), 389–394.

[27] Biagi, J. J., Raphael, M. J., Mackillop, W. J., Kong, W., King, W. D., and Booth, C. M. Association between time to initiation of adjuvant chemotherapy and survival in colorectal cancer. *JAMA 305*, 22 (2011), 2335.

[28] BioCarta. BioCarta - Charting Pathways of Life. http://www.biocarta.com, 2004.

[29] Blute, M. L., Bostwick, D. G., Bergstralh, E. J., Slezak, J. M., Martin, S. K., Amling, C. L., and Zincke, H. Anatomic site-specific positive margins in organconfined prostate cancer and its impact on outcome after radical prostatectomy. *Urology 50*, 5 (1997), 733–739.

[30] Boja, E. S., Kinsinger, C. R., Rodriguez, H., Srinivas, P., and $author.lastName, a. f. Integration of omics sciences to advance biology and medicine. *Clinical Proteomics 11*, 1 (Dec. 2014), 45.

[31] Boland, M. R., Hripcsak, G., Shen, Y., Chung, W. K., and Weng, C. Defining a comprehensive verotype using electronic health records for personalized medicine. *Journal of the American Medical Informatics Association: JAMIA 20*, e2 (Dec 2013), e232–8.

[32] Bollig-Fischer, A. How the future of clinical cancer diagnostics can contribute to overcoming race-associated cancer disparities. *Expert Review of Molecular Diagnostics 16*, 12 (2016), 1233–1235.

[33] Botstein, D., and Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics 33 Suppl* (Mar 2003), 228–237.

[34] Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.-A. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research 31*, 1 (2003), 68–71.

[35] Breastcancer.org. Types of breast cancer: Non-invasive, invasive and more, Oct 2018.

[36] Breslin, T., Eden, P., and Krogh, M. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics 5*, 1 (2004), 193.

[37] Bro, R., and Kiers, H. A. L. A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics 17*, 5 (Jun 2003), 274–286.

[38] Bro, R., Kjeldahl, K., Smilde, A. K., and Kiers, H. A. L. Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry 390*, 5 (Mar 2008), 1241–1251.

[39] Bushel, P. R., Wolfinger, R. D., and Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. In *BMC Systems Biology* [40], p. 15.

[40] Bushel, P. R., Wolfinger, R. D., and Gibson, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology 1* (2007), 15.

[41] Calura, E., Martini, P., Sales, G., Beltrame, L., Chiorino, G., D'Incalci, M., Marchini, S., and Romualdi, C. Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Research* (2014), gku354.

[42] Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research 32*, suppl 1 (2004), D262–D266.

[43] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature 474*, 7353 (2011), 609–615.

[44] Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature 489*, 7417 (2012), 519–525.

[45] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature 487*, 7407 (2012), 330–337.

[46] Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. *Nature 497*, 7447 (2013), 67–73.

[47] Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. *Cell 159*, 3 (2014), 676–690.

[48] Cancer Genome Atlas Research Network, T., Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael,

J. F., Fulton, L. L., Dooling, D. J., Ding, L., and et al. Comprehensive molecular portraits of human breast tumours. In *Nature* [49].

[49] Cancer Genome Atlas Research Network, T., Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., and et al. Comprehensive molecular portraits of human breast tumours. *Nature 1* (2012).

[50] Carey, V. J., and Stodden, V. Reproducible Research Concepts and Tools for Cancer Bioinformatics. In *Biomedical Informatics for Cancer Research*, M. F. Ochs, J. T. Casagrande, and R. V. Davuluri, Eds. Springer US, 2010, pp. 149–175.

[51] Carroll, J. D., and Chang, J.-J. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika 35*, 3 (Apr 1970), 283–319.

[52] Carroll, R. J., Eyler, A. E., and Denny, J. C. Naïve Electronic Health Record Phenotype Identification for Rheumatoid Arthritis. *AMIA Annual Symposium Proceedings 2011* (2011), 189.

[53] Chaffer, C. L., and Weinberg, R. A. How does multistep tumorigenesis really proceed? *Cancer discovery 5*, 1 (2015), 22–24.

[54] Chalise, P., Koestler, D. C., Bimali, M., Yu, Q., and Fridley, B. L. Integrative clustering methods for high-dimensional molecular data. *Translational cancer research 3*, 3 (June 2014), 202–216.

[55] Chen, C.-Y., Lee, P. H., Castro, V. M., Minnier, J., Charney, A. W., Stahl, E. A., Ruderfer, D. M., Murphy, S. N., Gainer, V., Cai, T., et al. Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. *Translational psychiatry 8*, 1 (2018), 86.

[56] Chen, Y., Carroll, R. J., Hinz, E. R. M., Shah, A., Eyler, A. E., Denny, J. C., and Xu, H. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association 20*, e2 (Dec 2013), e253–e259.

[57] Cheng, T.-Y. D., Cramb, S. M., Baade, P. D., Youlden, D. R., Nwogu, C., and Reid, M. E. The international epidemiology of lung cancer: latest trends, disparities, and tumor characteristics. *Journal of Thoracic Oncology 11*, 10 (2016), 1653–1671.

[58] Chi, E. C., and Kolda, T. G. On Tensors, Sparsity, and Nonnegative Factorizations. *SIAM Journal on Matrix Analysis and Applications; Philadelphia 33*, 4 (2012), 1272–1299.

[59] Chi, E. C., and Kolda, T. G. On tensors, sparsity, and nonnegative factorizations. *SIAM Journal on Matrix Analysis and Applications; Philadelphia 33*, 4 (2012), 1272–1299.

[60] Chin, L., Andersen, J. N., and Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine 17*, 3 (Mar 2011), 297–303.

[61] Cleveland, W., Grosse, E., and Shyu, W. Local regression models. in 'statistical models in s'.(eds jm chambers and tj hastie.) pp. 309–373, 1992.

[62] Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., Thürlimann, B., Senn, H.-J., Members, P., André, F., et al. Tailoring therapies improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. *Annals of Oncology 26*, 8 (2015), 1533–1546.

[63] Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., et al.

Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research 44*, 8 (2015), e71–e71.

[64] Collins, F. S., and Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine 372*, 9 (Feb 2015), 793–795.

[65] Collins, F. S., and Varmus, H. A new initiative on precision medicine. *New England Journal of Medicine 372*, 9 (2015), 793–795.

[66] Collisson, E. A., Sadanandam, A., Olson, P., Gibb, W. J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G. E., Jakkula, L., and et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature Medicine 17*, 4 (Apr 2011), 500–503.

[67] Consortium, G. O., et al. The gene ontology (go) database and informatics resource. *Nucleic acids research 32*, suppl 1 (2004), D258–D261.

[68] Cook, M. B., Dawsey, S. M., Freedman, N. D., Inskip, P. D., Wichner, S. M., Quraishi, S. M., Devesa, S. S., and McGlynn, K. A. Sex disparities in cancer incidence by period and age. *Cancer Epidemiology and Prevention Biomarkers 18*, 4 (2009), 1174–1182.

[69] Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*, 2 (1972), 187–220.

[70] Cox, D. R. *Analysis of survival data*. Routledge, 2018.

[71] Creighton, C. J., Nagaraja, A. K., Hanash, S. M., Matzuk, M. M., and Gunaratne, P. H. A bioinformatics tool for linking gene expression profiling results with public databases of microRNA target predictions. *RNA 14*, 11 (Nov. 2008), 2290–2296.

[72] Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., and D'Eustachio, P. The Reactome pathway knowledgebase. *Nucleic Acids Research 42*, D1 (2014), D472–D477.

[73] Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., and Shi, B. Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research 5*, 10 (2015), 2929.

[74] Das, J., Podder, S., and Ghosh, T. C. Insights into the miRNA regulations in human disease genes. *BMC Genomics 15* (2014), 1010.

[75] Dave, B., Mittal, V., Tan, N. M., and Chang, J. C. Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. *Breast Cancer Research 14* (Jan 2012), 202.

[76] Davis, C. F., Ricketts, C. J., Wang, M., Yang, L., Cherniack, A. D., Shen, H., Buhay, C., Kang, H., Kim, S. C., Fahey, C. C., et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer cell 26*, 3 (2014), 319–330.

[77] De Keersmaecker, S. C. J., Thijs, I. M. V., Vanderleyden, J., and Marchal, K. Integration of omics data: how well does it work for bacteria? *Molecular Microbiology 62*, 5 (Dec. 2006), 1239–1250.

[78] Del Priore, G., Zandieh, P., and Lee, M.-J. Treatment of continuous data as categoric variables in obstetrics and gynecology. *Obstetrics & Gynecology 89*, 3 (1997), 351–354.

[79] Dembele, D., and Kastner, P. Fuzzy C-means method for clustering microarray data. *Bioinformatics 19*, 8 (2003), 973–980.

[80] Denny, J. C., Bastarache, L., and Roden, D. M. Phenome-wide association studies as a tool to advance precision medicine. *Annual review of genomics and human genetics 17* (2016), 353–373.

[81] Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., and Crawford, D. C. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics 26*, 9 (May 2010), 1205–1210.

[82] DePinho, R. A. The age of cancer. *Nature 408*, 6809 (2000), 248–254.

[83] Diaz, D., Bavota, G., Marcus, A., Oliveto, R., Takahashi, S., and De Lucia, A. Using code ownership to improve IR-based traceability link recovery. In *2013 21st International Conference on Program Comprehension (ICPC)* (2013), IEEE, pp. 123–132.

[84] Diaz, D., Bollig-Fischer, A., and Kotov, A. Tensor decomposition for subtyping of complex diseases based on clinical and genomic data. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)* (2019), IEEE.

[85] Diaz, D., Bollig-Fischer, A., and Kotov, A. Tensor decomposition for subtyping of complex diseases based on clinical and genomic data. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)* [84].

[86] Diaz, D., Bollig-Fischer, A., and Kotov, A. Tensor decomposition for subtyping of complex diseases based on clinical and genomic data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2019)* (2019), IEEE.

[87] Diaz, D., Donato, M., Nguyen, T., and Draghici, S. MicroRNA-augmented pathways(mirAP) and their applications to pathway analysis and disease subtyping. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 22* (2016).

[88] Diaz, D., Donato, M., Nguyen, T., and Draghici, S. Microrna-augmented pathways (mirap) and their applications to pathway analysis and disease subtyping. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017* (2017), World Scientific, pp. 390–401.

[89] Diaz, D., and Draghici, S. *mirIntegrator: Integrating miRNAs into signaling pathways*, 2015. R package.

[90] Diaz, D., Nguyen, T., and Draghici, S. A systems biology approach for unsupervised clustering of high-dimensional data. In *Machine Learning, Optimization, and Big Data* (Aug 2016), Lecture Notes in Computer Science, Springer, Cham, pp. 193–203.

[91] Diaz Herrera, D. M. Integrative pathway analysis pipeline for mirna and mrna data. Master's thesis, Wayne State University, 2017.

[92] Diaz-Uriarte, R., and Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7* (2006), 3.

[93] Ding, C., and He, X. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning* (2004), ACM, p. 29.

[94] Ding, C., He, X., and Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining* (2005), SIAM, pp. 606–610.

[95] Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. Improving

gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics 8*, 1 (2007), 242.

[96] Donahue, T. R., Tran, L. M., Hill, R., Li, Y., Kovochich, A., Calvopina, J. H., Patel, S. G., Wu, N., Hindoyan, A., Farrell, J. J., et al. Integrative survival-based molecular profiling of human pancreatic cancer. *Clinical Cancer Research 18*, 5 (2012), 1352–1363.

[97] Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene expression profile from microarray data. *Genome biology 4*, 1 (2003), R7.

[98] Drăghici, S., Chen, D., and Reifman, J. Applications and challenges of DNA microarray technology in military medical research. *Military Medicine 169*, 8 (2004), 654–659.

[99] Drăghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichiţa, C., Georgescu, C., and Romero, R. A systems biology approach for pathway level analysis. *Genome Research 17*, 10 (2007), 1537–1545.

[100] Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. Global functional profiling of gene expression. *Genomics 81*, 2 (2003), 98–104.

[101] Drăghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichiţa, C., Georgescu, C., and Romero, R. A systems biology approach for pathway level analysis. *Genome Research 17*, 10 (2007), 1537–1545.

[102] Dudoit, S., and Fridlyand, J. A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology 3*, 7 (2002), 0036.1–0036.21.

[103] Dudoit, S., and Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology 3*, 7 (June 2002), research0036.1.

[104] Dumitrescu, L., Ritchie, M. D., Denny, J. C., El Rouby, N. M., McDonough, C. W., Bradford, Y., Ramirez, A. H., Bielinski, S. J., Basford, M. A., Chai, H. S., and et al. Genome-wide study of resistant hypertension identified from electronic health records. *PLOS ONE 12*, 2 (Feb 2017), e0171745.

[105] Dweep, H., and Gretz, N. miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. *Nature Methods 12*, 8 (2015), 697–697.

[106] Edgar, R., Domrachev, M., and Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research 30*, 1 (2002), 207–210.

[107] Efron, B., and Tibshirani, R. On testing the significance of sets of genes. *The Annals of Applied Statistics 1*, 1 (2007), 107–129.

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

[109] Ein-Dor, L., Zuk, O., and Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *In Proceedings of the National Academy of Sciences 103*, 15 (2006), 5923–5928.

[110] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences* [111], pp. 14863–14868.

[111] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences 95*, 25 (1998), 14863–14868.

[112] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences 95*, 25 (1998), 14863–14868.

[113] Emmert-Streib, F., and V. Glazko, G. Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Computational Biology 7*, 5 (2011), e1002053.

[114] Esteva, F. J., Sahin, A. A., Cristofanilli, M., Coombes, K., Lee, S.-J., Baker, J., Cronin, M., Walker, M., Watson, D., Shak, S., and et al. Prognostic role of a multigene reverse transcriptase-pcr assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. *Clinical cancer research: an official journal of the American Association for Cancer Research 11*, 9 (May 2005), 3315–9.

[115] Fakoor, R., Ladhak, F., Nazi, A., and Huber, M. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning* (2013).

[116] Filipczak, P. T., Leng, S., Tellez, C. S., Do, K. C., Grimes, M. J., Thomas, C. L., Walton-Filipczak, S. R., Picchi, M. A., and Belinsky, S. A. p53-suppressed oncogene tet1 prevents cellular aging in lung cancer. *Cancer research 79*, 8 (2019), 1758–1768.

[117] Fisher, R. A. *Statistical methods for research workers.* Oliver & Boyd, Edinburgh, 1925.

[118] Fleischer, T., Frigessi, A., Johnson, K. C., Edvardsen, H., Touleimat, N., Klajic, J., Riis, M. L., Haakensen, V. D., Wärnberg, F., Naume, B., et al. Genome-wide dna methylation profiles in progression to in situand invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome biology 15*, 8 (2014), 435.

[119] Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemporary clinical trials communications 11* (2018), 156–164.

[120] Foundation, N. B. C. Types of breast cancer, 2019.

[121] Freeman, H. Race, poverty, and cancer. *J Natl Cancer Inst 83*, 8 (1991), 526–527.

[122] Friedman, A. A., Letai, A., Fisher, D. E., and Flaherty, K. T. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer 15*, 12 (Nov 2015), 747–756.

[123] Friedman, J. H., and Meulman, J. J. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*, 4 (Nov. 2004), 815–849.

[124] Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science 303*, 5659 (Feb. 2004), 799–805.

[125] Fu, L., and Medico, E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics 8*, 1 (2007), 3.

[126] G. S. Firestein, D. S. P. DNA microarrays: boundless technology or bound by technology? Guidelines for studies using microarray technology. *Arthritis and Rheumatism 46*, 4 (2002), 859–861. Available online at http://linkage.rockefeller.edu/wli/microarray/firestein02.pdf.

[127] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., and Schultz, N. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science signaling 6*, 269 (Apr. 2013), pl1.

[128] Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal. 6*, 269 (2013), pl1–pl1.

[129] Gaujoux, R., and Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics 11* (July 2010), 367.

[130] Ge, H., Walhout, A. J. M., and Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends in Genetics 19*, 10 (Oct. 2003), 551–560.

[131] Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Della Porta, M. G., Jädersten, M., Dolatshad, H., Verma, A., Cross, N. C., Vyas, P., et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature communications 6* (2015), 5901.

[132] Ghoncheh, M., Pournamdar, Z., and Salehiniya, H. Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pacific Journal of Cancer Prevention 17*, S3 (2016), 43–46.

[133] Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics 28*, 18 (2012), i451–i457.

[134] Glaab, E., Baudot, A., Krasnogor, N., and Valencia, A. TopoGSA: network topological gene set analysis. *Bioinformatics 26*, 9 (2010), 1271–1272.

[135] GO. Gene Ontology. Tech. rep., Gene Ontology Consortium, 2001. http://www.geneontology.org/.

[136] Goel, M. K., Khanna, P., and Kishore, J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res 1*, 4 (2010), 274–278.

[137] Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics 20*, 1 (2004), 93–99.

[138] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 5439 (1999), 531–537.

[139] Gosset, W. S. The Probable Error of a Mean. *Biometrika 6* (1908), 1–25.

[140] Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., and et al. Patterns of somatic mutation in human cancer genomes. *Nature 446*, 7132 (Mar 2007), 153–158.

[141] Grumolato, L., Elkahloun, A., Ghzili, H., Alexander, D., Coulouarn, C., Yon, L., Salier, J., Eiden, L., Fournier, A., Vaudry, H., and Anouar, Y. Microarray and suppression subtractive hybridization analyses of gene expression in pheochromocytoma cells reveal pleiotropic effects of pituitary adenylate cyclase-activating polypeptide on cell proliferation, survival, and adhesion. *Endocrinology. 144*, 6 (June 2003), 2368–2379.

[142] Hall, M. A. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[143] Hall, M. A. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[144] Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. Co-clustering of biological networks and gene expression data. *Bioinformatics 18*, suppl 1 (July 2002), S145–S154.

[145] Harshman, R. A. Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis. *UCLA Working Papers in Phonetics 16*, 16 (1970), 1–84.

[146] Håstad, J. Tensor rank is NP-complete. *Journal of Algorithms 11*, 4 (Dec. 1990), 644–654.

[147] Hatfield, G., Hung, S.-p., and Baldi, P. Differential analysis of dna microarray gene expression data. *Molecular microbiology 47*, 4 (2003), 871–877.

[148] Henegar, C., Cancello, R., Rome, S., Vidal, H., Clément, K., and Zucker, J.-D. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. *Journal of bioinformatics and computational biology 4*, 04 (2006), 833–852.

[149] Hernández-Torruco, J., Canul-Reich, J., Frausto-Solis, J., and Mendez-Castillo, J. J. Feature Selection for Better Identification of Subtypes of Guillain-Barre Syndrome. *Computational and Mathematical Methods in Medicine, Computational and Mathematical Methods in Medicine 2014, 2014* (Sept. 2014), e432109.

[150] Hernández-Torruco, J., Canul-Reich, J., Frausto-Solís, J., and Méndez-Castillo, J. J. Feature Selection for Better Identification of Subtypes of Guillain-

Barré Syndrome. *Computational and Mathematical Methods in Medicine, Computational and Mathematical Methods in Medicine 2014, 2014* (Sept. 2014), e432109.

[151] Herrero, J., Valencia, A., and Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics 17*, 2 (2001), 126–136.

[152] Higginson, J., Muir, C. S., and Munoz, N. *Human cancer: epidemiology and environmental causes*. Cambridge University Press, 1992.

[153] Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America 106*, 23 (Jun 2009), 9362–7.

[154] Hira, Z. M., Gillies, D. F., Hira, Z. M., and Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Advances in Bioinformatics, Advances in Bioinformatics 2015, 2015* (June 2015), e198363.

[155] Hitchcock, F. L. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics 6*, 1-4 (1927), 164–189.

[156] Ho, J. C., Ghosh, J., Steinhubl, S. R., Stewart, W. F., Denny, J. C., Malin, B. A., and Sun, J. Limestone: High-throughput candidate phenotype generation via tensor factorization. *Journal of Biomedical Informatics 52* (2014), 199–211.

[157] Ho, J. C., Ghosh, J., and Sun, J. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of*

the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2014), pp. 115–124.

[158] Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell 158*, 4 (2014), 929–944.

[159] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. Network-based stratification of tumor mutations. In *Nature Methods* [160], pp. 1108–1115.

[160] Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. Network-based stratification of tumor mutations. *Nature Methods 10*, 11 (2013), 1108–1115.

[161] Hripcsak, G., and Albers, D. J. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association 20*, 1 (Jan 2013), 117–121.

[162] Hsu, C.-H., Peng, K.-L., Kang, M.-L., Chen, Y.-R., Yang, Y.-C., Tsai, C.-H., Chu, C.-S., Jeng, Y.-M., Chen, Y.-T., Lin, F.-M., et al. Tet1 suppresses cancer invasion by activating the tissue inhibitors of metalloproteinases. *Cell reports 2*, 3 (2012), 568–579.

[163] Hsu, J. J., Finkelstein, D. M., and Schoenfeld, D. A. Outcome-driven cluster analysis with application to microarray data. *PLOS ONE 10*, 11 (Nov 2015), e0141874.

[164] Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., and Chiu, C.-M. miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Research* (2010), D163–D169.

[165] Hsu, S.-D., Tseng, Y.-T., Shrestha, S., Lin, Y.-L., Khaleel, A., Chou, C.-H., Chu, C.-F., Huang, H.-Y., Lin, C.-M., Ho, S.-Y., Jian, T.-Y., Lin, F.-M., Chang, T.-H., Weng, S.-L., Liao, K.-W., Liao, I.-E., Liu, C.-C., and Huang, H.-D. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research 42*, D1 (Jan. 2014), D78–D85.

[166] Huang, D., and Pan, W. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics 22*, 10 (May 2006), 1259–1268.

[167] Huang, D. W., Sherman, B. T., and Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research 37*, 1 (2009), 1–13.

[168] Huang, G. T., Cunningham, K. I., Benos, P. V., and Chennubhotla, C. S. Spectral clustering strategies for heterogeneous disease expression data. *Pacific Symposium on Biocomputing* (2013), 212–223.

[169] Huang, H., Jiang, X., Li, Z., Li, Y., Song, C.-X., He, C., Sun, M., Chen, P., Gurbuxani, S., Wang, J., et al. Tet1 plays an essential oncogenic role in mll-rearranged leukemia. *Proceedings of the National Academy of Sciences 110*, 29 (2013), 11994–11999.

[170] Husain, I., Mohler, J. L., Seigler, H. F., and Besterman, J. M. Elevation of topoisomerase i messenger rna, protein, and catalytic activity in human tumors: demonstration of tumor-type specificity and implications for cancer chemotherapy. *Cancer research 54*, 2 (1994), 539–546.

[171] Ibrahim, R., Yousri, N. A., Ismail, M. A., and El-Makky, N. M. Multi-level gene/MiRNA feature selection using deep belief nets and active learning. In *2014 36th*

Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Aug. 2014), pp. 3957–3960.

[172] Institute, N. C. Immunotherapy for cancer, 2019. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy.

[173] Isci, S., Ozturk, C., Jones, J., and Otu, H. H. Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics 27*, 12 (2011), 1667–1674.

[174] Jacob, L., Neuvial, P., and Dudoit, S. Gains in power from structured two-sample tests of means on graphs. *Arxiv preprint arXiv:1009.5173* (2010).

[175] Jedroszka, D., Orzechowska, M., Baryla, I., and Bednarek, A. Po-152 differentiation of lung squamous cell carcinoma (lusc) and lung adenocarcinoma (luad) by gene co-expression analysis of notch signalling targets, 2018.

[176] Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., and Liu, Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research 37*, suppl 1 (2009), D98–D104.

[177] Jiang, Z., and Gentleman, R. Extensions to gene set enrichment. *Bioinformatics 23*, 3 (2007), 306–313.

[178] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. Human MicroRNA Targets. *PLOS Biology 2*, 11 (Oct. 2004), e363.

[179] Jørgensen, K. J., Gøtzsche, P. C., Kalager, M., and Zahl, P.-H. Breast cancer screening in denmark: a cohort study of tumor size and overdiagnosis. *Annals of Internal Medicine 166*, 5 (2017), 313–323.

[180] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S.,

Birney, E., and Stein, L. REACTOME: a knowledgebase of biological pathways. *Nucleic Acids Research 33*, Database issue (2005), D428–432.

[181] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Research 36*, suppl 1 (2008), D480–D484.

[182] Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research 28*, 1 (2000), 27–30.

[183] Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. In *Nucleic Acids Research* [182], pp. 27–30.

[184] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. From genomics to chemical genomics: new developments in kegg. *Nucleic acids research 34*, suppl 1 (2006), D354–D357.

[185] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Research 30*, 1 (January 2002), 42–46.

[186] Kapp, A. V., and Tibshirani, R. Are clusters found in one dataset present in another dataset? *Biostatistics 8*, 1 (2006), 9–31.

[187] Khatri, P., and Drăghici, S. A comparison of existing tools for ontological analysis of gene expression data. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Wiley Online Library, 2005, ch. 4. 4.5:54.

[188] Khatri, P., Drăghici, S., Ostermeier, G. C., and Krawetz, S. A. Profiling gene expression using Onto-Express. *Genomics 79*, 2 (2002), 266–270.

[189] Khatri, P., Sirota, M., and Butte, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology 8*, 2 (2012), e1002375.

[190] Kho, A. N., Hayes, M. G., Rasmussen-Torvik, L., Pacheco, J. A., Thompson, W. K., Armstrong, L. L., Denny, J. C., Peissig, P. L., Miller, A. W., Wei, W.-Q., and et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the American Medical Informatics Association 19*, 2 (Mar 2012), 212–218.

[191] Kho, A. N., Pacheco, J. A., Peissig, P. L., Rasmussen, L., Newton, K. M., Weston, N., Crane, P. K., Pathak, J., Chute, C. G., Bielinski, S. J., and et al. Electronic medical records for genetic research: Results of the eMERGE consortium. *Science Translational Medicine 3*, 79 (Apr 2011), 79re1–79re1.

[192] Kim, E.-Y., Kim, S.-Y., Ashlock, D., and Nam, D. MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. In *BMC Bioinformatics* [193], p. 260.

[193] Kim, E.-Y., Kim, S.-Y., Ashlock, D., and Nam, D. MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics 10* (2009), 260.

[194] Kim, H.-S., Oh, S. H., Kim, J.-H., Kim, J.-Y., Kim, D.-H., Lee, S.-J., Choi, S.-U., Park, K. M., Ryoo, Z. Y., Park, T. S., et al. Mll-tet1 fusion protein promotes immortalization of myeloid progenitor cells and leukemia development. *haematologica 102*, 11 (2017), e434.

[195] Kim, S., Herazo-Maya, J. D., Kang, D. D., Juan-Guardela, B. M., Tedrow, J., Martinez, F. J., Sciurba, F. C., Tseng, G. C., and Kaminski, N. Integrative phenotyping framework (iPF): integrative clustering of multiple

omics data identifies novel lung disease subphenotypes. *BMC Genomics 16*, 1 (Dec. 2015), 924.

[196] Kim, S.-Y., and Volsky, D. J. Page: parametric analysis of gene set enrichment. *BMC bioinformatics 6*, 1 (2005), 144.

[197] Kim, T. Y., Kim, H. U., and Lee, S. Y. Data integration and analysis of biological networks. *Current Opinion in Biotechnology 21*, 1 (Feb. 2010), 78–84.

[198] Kohonen, T. The self-organizing map. *Proceedings of the IEEE 78*, 9 (1990), 1464–1480.

[199] Kolda, T. G., and Bader, B. W. Tensor decompositions and applications. *SIAM Review 51*, 3 (2009), 455–500.

[200] Kong, S. W., Pu, W. T., and Park, P. J. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics 22*, 19 (2006), 2373–2380.

[201] Kotelnikova, E., Shkrob, M. A., Pyatnitskiy, M. A., Ferlini, A., and Daraselia, N. Novel approach to meta-analysis of microarray datasets reveals muscle remodeling-related drug targets and biomarkers in Duchenne muscular dystrophy. *PLoS Computational Biology 8*, 2 (2012), e1002365.

[202] Kramer, A., Green, J., Pollard, J., and Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics (Oxford, England) 30*, 4 (Feb 2014), 523–530.

[203] Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., et al. Combinatorial microRNA target predictions. *Nature genetics 37*, 5 (2005), 495–500.

[204] Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J. P., Heese, O., Reifenberger, G., et al. Long-term survival with glioblastoma multiforme. *Brain 130*, 10 (2007), 2596–2606.

[205] Krzanowski, W. J., and Lai, Y. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics* (1988), 23–34.

[206] Kullo, I. J., Fan, J., Pathak, J., Savova, G. K., Ali, Z., and Chute, C. G. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *Journal of the American Medical Informatics Association 17*, 5 (Sep 2010), 568–574.

[207] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. In *Nature* [208], pp. 495–501.

[208] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature 505*, 7484 (Jan 2014), 495–501.

[209] Lay Jr., J. O., Liyanage, R., Borgmann, S., and Wilkins, C. L. Problems with the "omics". *TrAC Trends in Analytical Chemistry 25*, 11 (Dec. 2006), 1046–1056.

[210] Lee, U., Frankenberger, C., Yun, J., Bevilacqua, E., Caldas, C., Chin, S.-F., Rueda, O. M., Reinitz, J., and Rosner, M. R. A Prognostic Gene Signature for Metastasis-Free Survival of Triple Negative Breast Cancer Patients. *PLOS ONE 8*, 12 (Dec. 2013), e82125.

[211] Lee, Y. S., and Dutta, A. MicroRNAs in cancer. *Annual Review of Pathology 4* (2009).

[212] Leung, M. K. K., Xiong, H. Y., Lee, L. J., and Frey, B. J. Deep learning of the tissue-regulated splicing code. *Bioinformatics 30*, 12 (June 2014), i121–i129.

[213] Lewis, B. P., Burge, C. B., and Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell 120*, 1 (2005), 15–20.

[214] Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One 10*, 1 (2015), e0116774.

[215] Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., and Dudley, J. T. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine 7*, 311 (2015), 311ra174–311ra174.

[216] Li, T., Zhang, C., and Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics 20*, 15 (2004), 2429–2437.

[217] Li, Y., and Ngom, A. Non-negative matrix and tensor factorization based classification of clinical microarray gene expression data. In *2010 IEEE international conference on bioinformatics and biomedicine (BIBM)* (2010), IEEE, pp. 438–443.

[218] Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., and Cui, Q. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research 42*, Database issue (Jan. 2014), D1070–D1074.

[219] Liang, M., Li, Z., Chen, T., and Zeng, J. Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach. *IEEE/ACM Trans. Comput. Biol. Bioinformatics 12*, 4 (July 2015), 928–937.

[220] Lindeman, N. I., Cagle, P. T., Aisner, D. L., Arcila, M. E., Beasley, M. B., Bernicker, E. H., Colasacco, C., Dacic, S., Hirsch, F. R., Kerr, K., et al. Updated molecular testing guideline for the selection of lung cancer patients for treatment with targeted tyrosine kinase inhibitors: guideline from the college of american pathologists, the international association for the study of lung cancer, and the association for molecular pathology. *Journal of Thoracic Oncology 13*, 3 (2018), 323–358.

[221] Linn, R. J., Hall, D. L., and Llinas, J. Survey of multisensor data fusion systems. In *Data Structures and Target Classification* (1991), vol. 1470, pp. 13–29.

[222] Liu, H., Patel, M. R., Prescher, J. A., Patsialou, A., Qian, D., Lin, J., Wen, S., Chang, Y.-F., Bachmann, M. H., Shimono, Y., and et al. Cancer stem cells from human breast tumors are involved in spontaneous metastases in orthotopic mouse models. *Proceedings of the National Academy of Sciences of the United States of America 107*, 42 (2010), 18115–18120.

[223] Liu, Y., and Schumann, M. Data mining feature selection for credit scoring models. *Journal of the Operational Research Society 56*, 9 (Apr. 2005), 1099–1108.

[224] Lock, E. F., and Dunson, D. B. Bayesian consensus clustering. *Bioinformatics 29*, 20 (2013), 2610–2616.

[225] Lorsbach, R., Moore, J., Mathew, S., Raimondi, S., Mukatira, S., and Downing, J. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia 17*, 3 (2003), 637–641.

[226] Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., and Cui, Q. An analysis of human microRNA and disease associations. *PloS One 3*, 10 (2008), e3420.

[227] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. Extracting insights from the shape of complex data using topology. *Scientific reports 3* (2013), 1236.

[228] Luo, F., Khan, L., Bastani, F., Yen, I.-L., and Zhou, J. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. *Bioinformatics 20*, 16 (2004), 2605–2617.

[229] Luo, Y., Wang, F., and Szolovits, P. Tensor factorization toward precision medicine. *Briefings in Bioinformatics 18*, 3 (May 2017), 511–514.

[230] Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biology 5* (2004), R101.

[231] Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research 41*, 1 (2013), e19–e19.

[232] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., et al. Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research 37*, suppl 1 (2009), D619–D622.

[233] McCarthy, D. J., Chen, Y., and Smyth, G. K. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research 40*, 10 (2012), 4288–4297.

[234] McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., and et al. The emerge network: A consortium of biorepositories linked to electronic

medical records data for conducting genomic studies. *BMC Medical Genomics 4* (Jan 2011), 13.

[235] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., et al. The panther database of protein families, subfamilies, functions and pathways. *Nucleic acids research 33*, suppl 1 (2005), D284–D288.

[236] Mi, H., Muruganujan, A., and Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research 41*, D1 (2013), D377–D386.

[237] Milligan, G. W., and Cooper, M. C. An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*, 2 (1985), 159–179.

[238] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichiţa, C., and Drăghici, S. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology 4* (2013), 278.

[239] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning 52*, 1-2 (2003), 91–118.

[240] Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics 34*, 3 (Jul 2003), 267–273.

[241] Moreno-Risueno, M. A., Busch, W., and Benfey, P. N. Omics meet networks — using systems approaches to infer regulatory networks in plants. *Current Opinion in Plant Biology 13*, 2 (Apr. 2010), 126–131.

[242] Mørup, M. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1*, 1 (Jan 2011), 24–40.

[243] Mørup, M., and Hansen, L. K. Automatic relevance determination for multi-way models. *Journal of Chemometrics 23*, 7–8 (Jul 2009), 352–363.

[244] Mouw, K. W., Goldberg, M. S., Konstantinopoulos, P. A., and D'Andrea, A. D. Dna damage and repair biomarkers of immunotherapy response. *Cancer discovery 7*, 7 (2017), 675–693.

[245] Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., and et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature 487*, 7407 (Jul 2012), 330–337.

[246] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology 7*, 1 (2011), 548.

[247] Nam, S., Li, M., Choi, K., Balch, C., Kim, S., and Nephew, K. P. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Research 37*, suppl 2 (May 2009), W356–W362.

[248] Narod, S., Tung, N., Lubinski, J., Huzarski, T., Robson, M., Lynch, H. T., Neuhausen, S., Ghadirian, P., Kim-Sing, C., Sun, P., et al. A prior diagnosis of breast cancer is a risk factor for breast cancer in brca1 and brca2 carriers. *Current Oncology 21*, 2 (2014), 64.

[249] Network, C. G. A. R., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., and et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics 45*, 10 (2013), 1113.

[250] Nguyen, T., Tagett, R., Diaz, D., and Draghici, S. A novel approach for data integration and disease subtyping. *Genome Research 27*, 12 (2017), 2025–2039.

[251] Nicolau, M., Levine, A. J., and Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences 108*, 17 (2011), 7265–7270.

[252] Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA 13*, 11 (Nov. 2007), 1894–1910.

[253] Nielson, J. L., Cooper, S. R., Yue, J. K., Sorani, M. D., Inoue, T., Yuh, E. L., Mukherjee, P., Petrossian, T. C., Paquette, J., Lum, P. Y., et al. Uncovering precision phenotype-biomarker associations in traumatic brain injury using topological data analysis. *PloS one 12*, 3 (2017), e0169490.

[254] Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G. W., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell 17*, 5 (May 2010), 510–522.

[255] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research 27*, 1 (1999), 29–34.

[256] Omberg, L., Meyerson, J. R., Kobayashi, K., Drury, L. S., Diffley, J. F. X., and Alter, O. Global effects of dna replication and dna replication origin activity on eukaryotic gene expression. *Molecular Systems Biology 5*, 1 (Apr 2009), 312.

[257] Onitilo, A. A., Engel, J. M., Greenlee, R. T., and Mukesh, B. N. Breast cancer subtypes based on er/pr and her2 expression: comparison of clinicopathologic features and survival. *Clinical medicine & research 7*, 1-2 (2009), 4–13.

[258] Orzechowska, M., Jedroszka, D., Płuciennik, E., Kośla, K., Nowakowska, M., Baryła, I., Pospiech, K., Styczeń-Binkowska, E., and Bednarek, A. K. Notch signalling has predictive potential correlating with recurrence and stage of lung cancers. *Gliwise Scientific Meetings* (2016).

[259] Overhage, J. M., Ryan, P. B., Reich, C. G., Hartzema, A. G., and Stang, P. E. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association 19*, 1 (Jan. 2012), 54–60.

[260] Palsson, B., and Zengler, K. The challenges of integrating multi-omic data sets. *Nature Chemical Biology 6*, 11 (Nov. 2010), 787–789.

[261] Pan, W., and Shen, X. Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research 8*, May (2007), 1145–1164.

[262] Paoli, S., Jurman, G., Albanese, D., Merler, S., and Furlanello, C. Integrating gene expression profiling and clinical data. *International Journal of Approximate Reasoning 47*, 1 (Jan 2008), 58–69.

[263] Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Research 41*, Web Server issue (July 2013), W169–W173.

[264] Pathak, J., Kho, A. N., and Denny, J. C. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association 1* (Apr 2017), e206–e211.

[265] Pattabiraman, D. R., and Weinberg, R. A. Tackling the cancer stem cells?what challenges do they pose? *Nature reviews Drug discovery 13*, 7 (2014), 497–512.

[266] Pavlidis, P., Qin, J., Arango, V., Mann, J. J., and Sibille, E. Using the Gene Ontology for Microarray Data Mining: A Comparison of Methods and Application to Age Effects in Human Prefrontal Cortex. *Neurochemical Research 29*, 6 (June 2004), 1213 –1222.

[267] Peaeson, E., and Haetlet, H. Biometrika tables for statisticians. *Biometrika Trust* (1976).

[268] Pećina-Šlaus, N., and Pećina, M. Only one health, and so many omics. *Cancer Cell International 15*, 1 (June 2015), 64.

[269] Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O., and Botstein, D. Molecular portraits of human breast tumours. *Nature 406*, 6797 (2000), 747–752.

[270] Pok, G., Liu, J.-C. S., and Ryu, K. H. Effective feature selection framework for cluster analysis of microarray data. *Bioinformation 4*, 8 (Feb. 2010), 385–389.

[271] Prlić, A., and Procter, J. B. Ten Simple Rules for the Open Development of Scientific Software. *PLOS Comput Biol 8*, 12 (Dec. 2012), e1002802.

[272] Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E., and Kotelnikova, E. Clustering Gene Expression Regulators: New Approach to Disease Subtyping. *PLoS ONE 9*, 1 (Jan. 2014).

[273] Quitadamo, A., Tian, L., Hall, B., and Shi, X. An integrated network of microRNA and gene expression in ovarian cancer. *BMC Bioinformatics 16*, Suppl 5 (Mar. 2015), S5.

[274] Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology 3*, 1 (2004).

[275] Ramakrishnan, N., Hanauer, D., and Keller, B. Mining electronic health records. *Computer 43*, 10 (2010), 77–81.

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

[277] Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., Ledbetter, D. H., Maglott, D. R., Martin, C. L., Nussbaum, R. L., and et al. Clingen — the clinical genome resource. *New England Journal of Medicine 372*, 23 (Jun 2015), 2235–2242.

[278] Reis-Filho, J. S., and Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet 378*, 9805 (Nov 2011), 1812–1823.

[279] Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. Stem cells, cancer, and cancer stem cells. *nature 414*, 6859 (2001), 105–111.

[280] Riaz, M., van Jaarsveld, M. T., Hollestelle, A., Prager-van der Smissen, W. J., Heine, A. A., Boersma, A. W., Liu, J., Helmijr, J., Ozturk, B., Smid, M., et al. mirna expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific mirnas. *Breast cancer research 15*, 2 (2013), R33.

[281] Richesson, R. L., Hammond, W. E., Nahm, M., Wixted, D., Simon, G. E., Robinson, J. G., Bauck, A. E., Cifelli, D., Smerek, M. M., Dickerson, J., and et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the nih health care systems collaboratory. *Journal of the American Medical Informatics Association: JAMIA 20*, e2 (Dec 2013), e226–31.

[282] Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., Lee, W., Yuan, J., Wong, P., Ho, T. S., et al. Mutational landscape determines sensitivity to pd-1 blockade in non–small cell lung cancer. *Science 348*, 6230 (2015), 124–128.

[283] Robinson, D., Van Allen, E. M., Wu, Y.-M., Schultz, N., Lonigro, R. J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C. C., Attard, G., et al. Integrative clinical genomics of advanced prostate cancer. *Cell 161*, 5 (2015), 1215–1228.

[284] Robinson, P. N. Deep phenotyping for precision medicine. *Human Mutation 33*, 5 (May 2012), 777–780.

[285] Robinson, S. W., Fernandes, M., and Husi, H. Current advances in systems and integrative biology. *Computational and Structural Biotechnology Journal 11*, 18 (Aug. 2014), 35–46.

[286] Roe, M. T., Cyr, D. D., Eckart, D., Schulte, P. J., Morse, M. A., Blackwell, K. L., Ready, N. E., Zafar, S. Y., Beaven, A. W., Strickler, J. H., et al. Ascertainment, classification, and impact of neoplasm detection during prolonged treatment with dual antiplatelet therapy with prasugrel vs. clopidogrel following acute coronary syndrome. *European heart journal 37*, 4 (2016), 412–422.

[287] Roychowdhury, S., and Chinnaiyan, A. M. Advancing precision medicine for prostate cancer through genomics. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology 31*, 15 (May 2013), 1866–73.

[288] Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Ison, J., Keays, M., Kurbatova, N., Malone, J., Mani, R., Mupo, A., Pereira, R. P., Pilicheva, E., Rung, J., Sharma, A., Tang, Y. A., Ternent, T., Tikhonov, A., Welter, D., Williams, E., Brazma, A., Parkinson, H., and Sarkans, U. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research 41*, D1 (2013), D987–D990.

[289] Samur, M. K. Rtcgatoolbox: a new tool for exporting tcga firehose data. *PloS one 9*, 9 (2014), e106397.

[290] Samur, M. K., Yan, Z., Wang, X., Cao, Q., Munshi, N. C., Li, C., and Shah, P. K. canevolve: a web portal for integrative oncogenomics. *PLoS One 8*, 2 (2013), e56228.

[291] Saria, S., and Goldenberg, A. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems 30*, 4 (Jul 2015), 70–75.

[292] Schaefer, C., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. PID: the Pathway Interaction Database. *Nucleic acids research 37*, Database issue (2009), D674–D679.

[293] Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv. 34*, 1 (Mar. 2002), 1–47.

[294] Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA 12*, 2 (Feb. 2006), 192–197.

[295] Shai, R., Shi, T., Kremen, T. J., Horvath, S., Liau, L. M., Cloughesy, T. F., Mischel, P. S., and Nelson, S. F. Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene 22*, 31 (2003), 4918–4923.

[296] Sharma, A., Imoto, S., Miyano, S., and Sharma, V. Null space based feature selection method for gene expression data. *International Journal of Machine Learning and Cybernetics 3*, 4 (Nov. 2011), 269–276.

[297] Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., and Sander, C. Integrative Subtype Discovery in Glioblastoma Using iCluster. *PLOS ONE 7*, 4 (Apr. 2012), e35236.

[298] Shen, R., Olshen, A. B., and Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics 25*, 22 (2009), 2906–2912.

[299] Shen, R., Olshen, A. B., and Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics 25*, 22 (2009), 2906–2912.

[300] Shen, R., Wang, S., and Mo, Q. Sparse integrative clustering of multiple omics data sets. *The annals of applied statistics 7*, 1 (Apr. 2013), 269–294.

[301] Shi, Z., Wang, J., and Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nature Methods 10*, 7 (2013), 597–598.

[302] Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association: JAMIA 21*, 2 (2014), 221–30.

[303] Shojaie, A., and Michailidis, G. Analysis of Gene Sets Based on the Underlying Regulatory Network. *Journal of Computational Biology 16*, 3 (2009), 407–426.

[304] Si, Y., Liu, J., Shen, H., Zhang, C., Wu, Y., Huang, Y., Gong, Z., Xue, J., and Liu, T. Fisetin decreases tet 1 activity and ccny/cdk 16 promoter 5hmc levels to inhibit the proliferation and invasion of renal cancer stem cell. *Journal of cellular and molecular medicine 23*, 2 (2019), 1095–1105.

[305] Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing 65*, 13 (2017), 3551–3582.

[306] Siewert, J. R., Stein, H. J., Feith, M., Bruecher, B. L., Bartels, H., and Fink, U. Histologic tumor type is an independent prognostic parameter in esophageal cancer: lessons from more than 1,000 consecutive resections at a single center in the western world. *Annals of surgery 234*, 3 (2001), 360.

[307] Singh, H., Nugent, Z., Demers, A. A., Kliewer, E. V., Mahmud, S. M., and Bernstein, C. N. The reduction in colorectal cancer mortality after colonoscopy varies by site of the cancer. *Gastroenterology 139*, 4 (2010), 1128–1137.

[308] Sinicropi-Yao, S. L., Amann, J. M., Lopez, D. L. Y., Cerciello, F., Coombes, K. R., and Carbone, D. P. Co-expression analysis reveals mechanisms underlying the varied roles of notch1 in nsclc. *Journal of Thoracic Oncology 14*, 2 (2019), 223–236.

[309] Smid, M., Wang, Y., Zhang, Y., Sieuwerts, A. M., Yu, J., Klijn, J. G., Foekens, J. A., and Martens, J. W. Subtypes of breast cancer show preferential site of relapse. *Cancer research 68*, 9 (2008), 3108–3114.

[310] Smyth, G. K. Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, Eds. Springer, New York, 2005, pp. 397–420.

[311] Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., Walsh, L. A., Postow, M. A., Wong, P., Ho, T. S., et al. Genetic basis for clinical response to ctla-4 blockade in melanoma. *New England Journal of Medicine 371*, 23 (2014), 2189–2199.

[312] Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer 18*, 11 (2018), 696–705.

[313] Sotiriou, C., and Pusztai, L. Gene-expression signatures in breast cancer. *New England Journal of Medicine 360*, 8 (2009), 790–800.

[314] Stessman, H. A., Bernier, R., and Eichler, E. E. A Genotype-First Approach to Defining the Subtypes of a Complex Disease. *Cell 156*, 5 (Feb. 2014), 872–877.

[315] Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *science 331*, 6024 (2011), 1553–1558.

[316] Su, P.-H., Hsu, Y.-W., Huang, R.-L., Chen, L.-Y., Chao, T.-K., Liao, C.-C., Chen, C.-W., Wu, T.-I., Mao, S.-P., Balch, C., et al. Tet1 promotes 5hmc-dependent stemness, and inhibits a 5hmc-independent epithelial-mesenchymal transition, in cervical precancerous lesions. *Cancer letters 450* (2019), 53–62.

[317] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences of the Unites States of America 102*, 43 (2005), 15545–15550.

[318] Sugar, C. A., and James, G. M. Finding the number of clusters in a dataset. *Journal of the American Statistical Association 98*, 463 (2003).

[319] Sun, J., Bi, J., and Kranzler, H. R. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC Genetics 15*, 1 (2014), 1.

[320] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences 96*, 6 (1999), 2907–2912.

[321] Tan, P. K., Downey, T. J., Spitznagel Jr, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research 31*, 19 (2003), 5676–5684.

[322] Tarca, A. L., Drăghici, S., Bhatti, G., and Romero, R. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics 13*, 1 (2012), 136.

[323] Tarca, A. L., Drăghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. A novel signaling pathway impact analysis. *Bioinformatics 25*, 1 (2009), 75–82.

[324] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. *Nature Genetics 22* (1999), 281–285.

[325] TCGA Research Network. The Cancer Genome Atlas. http://cancergenome.nih.gov/.

[326] Therneau, T. M., and Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health. Springer New York, New York, NY, 2000.

[327] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. Discovering statistically significant pathways in expression profiling studies. *Proceeding of The National Academy of Sciences of the USA 102*, 38 (2005), 13544–13549.

[328] Tibshirani, R., and Walther, G. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics 14*, 3 (Sept. 2005), 511–528.

[329] Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*, 2 (2001), 411–423.

[330] Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. Global cancer incidence and mortality rates and trends?an update. *Cancer Epidemiology and Prevention Biomarkers 25*, 1 (2016), 16–27.

[331] Tseng, G. C., Ghosh, D., and Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research 40*, 9 (2012), 3785–3799.

[332] Tseng, G. C., Ghosh, D., and Zhou, X. J. *Integrating Omics Data*. Cambridge University Press, Aug. 2015.

[333] Tseng, G. C., and Wong, W. H. Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics 61*, 1 (2005), 10–16.

[334] Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika 31*, 3 (Sept. 1966), 279–311.

[335] Turner, M. C. Epidemiology: allergy history, ige, and cancer. *Cancer Immunology, Immunotherapy 61*, 9 (2012), 1493–1510.

[336] Uppu, S., Krishna, A., and Gopalan, R. Towards deep learning in genome-wide association interaction studies. *PACIS 2016 Proceedings* (June 2016).

[337] Vainio, H., and Wilbourn, J. Cancer etiology: Agents causally associated with human cancer. *Pharmacology & toxicology 72* (1993), 4–11.

[338] Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Foppen, M. H. G., Goldinger, S. M., et al. Genomic correlates of response to ctla-4 blockade in metastatic melanoma. *Science 350*, 6257 (2015), 207–211.

[339] Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J. M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics 26*, 12 (2010), i237–i245.

[340] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., Hayes, D. N., and Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell 17*, 1 (2010), 98–110.

[341] Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., and Hayes, D. N. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell 17*, 1 (Jan. 2010), 98–110.

[342] Vernon, S. W., Tilley, B. C., Neale, A. V., and Steinfeldt, L. Ethnicity, survival, and delay in seeking treatment for symptoms of breast cancer. *Cancer 55*, 7 (1985), 1563–1571.

[343] Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T., and Hatzigeorgiou, A. G. DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic Acids Research 43*, W1 (2015), W460–W466.

[344] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. Cancer genome landscapes. *science 339*, 6127 (2013), 1546–1558.

[345] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods 11*, 3 (2014), 333–337.

[346] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. In *Nature Methods* [345], pp. 333–337.

[347] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. In *Nature Methods* [345], pp. 333–337.

[348] Wang, S., and Zhu, J. Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics 64*, 2 (June 2008), 440–448.

[349] Wang, X., Min, S., Liu, H., Wu, N., Liu, X., Wang, T., Li, W., Shen, Y., Wang, H., Qian, Z., et al. Nf1 loss promotes kras-driven lung adenocarcinoma and results in psat1-mediated glutamate dependence. *EMBO molecular medicine 11*, 6 (2019).

[350] Wang, Y., Chen, R., Ghosh, J., Denny, J. C., Kho, A., Chen, Y., Malin, B. A., and Sun, J. Rubik: Knowledge guided tensor factorization and completion for health data analytics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), pp. 1265–1274.

[351] Wang, Y. C., Deng, N., Chen, S., and Wang, Y. Computational Study of Drugs by Integrating Omics Data with Kernel Methods. *Molecular Informatics 32*, 11-12 (Dec. 2013), 930–941.

[352] Ward, E., Jemal, A., Cokkinides, V., Singh, G. K., Cardinez, C., Ghafoor, A., and Thun, M. Cancer disparities by race/ethnicity and socioeconomic status. *CA: a cancer journal for clinicians 54*, 2 (2004), 78–93.

[353] Watson, J. D., and Crick, F. H. The structure of dna. In *Cold Spring Harbor symposia on quantitative biology* (1953), vol. 18, Cold Spring Harbor Laboratory Press, pp. 123–131.

[354] WEST, D. W., SLATTERY, M. L., ROBISON, L. M., SCHUMAN, K. L., FORD, M. H., MAHONEY, A. W., LYON, J. L., and SORENSEN, A. W.

Dietary intake and colon cancer: sex-and anatomic site-specific associations. *American journal of epidemiology 130*, 5 (1989), 883–894.

[355] Wishart, G., Bajdik, C., Dicks, E., Provenzano, E., Schmidt, M., Sherman, M., Greenberg, D., Green, A., Gelmon, K., Kosma, V., et al. Predict plus: development and validation of a prognostic model for early breast cancer that includes her2. *British Journal of Cancer 107*, 5 (2012), 800.

[356] Witten, D. M., and Tibshirani, R. A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association 105*, 490 (June 2010), 713–726.

[357] Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*, 3 (July 2009), 515–534.

[358] Wu, M.-Z., Chen, S.-F., Nieh, S., Benner, C., Ger, L.-P., Jan, C.-I., Ma, L., Chen, C.-H., Hishida, T., Chang, H.-T., et al. Hypoxia drives breast tumor malignancy through a tet–tnfα–p38–mapk signaling axis. *Cancer research 75*, 18 (2015), 3912–3924.

[359] Xia, J., and Wishart, D. S. MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics 26*, 18 (2010), 2342–2344.

[360] Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research 37*, Database issue (Jan. 2009), D105–110.

[361] Xie, B., Pan, W., and Shen, X. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic journal of statistics 2* (2008), 168–212.

[362] Xie, Y., and Ahn, C. Statistical Methods for Integrating Multiple Types of High-Throughput Data. *Methods in molecular biology (Clifton, N.J.) 620* (2010), 511–529.

[363] Yang, D., Sun, Y., Hu, L., Zheng, H., Ji, P., Pecot, C. V., Zhao, Y., Reynolds, S., Cheng, H., Rupaimoole, R., et al. Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell 23*, 2 (2013), 186–199.

[364] Yang, L., Wang, S., Zhou, M., Chen, X., Jiang, W., Zuo, Y., and Lv, Y. Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network. *Scientific reports 7*, 1 (2017), 738.

[365] Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. Normalization for cdna microarry data. In *Microarrays: optical technologies and informatics* (2001), vol. 4266, International Society for Optics and Photonics, pp. 141–152.

[366] Yener, B., Acar, E., Agius, P., Bennett, K., Vandenberg, S. L., and Plopper, G. E. Multiway modeling and analysis in stem cell systems biology. *BMC Systems Biology 2*, 1 (Jul 2008), 63.

[367] Zheng, Z., Wu, X., and Srihari, R. Feature Selection for Text Categorization on Imbalanced Data. *SIGKDD Explor. Newsl. 6*, 1 (June 2004), 80–89.

[368] Zhou, Z., Zhang, H.-S., Liu, Y., Zhang, Z.-G., Du, G.-Y., Li, H., Yu, X.-Y., and Huang, Y.-H. Loss of tet1 facilitates dld1 colon cancer cell migration via h3k27me3-mediated down-regulation of e-cadherin. *Journal of cellular physiology 233*, 2 (2018), 1359–1369.

[369] Zhu, Y., Qiu, P., and Ji, Y. Tcga-assembler: open-source software for retrieving and processing tcga data. *Nature methods 11*, 6 (2014), 599.

ABSTRACT

METHODS TO INTEGRATE GENETIC AND CLINICAL DATA FOR DISEASE SUBTYPING

by

DIANA MABEL DIAZ HERRERA

August 2020

Advisor: Dr. Alexander Kotov

Major: Computer Science

Degree: Doctor of Philosophy

Enormous efforts have been made to collect genetic and clinical data from cancer patients to advance the understanding of disease development and progression. Processing and analyzing these flows of data is challenging. This thesis is a contribution towards the integration of clinical and genetic data using computational methods. We present here three new data integration approaches to elucidate granular and meaningful disease sub-types from high-dimensional complex genetic and clinical variables, an essential step towards personalized medicine which is considered the future for oncology studies.

First we proposed disSuptyper, a pipeline that integrates biological knowledge, gene expression data, survival data, and biological pathways using statistical analyzes and unsupervised methods. Second we proposed CLIGEN, a tensor-based method that analyzes somatic mutation and clinical variables jointly using CP tensor factorization. Third we proposed TGENEX, a method that integrates gene expression, somatic mutation and clinical variables to identify candidate disease sub-types.

AUTOBIOGRAPHICAL STATEMENT

DIANA DIAZ

EDUCATION

- Doctor of Philosophy (Computer Science), 2019
  Wayne State University, Detroit, MI, USA
- Master of Science (Computer Science), 2017
  Wayne State University, Detroit, MI, USA
- Master of Science (Computer Systems Engineering), 2009
  The Andes University, Colombia
- Bachelor of Engineering (Computer Systems Engineering), 2006
  The Piloto University of Colombia, Colombia

PUBLICATIONS

1. **Diaz D**, Bollig-Fischer, A., **Kotov, A** Tensor Decomposition for Sub-typing of Complex Diseases based on Clinical and Genomic Data. BIBM 2019: International Conference on Bioinformatics & Biomedicine
2. **Diaz D**, **Kotov, A** Computational Methods for Cancer Subtyping: A Systematic Literature Review. Accepted to the 19th International Conference on Computational Science and its Applications (2019).
3. **Diaz D**, **Kotov, A**, Bollig-Fischer, A. Joint analysis of clinical records and genetic data using CP tensor decomposition. *NeurIPS 2018 Workshops ML4H and WIML* (2018).
4. Nguyen, T., Tagett, R., **Diaz D**, Draghici S. A novel approach for data integration and disease sub-typing.. *Genome research* (2017).
5. **Diaz D**, Donato M, Nguyen T, Draghici S. MicroRNA-Augmented Pathways (mirAP) and their applications to pathway analysis and disease sub-typing. In proceedings of *Pacific Symposium on Biocomputing (PSB) 2017.* (2017).
6. **Diaz D**, Nguyen T, Draghici S. A Systems Biology Approach for Unsupervised Clustering of High-Dimensional Data. In proceedings of *International Workshop on Machine Learning, Optimization and Big Data.* (2016).
7. Nguyen T, **Diaz D**, Tagett R, Draghici S. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Scientific Reports, 6, 29251* (2016).
8. **Diaz D**, Draghici S. mirIntegrator: Integrating miRNAs into signaling pathways. *Bioconductor* (2015).